AI Guidelines for Business
Appendix Ver1.1
April 4, 2025

Appendix. Preface

Structure of appendix and expectations for readers

The main part of the AI Guidelines for Business (hereinafter referred to as the “main part”) presented the basic philosophies (= why) that should be kept in mind by AI business actors (AI developers, AI providers, and AI business users) who were intended readers of this guideline, and the guiding principles (= what) in actions that should be taken on AI based on the philosophies. AI business actors need to determine what specific approach should be taken to implement the guiding principles. This appendix to the guideline (attached material; hereinafter referred to as the “appendix”) deals with implementations (= how) for reference, and is intended as a reference for specific actions.

Appendix 1 presents examples of AI systems and services that are assumed by this guideline, specific use examples, examples of patterns of AI business actors, examples of benefits from AI to each industry and business operation, and risks taking actual cases as examples. Appendix 2 presents contents for deepening understanding of actions to be taken by business operators for building AI governance through behavioral goals and practical examples.

Appendices 3, 4, and 5 describe important matters for AI developers, AI providers, and AI business users, respectively. Each of the appendices is divided into Parts A and B. Part A gives supplemental descriptions of important matters for each AI business actor described in one of Parts 3 to 5 of the main part and describes specific methods for the implementation. Part B describes specific methods for specifically important items in “Common guiding principles” described in Part 2 of the main part, though they are not described in Parts 3 to 5 of the main part.

Appendix 6 describes matters to be kept in mind when you refer to “Contract Guidelines on Utilization of AI and Data” which can be used as a reference when closing a contract for use of data. (Appendices 7 to 9(Japanese Only) shown in “Figure 1. Structure of this guideline” are contained in another material separate from this material.)

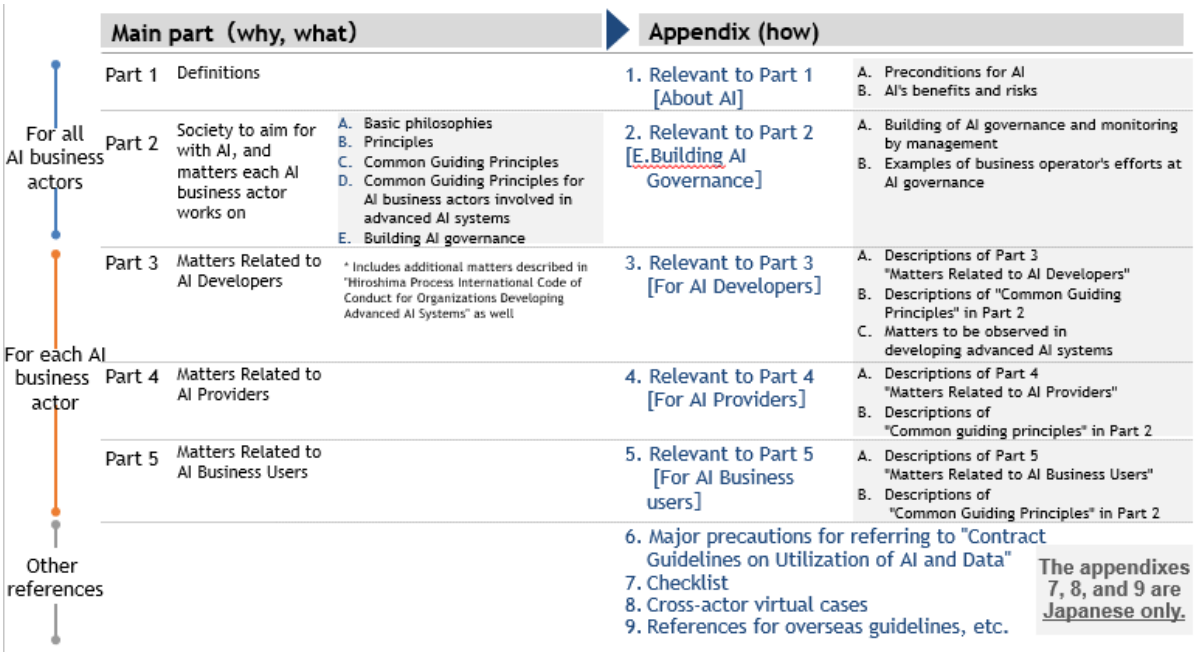


Figure 1. Structure of this guideline

It is assumed that reading Appendix 1 along with the main part and other appendices enables to specifically understand the AI assumed by the descriptions and benefits and risks to be brought about by AI, helping more deeply understand the descriptions. Reading Appendix 2 also enables to understand the behavioral goals for building AI governance on the business operator's AI use through specific practical examples, so it is important for all the AI developers, AI providers, and AI business users to read Appendices 1 and 2.

Appendices 3 to 5 are specifically intended for AI developers, AI providers, and AI business users, respectively. Therefore, it is important for AI business actors to read the one intended for them and determine and take actions referring to the practical examples in the appendix. It is also expected to understand the descriptions in the appendices for other AI business actors as much as possible along with the main part, as doing so would lead to the consideration of measures for reducing risks in the whole value chain.

In addition, to stably devise and take actions for gaining benefits while diminishing the risk in AI, it is important to create and effectively use a checklist suited to the business and circumstances of each business operator referring to the checklist shown in Appendix 7 (separate material). The checklist format contains ten guiding principles and important matters to be checked, to enable to check whether or not the guiding principles and important matters described in Part 2. C. of the main part are implemented. It has been created assuming that each business operator would customize it in accordance with the circumstances as necessary. Examples for AI developers, AI providers, and AI business users, respectively, are also presented as references. Furthermore, the appendix contains a format that can be used by business operators involved in advanced AI systems described in Part 2. D. of the main part for checking the implementation of important matters and a format that can be used for checking the building of AI governance described in Appendix 2. It has a structure that can contribute to the implementation checks ranging from actions to AI governance. Actual AI services are assumed to be used in various cases depending on the purpose, used technology, data, usage environment, etc. Therefore, it is expected that AI developers, AI providers, and AI business users cooperate with each other to devise the optimum approach while considering the advancement of technologies, changes in the external environment, etc. Doing so is assumed to help the effective cooperation.

Incidentally, this guideline, throughout the main part and appendix, has been compiled under the concept of risk-based approach. It is expected that business operators also identify matters to which they should concentrate their efforts and matters that do not require such efforts, effectively take measures and build AI governance. The appendix shows a mere example of means for achieving the course of action presented in the main part and does not provide comprehensive implementations and descriptions of all the guiding principles described in the main part, and the business operation style is assumed to vary with the business operator. Therefore, it is not required to implement all the descriptions in this appendix as they are.

Descriptions of expressions in this guideline

Hereinafter, in the same way as for the main part, each matter (item) described in "Table 1. Important matters for each AI business actor in addition to common guiding principles" will be identified and indicated with the notation [AI business actor - Guiding principle number) Description.].

- An AI business actor is indicated by its initial: AI **D**eveloper, AI **P**rovider, and AI Business **U**ser. A guiding principle number and description number are indicated by numbers, respectively, given in the table.

"D-2) i," for example, refers to the important matter for AI developers about the proper data training regarding safety.

As for the matters expressed as "-" in the table, AI business actors are expected to implement the actions described in the "Part 2. C. Common guiding principles" column of the main part, rather than doing nothing.

Table 1. Important matters for each AI business actor in addition to common guiding principles

	Part 2. C. Common guiding principles	Important matters for each AI business actor in addition to common guiding principles		
		Part 3. AI Developer (D)	Part 4. AI Provider (P)	Part 5. AI Business User (U)
1) Human-centric	(1) Human dignity and autonomy of individuals (2) Paying attention to manipulations by AI on decision-makings and emotions (3) Countermeasures against disinformation (4) Ensuring diversity/inclusion (5) Providing user support (6) Ensuring sustainability	-	-	-
2) Safety	(1) Taking into consideration the lives, bodies, properties and minds of humans and the environment (2) Proper use (of AI) (3) Proper training	i. Proper data training ii. Development that takes into consideration the lives, bodies, properties and minds of humans and the environment iii. Development contributing to proper use (of AI)	i. Actions against risks that consider the lives, bodies, properties, and minds of human and the environment ii. Provision contributing to proper use (of AI)	i. Proper use (of AI) that considers safety
3) Fairness	(1) Consideration for bias in technologies forming AI models (2) Intervention by decisions made by humans	i. Consideration for bias in data ii. Consideration for bias in algorithms, etc., of AI models	i. Consideration for bias in configurations and data of AI systems and services	i. Consideration for bias in input data or prompt
4) Privacy protection	(1) Protection of privacy across AI systems and services in general	i. Proper data training (Repeat of D-2) i.)	i. Deployment of mechanisms and measures for protecting privacy ii. Countermeasures against privacy violation	i. Countermeasures against inappropriate input of personal data and privacy violation
5) Ensuring security	(1) Security measures relevant to AI systems and services (2) Consideration for the latest trends	i. Deployment of mechanisms for security measures ii. Consideration for the latest trends	i. Deployment of mechanisms for security measures ii. Handling of vulnerabilities	i. Implementation of security measures
6) Transparency	(1) Ensuring verifiability (2) Providing relevant stakeholders with information (3) Reasonable and truthful support (4) Improving explainability and interpretability for relevant stakeholders	i. Ensuring verifiability ii. Providing relevant stakeholders with information	i. Documentation of system architectures and the like ii. Providing relevant stakeholders with information	i. Providing relevant stakeholders with information
7) Accountability	(1) Improving traceability (2) Explanation of conformity to common guiding principles (3) Designation of responsible persons (4) Sharing responsibilities among actors (5) Specific actions for stakeholders (6) Documentation	i. Explanation to AI providers of conformity to common guiding principles ii. Documentation of development-related information	i. Explanation to AI business users of conformity to common guiding principles ii. Documentation of service agreements or the like	i. Explanation to relevant stakeholders ii. Effective use of provided documents and conformity to agreements
8) Education/ literacy	(1) Ensuring AI literacy (2) Education and reskilling (3) Support for stakeholders	-	-	-
9) Ensuring fair competition	-	-	-	-
10) Innovation	(1) Promoting open innovation, etc. (2) Consideration for interconnectivity and interoperability (3) Providing information appropriately	i. Contribution to creation of opportunities for innovation	-	-

Appendix 1. Relevant to Part 1

A. Preconditions for AI

Flow of training and use of AI

In general, to build AI, an AI model is built through a prior training process based on data, and to use AI, the AI model is used to make an inference or prediction and output a result. In addition to conventional AI that uses an AI model that uses specific numeral data, image data, etc., this guideline also covers generative AI that learns a large amount of texts, images, or information posted on the Internet. In some cases, data obtained as outputs is used as inputs for re-training, outputs of an AI model are used as training data for another AI model, or an original AI model is used to create another AI model (see “Figure 2. Examples of flow of training and use of AI”).

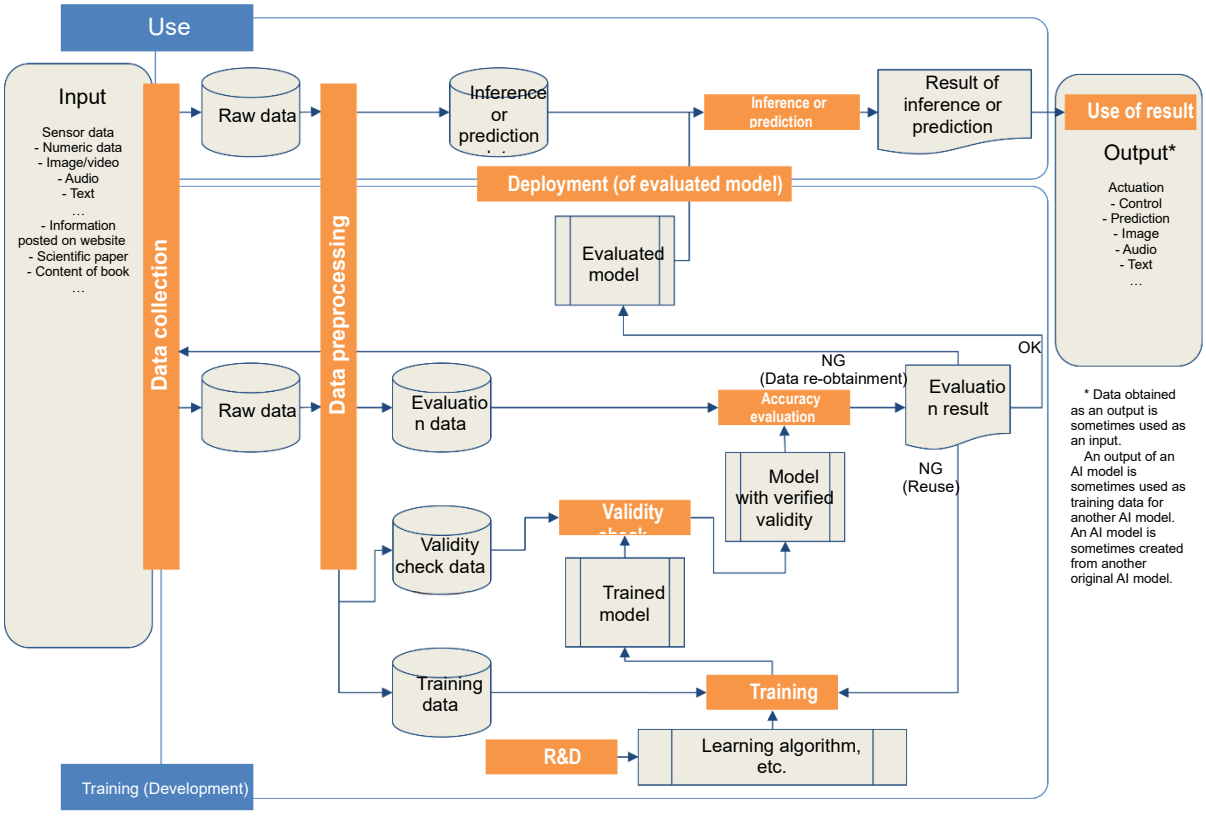
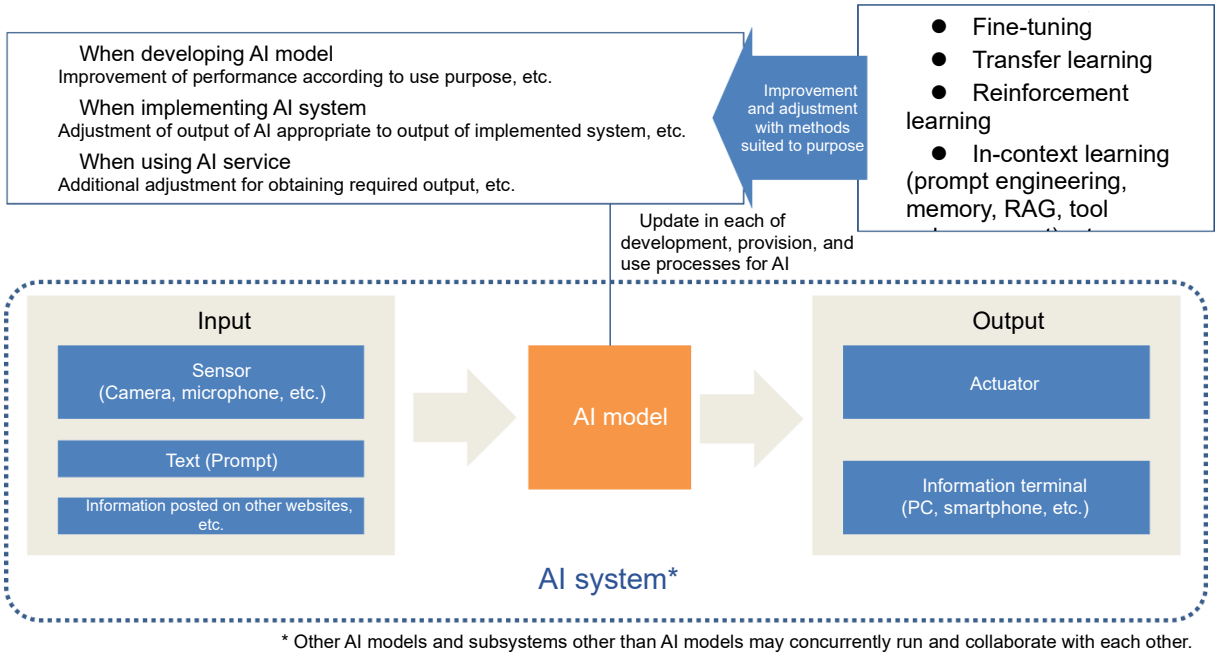


Figure 2. Examples of flow of training and use of AI

Overview of AI system

A system with incorporated software that has AI functions is considered as an AI system. An AI system inputs sensor data, texts, etc., and based on them, outputs via actuators or information terminals. Note that the appendix uses the term "actuator" as a general name for devices that output images, audios, texts, or prediction results as well as driving devices such as a motor and engine and physical devices that carry out control processes through the actions of the driving devices.

In some cases, an AI system is updated through improvement and adjustment in AI development, provision, and use phases via some methods including fine-tuning, transfer learning, reinforcement learning, and in-context learning (prompt engineering, memory, retrieval-augmented generation (RAG[1]), and tool enhancement) (see "Figure 3. Overview of AI system").



* Other AI models and subsystems other than AI models may concurrently run and collaborate with each other.

Figure 3. Overview of AI system

[1] The concept of RAG as defined by Patrick Lewis's "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," involves a language generation model that combines pre-trained parametric memory with non-parametric memory, specifically search-based memory. In a corporate setting, RAG is utilized to enhance the accuracy of generative AI responses by searching internal documents and databases. Additionally, it is employed to search the internet in real-time, allowing the model to provide answers based on the most current data available.

Value chain of AI from development to use

An AI developer builds an AI model using collected data, and an AI provider incorporates the AI model into an existing or new system to build an AI system. The built AI system or an AI service that uses the system is provided to AI business users and used by them (see "Figure 4. Correlation between AI business actor and general AI use flow").[2]

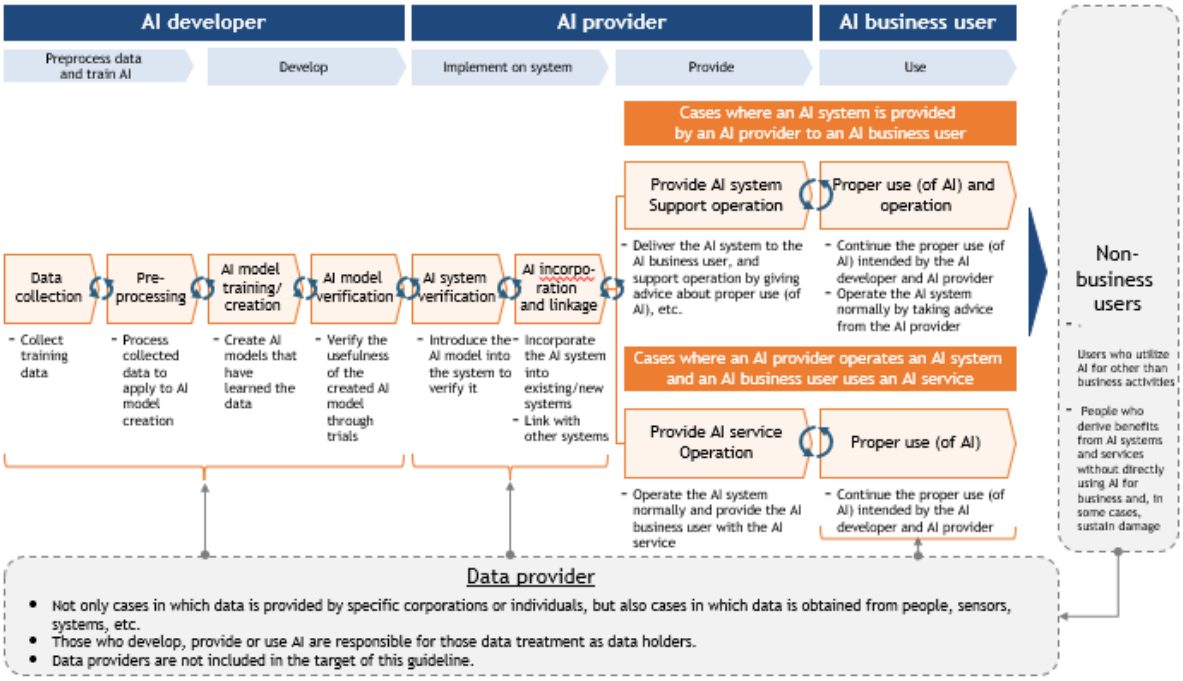


Figure 4. Correlation between AI business actor and general AI use flow

[2] The entity operating the AI system can be either the "provider" or the "user" depending on the form of provision. When AI users receive AI systems from AI providers, the systems become the assets of the AI users, and the ongoing operation for continuous use is organized as the user's role. On the other hand, when AI users receive AI services from AI providers, the AI systems necessary for providing those services remain the assets of the AI providers and are not transferred to the AI users, thus the operation of the AI systems is organized as the provider's role.

Examples of AI systems and services

Typical AI systems and services are shown in “Table 2. Examples of AI systems and services.”

Table 2. Examples of AI systems and services[3]

Case name	Used AI	Overview	AI developer	AI provider	AI business user	Non-business user
Recruitment AI	Text analysis	The recruit department of each foreign subsidiary of Company A Group **uses an AI service that provides reference information for conducting the applicant screening process based on applicants' applications**. The AI development department of Company A has created an AI model that receives past application data and decision of acceptance (judgment on whether to employ each applicant) from the recruit department of Company A (AI business user; including the recruit departments of overseas group companies) and processes them through machine learning (classification model) for supporting in making acceptance decisions.	Company A (Development department)	Company A (System department and human resource development department)	Company A Group (Recruit department)	Applicants for recruitment
Unmanned convenience store	Image analysis	Company J, which holds convenience store franchises across Japan, operates **unmanned convenience stores in which image recognition AI is used. In the unmanned convenience stores, AI calculates the price for items taken by each customer and carries out the payment process for all the items through digital money, etc., when the customer leaves the store**. An AI system for unmanned convenience stores developed by Company X is incorporated into the AI service.	Company X	Company J (AI system development department and convenience store business division)	Convenience stores	Customers of convenience stores
Cancer diagnosis AI	Text and image analyses	Using the multimodal learning, this system **imports “information of the medical history, genes, etc., of a patient (data 1)” and “endoscopic image (data 2)” to highlight areas that are highly possibly affected by cancer in real time during an endoscopic examination**. It enables physicians to observe output images and diagnose potential cancer. Company A has developed AI and provides the cancer diagnosis AI system to health facilities.	Company A (AI development department)	Company A (Healthcare IT service department)	Health facilities (System department and gastroenterology)	Patients examined
Defective detection AI	Image analysis	**This is an quality inspection system for finished goods using deep learning image generation and recognition model.** Conventionally, finished goods (industrial parts) are inspected through visual inspections, requiring considerable labor cost. Therefore, it was decided to incorporate an automatic inspection system for finished goods using deep learning into manufacturing lines. **The system detects appearance defects in finished goods (industrial parts) manufactured in factories of Industry Corporation A**. The number of defects detected in a factory is extremely small compared to the total number of finished goods shipped usually. Therefore, the system uses an AI model that generates images different from finished goods and an AI model that can properly recognize normal goods. The development of the deep learning models was outsourced to Company B.	Company B (Solutions for manufacturers)	Industry Corporation A (Manufacturing management department)	Industry Corporation A (Manufacturing lines in factories)	-

[3] Excerpted from “Risk Chain Model – Posted Case Studies” by the Institute for Future Initiatives, the University of Tokyo. In accordance with the classification of AI business actors in this guideline, the columns for the AI developer, AI provider, AI business user, and non-business user have been added.

Overhead power line inspection AI	Image analysis	**This is a diagnostic service for overhead power lines using an image analysis technology through deep learning. It analyzes images to inspect overhead power lines and detect abnormal positions automatically**. Usually, maintenance personnel inspect overhead power lines through visual inspections using high-powered telescopes. Overhead power lines in an environment where it is difficult to conduct visual inspections such as a mountain area, however, it is necessary to record videos from a helicopter and have them visually inspected by experienced maintenance personnel through a slow playback, taking a long time. Against this background, Company P decided to introduce image recognition AI of Company X to automate the detection of abnormal positions of overhead power lines and the report creation. Videos are recorded using a drone or helicopter. Although videos are not inspected in real time, image recognition AI swiftly detects abnormal positions and creates a report after a video recording work is complete.	Company X (AI development department)	Company P (System department and power service maintenance department)	Company P (Maintenance personnel)	-
Smart home appliance optimization AI	Sensor data analysis	**An AI model optimizes smart home appliances by analyzing environment information, user behaviors, etc. Company A's AI service** collects data of sensors installed by the user (location and status of the user, temperature, humidity, illuminance, and CO_2 level), open data (weather information), and feedbacks from the user (opinions about stresses, comfortableness, etc.), analyzes them using an AI model, and automatically controls smart home appliances (smart refrigerator (food management, recipe recommendation, etc.), air conditioning, underfloor heating, air purifier, robotic vacuum cleaner, ventilation system, etc.). AI provider is the Company A (appliance business division) , but it may be a reseller and is thus expected to act according to what was explained to consumers.	Company A (AI development department)	Company A (Appliance business division)	-	Consumers
In-house introduction of dialogue-type AI	Generation of texts, etc.	**This is an in-house AI assistance service using generative AI.** Employees of Company B can get answers by entering prompts (directions or questions) in dialogue-type AI. It is used for every purpose and usage in in-house business operations including questions, programming, document generation, translation, and summarization, contributing to the improvement of productivity in business operations. Company C, which is a group company of Company B, has implemented the AI assistant service using Company A's cloud platform and generative AI model and provides it to the employees of the Company B Group (including Company C).	Company A	Company C which is a group company of Company B	Employees of Company B Group (including Company C)	-

Patterns of AI companies

There are three patterns of value chains of AI used in business: pattern 1 in which AI business users use AI offering benefits to AI business users and also non-business users,[4] pattern 2 in which AI business users use AI and gain benefits, and pattern 3 in which non-business users use an AI system or service provided by an AI provider and gain benefits (see "Figure 5. Patterns of AI companies").

In pattern 1, only benefits are provided to non-business users, without providing AI systems (services) to them.

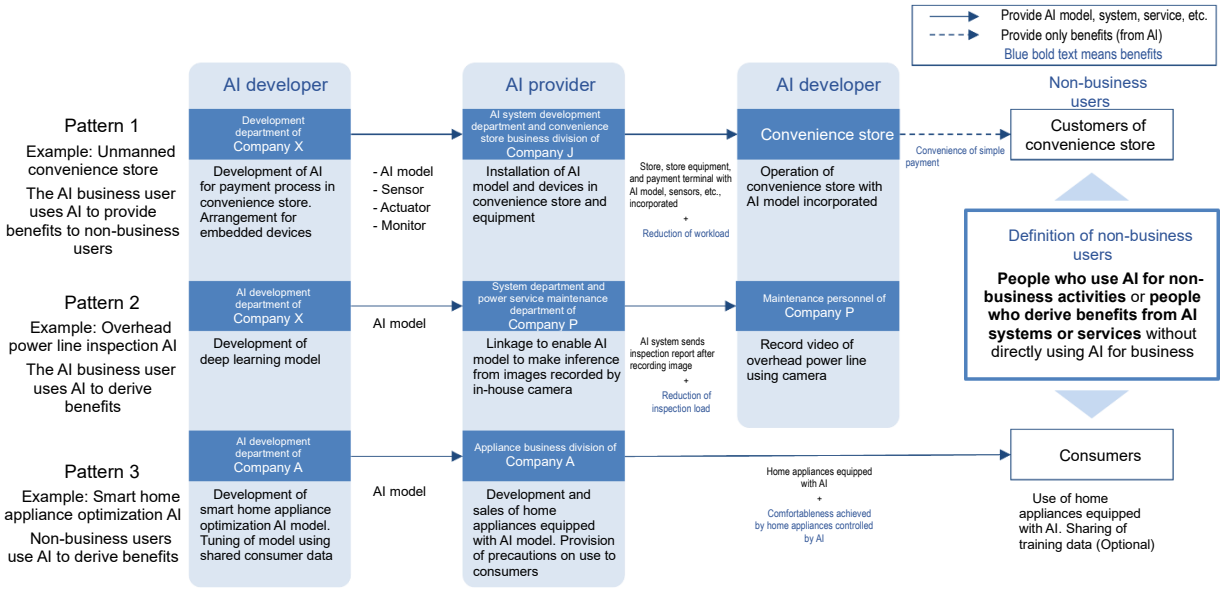


Figure 5. Patterns of AI companies

[4] Those who use AI for non-business activities or those who derive benefits from AI systems and services without directly using AI for business and, in some cases, sustain damage (defined in the main part of this guideline).

About data provider

In the AI development, provision, and use phases, data is used for training AI models and using AI. In some cases, when building or using an AI model using data, an AI developer, AI provider, or AI business user uses their own data without using external data. In other cases, they use data provided by a specific company or individual or data obtained from sources such as specific groups, sensors, systems and so on. Although it is not feasible to describe all process patterns to treat such data on this guideline, only those concerning AI developer, AI provider and AI user who are supposed to be provided or received data are described. (see “Figure 6. Concepts of data provision”). However, when providing or receiving data to or from a specific company or individual, it is important to refer to Appendix 6 and the “Contract Guidelines on Utilization of AI and Data” mentioned in it to conclude an agreement and contract between the party who is provided with data and the party who provides data (data provider) before using data.

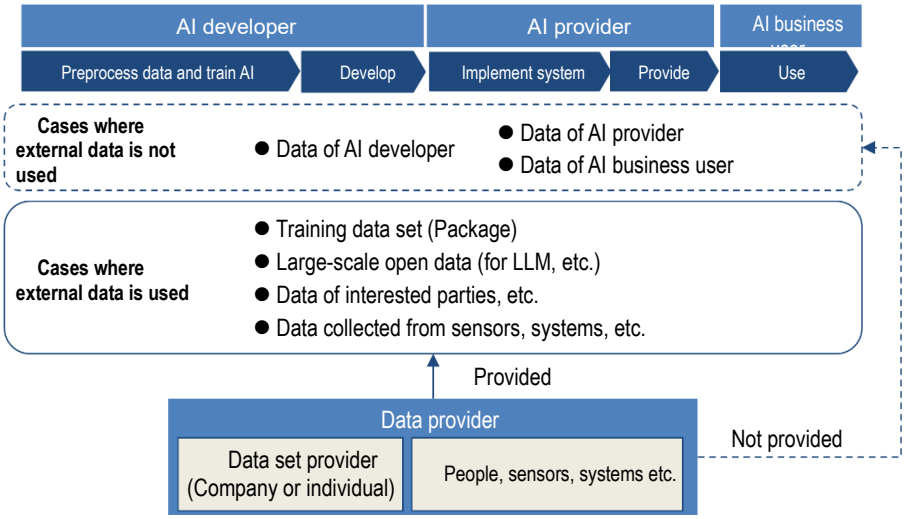


Figure 6. Concepts of data provision

B. AI's benefits and risks

AI brings benefits such as creating new business, adding values to existing business, and improving productivity, whereas it also carries risks.

It is expected to reduce those risks as much as possible. On the other hand, taking too many measures against risks might decrease the benefits of using AI due to increased costs, etc. Therefore, the concept of the risk-based approach, in which the degree of measures against risks is proportionate to the characteristics and probability of risks, is important.

AI's benefits

There are a variety of benefits from AI use, and they have been enhancing as technologies are advanced.

AI business actors can use AI to create value. The following can be expected as the result of AI use:

- Reduction of operation costs
- Creation of new products and services that accelerate innovation in existing business
- Renovation of organization

Furthermore, it is conceivable that AI is applied to various fields (agriculture, education, healthcare, manufacturing, transportation, etc.) and various deployment models (cloud service, on-premise system, cyber-physical system, etc.) are used.[5]

Examples of benefits

The following "Figure 7. Examples of benefits from AI to corporate activities" shows just a small part of examples of benefits from AI to corporate activities. AI can bring about effects to the overall corporate activities.

	Development	Marketing	Sales	Logistics/distributio	Customer support	Legal	Finance	HR
Examples of benefits available conventionally (Improved by generative AI)	Automation of code verification and documentation	Automatic distribution of ad emails	Support after order intake Automatic transmission of emails, etc.	Optimization of production and inventory based on demand prediction	Automatic response through chat bot	Translation	Automatic creation of financial statements	Automation of payroll calculation, etc.
	Extraction and verification of similar code and data	Personalized ad based on data	Sales prediction for each channel and need	Optimization of delivery route	Creation of FAQ based on past inquiries	Review of legal text	Future prediction based on past records, and detection of malpractices	Human resources demand matching based on résumés, etc.
Examples of benefits unique to generative AI	Generation of training data, coding assistant, brainstorming for new products	Automatic creation of sales promotion (marketing materials, sales	Automatic creation of sales talk script	Assistant for negotiation for logistics conditions	Automatic generation and summarization of transcription of support	Automatic generation of draft of contract based on stipulations	Response to in-house inquiry according to context	Holding human resources interview according to context

Figure 7. Examples of benefits from AI to corporate activities

For example, in the logistics field, AI is used to automate the distribution with robots and optimize the value chain through the demand prediction. In the human resources field, AI is used for improving efficiency using data such as the automation of the payroll calculation and the human resources demand matching based on résumés. AI is used for various use purposes to improve efficiency in business operations and optimize them.

In fields other than companies, AI is also used for the automation of administrative procedures, farm work aid systems using sensors and image data, and the utilization of medical histories, etc., in healthcare field.

[5] "ISO/IEC TR 24030" contains a wide range of use cases that cover those fields and deployment models.

In the B2C field, various services have been deployed including chatbots, self-driving, search systems, and voice assistants.

Potential of generative AI

In addition to the circumstances described above, generative AI has emerged recently. There is a high possibility that generative AI might trigger Japanese companies, which are left behind in DX, to regain momentum.

The advantages of Japanese companies include accumulated operational technology (OT) data of good quality and careful and courteous services and works. When you try to implement them using conventional AI, a lot of time and special knowledge are required, for example, the use of OT data in a cross-organization and cross-industry manner, integration of data interfaces for using AI for those services and works, preparation of a large amount of data, creation of scenarios and cases assuming a lot of patterns, and development based on them. The use of generative AI for those works enables to automatically create scenarios and cases (self-supervised learning), promoting AI use in a wide range of companies. There is an actual case where generative AI creates replies and materials for call centers and sales supports of retail companies to improve productivity. In addition, it is feasible that a system creates multiple patterns of replies and materials for entered inquires and requests from customers by referring to in-house data.

Multimodal generative AI, which can collect and integrate information from two or more different modalities such as text, audio, images, video, and sensor data, has been emerging. The advent of multimodal generative AI is expected to expand the range of AI applications in fields such as healthcare, drug discovery, education, and entertainment, as well as improve processing capabilities in general tasks like inference and analysis beyond just generation.

Additionally, the use of Retrieval-Augmented Generation (RAG) is expanding. By combining language generation with external information retrieval, RAG can suppress hallucinations, specify information sources for searches, and clearly indicate references in output texts, thereby enhancing transparency in the output process and rationale. Unlike regular fine-tuning, it allows for the addition of data sources without retraining the model, which is expected to reduce costs.

The application of generative AI in program code generation has also been advancing. It is expected to enable low-cost and rapid code generation, avoid human errors, and allow programming without advanced skills or knowledge.

Furthermore, autonomous AI systems, known as AI agents, have also been emerging. Compared to traditional AI and generative AI, AI agents are expected to offer greater efficiency and automation, leading to increased productivity.

To become a winner of global competition, it is expected to correctly understand derivable benefits, explore the potential, and keep active commitment, for example, changing the digital strategy by actively adopting generative AI.

AI's risks

Whereas the benefits are increasing, as AI use becomes widespread and new technologies emerge, the risks they incur are also enhancing. As generative AI becomes widespread, in particular, risks are diversified and increased such as the generation and distribution of disinformation and misinformation, and demands to respect intellectual property rights are increased.

Specifically, the issues shown below have come up.[6] Note that the risks mentioned below are just typical ones and do not include every risk in AI, and some of them are based on assumptions. Therefore, it is expected to consider them as mere examples. Hence, the development, provision, and use of AI should not be inhibited even if there are risks mentioned below.[7] On the

[6] For overseas case studies, "The AI Incident Database (AIID)" of Partnership on AI (http://incidentdatabase.ai) is a good reference with more than 2,000 reports posted. For details, see "Column 1: Sharing information on incidents" provided later. Incidentally, items in the parentheses for each issue indicate the corresponding common guiding principles described in the main part.

[7] Foreign laws and regulations should be obeyed as well. For example, the " Artificial Intelligence Act" of EU, enacted on May 21, 2024, defines AI systems that can be considered to directly pose a threat to the lives of humans and basic human rights (for example, subliminal manipulations (excluding ones used for remedies)), social scoring by the government, and voice assistance that encourage dangerous behavior, as "Prohibited AI" with unacceptable risks, prohibiting their placing on the market, implementation, and use. For other countries' AI-related laws and policies, OECD's "National AI policies & strategies" at https://oecd.ai/en/dashboards/overview is a useful reference.

contrary, it is expected to lead to strengthening of competitiveness, creation of value, and innovation through the active development, provision, and use of AI by recognizing the risks and considering the tolerance for risks and balance between benefits and risks.

Note that, in addition to disadvantages for business operators, risks to stakeholders[8] or the whole society are also studied. In the following, risks that are difficult for businesses themselves to address alone, requiring responses and discussions from society as a whole, including governmental and public institutions, are also described with reference to discussions in other countries. Therefore, it is not necessary for businesses to immediately address all risks, but it is important to recognize them as potential occurrences in society.

Attacks on AI systems such as data poisoning attack

- During the AI training, there is a risk of intrusion of invalid data into training data, causing performance degradation and misclassification. During the service operation, there is a risk of cyberattacks that aim at the application itself and a risk of attacks through AI inference results or prompts that are directions to AI. A chat bot, for example, was trained by a malicious group using racist questions in an organized manner, and as a result, it got to repeat hate speech.[9]
- There is a risk of RAG being misused by malicious third parties, through such as indirect prompt injection or malware generation.

Biased outputs, discriminatory outputs, and inconsistent outputs

- An IT company developed an AI human resources recruitment system by itself, but it was revealed that the system had a defect in machine learning that discriminated against women. For the training of the AI system, résumés of the applicants over the past 10 years were used. However, because almost all the applicants were men, so it was said that the AI decided that recruiting men was preferable. The company attempted to correct the program so that it did not discriminate women, but it decided to stop using the system because other discriminations could be produced.

Incorrect outputs due to Hallucinations and similar issues

- As for a hallucination which is a response given by generative AI which contains disinformation or misinformation presented as a fact, a lawsuit was filed against an AI developer and provider. A cast member of a TV program found disinformation being spread by generative AI stating that the cast member was sued for embezzlement. The generative AI even forged a fake complaint. The cast member sued the companies that developed and provided the generative AI for defamation.

Black-boxing and inadequate explanations of decisions

- Black-boxed AI's judgments caused a problem as well. As for a credit card, some reports were posted on SNS stating that the credit limit for a woman was less than that of a man with the same annual income. On this problem, the financial authority initiated an investigation and demand the credit card company to prove the validity of the algorithm. However, the company could not explain about the specific function and behavior of the algorithm.
- In the case of AI systems with complex mechanisms, such as multimodal generative AI, the difficulty of maintenance and troubleshooting may increase compared to regular AI systems.

Inappropriate use of personal data

[8] All the AI business actors who might be directly or indirectly affected by AI use including third parties other than AI developers, AI providers, AI business users, and non-business users (the same shall apply hereafter).
[9] A report targeting the U.S. on the types of security threats and risks to AI systems and services has been published. IPA, "Recognition Survey Report on AI Security Threats and Risks in the U.S." (May 2024)
https://www.ipa.go.jp/security/reports/technicalwatch/20240530.html

- There have been instances where a service using AI for human resources was abolished, as a result of the use of personal information lacking transparency being criticized. When the AI provided the information about the possibility that an applicant might withdraw his/her application and decline the unofficial job offer, no clear explanation was given to applicants such as students. Furthermore, the terms of use did not include stipulations about the provision of information to third parties under an agreement for a period of time.

Occurrence of accidents related to lives, etc.

- If AI makes an inappropriate judgment, for example, a self-driving car might cause an accident, seriously damaging lives and properties. In such a scenario, some people are concerned about a great risk of accident caused by a malfunction of AI. [10]
- In cases where generative AI is used to generate program code for machines and other equipment, there is a concern that incorrect or inefficient code could lead to performance degradation or accidents.

Discrimination in triage

- In triage by which individuals are prioritized upon an incident, if AI has an bias for determining the prioritization, fairness might be lost. When AI is used for triage in a healthcare scene, healthcare judgments with discrimination against specific human groups might be made, posing a threat to lives.

Excessive dependence

- There have been instances where companies face accountability issues or criticism due to inappropriate and excessive use of AI, such as relying solely on AI decisions in important decision-making scenarios like recruitment activities. Additionally, there have been reports of users becoming psychologically dependent on chatbot services using generative AI.

- Misuse The use of AI for frauds is also perceived as a problem. Among them, frauds using speech synthesized by AI are rapidly increasing. A woman got a call via which her daughter's voice asked for help demanding a ransom of a million dollars, though it was revealed that the voice was synthesized with AI and the call was a fraud imitating a kidnapping.

Infringement of intellectual property rights, etc.

- Some stakeholders have taken up the handling of intellectual property rights during the use of generative AI for discussion. In a foreign country, multiple artists filed a class-action lawsuit arguing that AI sometimes generated images similar to artists' works used for training the AI. [11]

---

[10] Taking into account these possibilities, the Digital Agency has established the "Sub-Working Group on Review of Social Rules for automated driving vehicles in age of AI" to enhance safety by ensuring AI makes more appropriate decisions in the social implementation of autonomous driving. This sub-working group proposes making systems which collect and share not only accident data but also other driving data (such as near-misses) among stakeholders for analysis. It also proposes clarifying and specifying rules to contribute to more appropriate program creation. Relevant ministries and agencies have been advancing their considerations based on these proposals.
https://www.digital.go.jp/councils/mobility-subworking-group

[11] In Japan, as stipulated by Article 30-4 of the Copyright Act, at the training and development stage, a copyrighted work can be used, to the extent considered necessary, without permission of the copyright holder, when the copyrighted work is used without the purpose to enjoy or to make others enjoy the ideas or sentiments expressed in the work, for example, for information analysis. On the other hand, at the generation and use stage, the legality is judged based on the dependency and similarity in the same way as for usual copyright infringement, except for cases where the Copyright Act permits the use of a copyrighted work. And discussions about AI and patent related laws are ongoing in some government offices including the Cabinet Office and the Agency for Cultural Affairs, and it is important to watch their status. Specifically, regarding the perspective on AI and copyright, it has been compiled by the Subcommittee on Legal Systems of the Copyright Subcommittee

- In addition to images, there is a possibility that outputs generated by generative AI, such as text, may infringe on others' intellectual property rights. For example, in cases where generative AI is used to generate program code for machines and other equipment, it is necessary to be mindful of the possibility that the generated code may infringe on others' intellectual property rights.

Financial Loss

- There may be instances where companies get held financially liable, through such as claims for damages, if the output of their AI systems or services significantly infringes on the rights of others.

Leak of confidential information

- During the use of AI, there is a risk that personal data and confidential information is entered as a prompt and leaks through an output from the AI, etc. For example, a case was revealed, in which an employee used an AI service for a business operation and entered source code that was confidential information in dialogue-type generative AI intended for non-business users. Generative AI services are easy to use, and especially if the company has not established rules and regulations, employees might use generative AI intended for non-business users during business operations in a risky way outside management by the company. When utilizing RAG and other external services or data, it is particularly important to be cautious about the unintended leakage of important information (such as personal or confidential information). Additionally, if program code generated using generative AI contains security vulnerabilities, there is a risk of information tampering or leakage. Note that there is dialogue-type generative AI with enterprise-grade security functions incorporated intended for the business use. It is recommended that companies use such services or applications especially when processing confidential information.

Unemployment of workers

- New technologies such as generative AI and AI are assumed to change the nature of tasks and alter the roles of workers. While the introduction of generative AI and AI is anticipated to reduce the workload of workers and improve labor productivity, there are also concerns about unemployment risks and widening disparities.[12]

Concentration of data and profits

---

of the Council for Cultural Affairs. Additionally, in the "Interim Report of the Study Group on Intellectual Property Rights in the AI Era" and the "Checklist & Guidance on AI and Copyright," unlike these guidelines, the "AI developers," "AI providers," and "AI users," as well as "non-business users (general users)" and "rights holders," are also considered, and examples of expected initiatives for each entity are organized. It is crucial that each AI business actor consider response policies based on those discussion contents.

・Agency for Cultural Affairs, "On the Perspective of AI and Copyright" (Subcommittee on Legal Systems of the Copyright Subcommittee, Council for Cultural Affairs, March 2024)
https://www.bunka.go.jp/seisaku/bunkashingikai/chosakuken/pdf/94037901_01.pdf

・Cabinet Office, "Interim Report of the Study Group on Intellectual Property Rights in the AI Era" (Intellectual Property Strategy Promotion Bureau, May 2024)
https://www.kantei.go.jp/jp/singi/titeki2/chitekizaisan2024/0528_ai.pdf

・Agency for Cultural Affairs, "Checklist & Guidance on AI and Copyright" (Copyright Division, Agency for Cultural Affairs, July 2024)
https://www.bunka.go.jp/seisaku/chosakuken/pdf/94097701_01.pdf

・Cabinet Office, "Interim Report of the Study Group on Intellectual Property Rights in the AI Era - Guide (for Rights Holders)" (Intellectual Property Strategy Promotion Bureau, November 2024)
https://www.kantei.go.jp/jp/singi/titeki2/chitekizaisan2024/2411_tebiki.pdf

[12] Quoted from the Ministry of Health, Labour and Welfare "The Employment Policy Study Group Report" (The Employment Policy Study Group, August 2024)
https://www.mhlw.go.jp/content/11602000/001294201.pdf

- Challenges have been pointed out, such as the concentration of data and profits among a few AI developers, and the lack of high-performance AI in minority language countries for their own languages.[13]

Infringement of qualifications, etc.

- There might be risks of infringement of legally prescribed licenses and qualifications caused by the use of generative AI. When generative AI answers to legal or healthcare questions, it might infringe legally prescribed licenses and qualifications, posing a legal problem. However, if an industry tries to avoid such risks, the introduction of generative AI may delay in the whole industry, restricting new services and improvement of efficiency.

Distribution and diffusion of disinformation and misinformation

- Blindly trusting misinformation produced by generative AI can be a risk. A lawyer in the US, for example, used generative AI for creating materials for a civil lawsuit on trial, quoting a precedent that did not exist in fact, causing a problem.
- Misuses of deepfakes have been committed one after another in various countries. In foreign countries, information manipulations and public opinion manipulations have been performed using fake images and videos. In a case, a fake image that showed an explosion near the Pentagon was produced by generative AI and instantly spread across SNS and the Internet. Fake accounts that posed as some overseas mass media and major financial media also spread this information, causing the average stock price to fall by more than 100 dollars temporarily. In some cases, company accounts erroneously spread disinformation of an incident, accident, natural disaster, etc.

Negative influence on democracy

- In various countries, the political use of personal data is also perceived as a problem. In an election campaign, for example, personal data was collected through profile information and a personality assessment app provided to non-business users of SNS. The personal data was used to grasp the personality of each individual and place a targeted ad that appealed to the personality so that the individual voted in favor of the client. Specifically, based on the collected data, individuals were classified into some groups such as a group who had an impulsive temper and tended to believe a conspiracy theory compared to average citizens and a group who had a neurosis and dark triad characteristics, and a large number of articles advantageous to the candidate's campaign were posted. Some people were concerned that this action intervened in an election campaign using personal data, undermining democracy which was the essence of a country.
- There have been instances of generating and disseminating disinformation, misinformation and deepfakes about other candidates during elections in various countries.

Filter bubble and echo chamber phenomena

- The social division caused by recommendations given by SNS, etc., is perceived as a problem. Some people are concerned that AI business users and non-business users may foster a tendency to think in extremes due to some phenomena, such as a filter bubble in which a person is surrounded by his/her favorite information only and an echo chamber in which only the same ideas as of a person are returned from the surroundings.

Loss of diversity and inclusion

[13] Quoted from the Cabinet Office "Policy on AI Systems"(AI Strategic Council, May 2024)
https://www8.cao.go.jp/cstp/ai/ai_senryaku/9kai/shiryo2-1.pdf

- If the whole society uses the same model in the same way, the derived opinions and replies might converge through LLM, losing the diversity. When utilizing RAG, while benefits such as improved response quality can be obtained, there is a possibility that the convergence of responses generated by AI will accelerate, further reducing diversity. Additionally, in financial transactions, the use of common algorithms and the automation and standardization of risk assessment and judgment may increase the instability of financial markets.

Reproduction of bias

- Because generative AI creates answers based on existing information, if those answers continue to be blindly trusted, biases contained in existing information might be amplified, continuing and enhancing unfair outputs containing discrimination. For example, when answers are created based on data in which gender discrimination is included and an increasing number of people believe the answers, the risk of fixing gender discrimination is increased.

Energy consumption and environmental load

- As the use of AI is spreading, the demands for calculation resources are also increasing. As a result, data centers are enhanced, and some people are concerned about the increase of the energy consumption by them. In models with high computational complexity, such as multimodal generative AI, the impact on the environment can be significant. Some people point out that the carbon dioxide emission caused by the large amount of power used for AI development is several tens of times greater than the carbon dioxide emission in the US per person and per year.[14] However, it should not be forgotten that AI has a potential for contributing to the environment, for example, introducing AI to energy management enables to use power more effectively.

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

To enable businesses to comprehensively understand and consider countermeasures for these risks as much as possible, the risks have been systematically classified.[15] (Table 3. Systematic Classification of AI-related Risk Examples（tentative ver.）)

Table 3. Systematic Classification of AI-related Risk Examples（tentative ver.）

- The table below does not cover all AI risks and includes hypothetical cases, and it is expected to be recognized as just one example.
- The table below also includes risks that require responses and discussions from society as a whole, including governmental and public institutions.

Major categories	Subcategories	Risk examples
Technical Risks (=risks primarily associated with AI systems)	Risks during the learning and input stages of AI	Attacks on AI systems such as data poisoning attacks
	Risks during the output stage of AI	Biased outputs, discriminatory outputs, and inconsistent outputs
		Incorrect outputs due to Hallucinations and similar issues
	Risks during the post-response stage	Black-boxing and inadequate explanations of decisions
Societal Risks (=existing risks that may also arise in AI or be amplified by AI)	Risks related to ethics and law	Inappropriate use of personal information
		Occurrence of accidents related to lives, etc.
		Discrimination in triage
		Excessive dependence
		Misuse
	Risks related to economic activities	Infringement of intellectual property rights, etc.
		Financial loss
		Leak of confidential information
		Unemployment of workers
		Concentration of data and profits
		Infringement of qualifications, etc.
	Risks related to the information space	Distribution and diffusion of disinformation
		Negative influence on democracy
		Filter bubble and echo chamber phenomena
		Loss of diversity and inclusion
		Reproduction of biases
	Risks related to the environment	Energy consumption and environmental load

To further connect the identified risks to the consideration of countermeasures by businesses, the main common guidelines corresponding to each risk and examples of countermeasures by businesses have been described. (In addition to the common guidelines listed, there may be other relevant common guidelines.) (Table 4. Mapping of AI's Risk Examples, Common Guiding Principles, and Important Matters for Each Business Actor) It is advisable to refer to the relevant sections of Parts 3 to 5 of the main text and Appendices 3 to 5 for measures and specific methods for each entity.

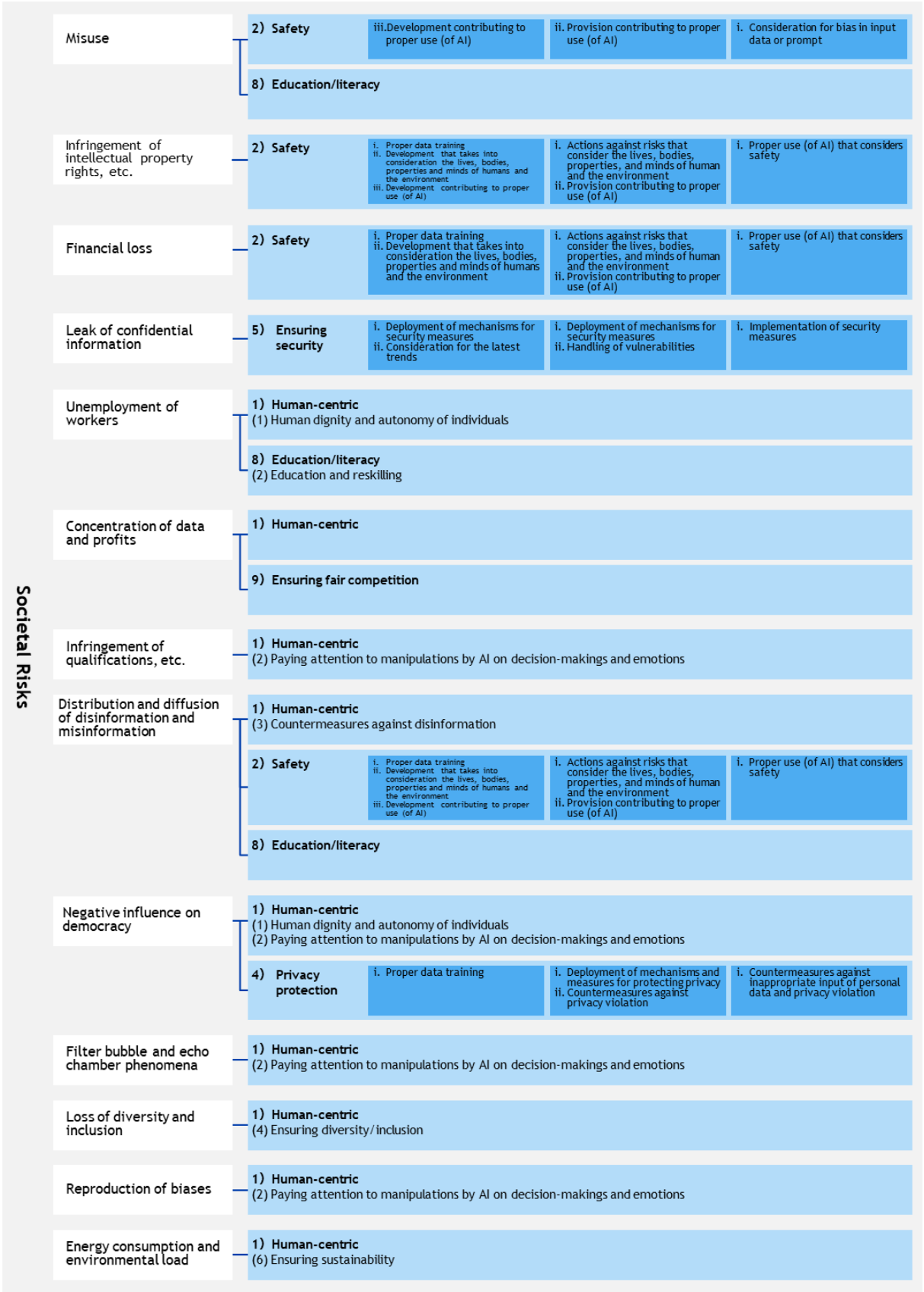
[15] In considering the classification framework, various classification methods were investigated and analyzed both domestically and internationally. For example, classifications were conducted with reference to the NIST AI Risk Management Framework (AI RMF 1.0) and Hiroki Habuka's "Introduction to AI Governance."

Table 4. Mapping of AI's Risk Examples, Common Guiding Principles, and Important Matters for Each AI Business Actor

- The table below does not cover all AI risks and includes hypothetical cases, and it is expected to be recognized as just one example.
- The table below also includes risks that require responses and discussions from society as a whole, including governmental and public institutions.

	Risk Examples	Key common guiding principles corresponding to each risk example	Important matters for each AI business actor in addition to common guiding principles		
			Part 3. AI Developer	Part 4. AI Provider	Part 5. AI Business User
Technical Risks	Attacks on AI systems such as data poisoning attacks	5) Ensuring security	i. Deployment of mechanisms for security measures ii. Consideration for the latest trends	i. Deployment of mechanisms for security measures ii. Handling of vulnerabilities	i. Implementation of security measures
	Biased outputs, discriminatory outputs, and inconsistent outputs	1) Human-centric (1) Human dignity and autonomy of individuals (3) Countermeasures against disinformation			
	Incorrect outputs due to Hallucinations and similar issues	2) Safety	i. Proper data training ii. Development that takes into consideration the lives, bodies, properties and minds of humans and the environment iii. Development contributing to proper use (of AI)	i. Actions against risks that consider the lives, bodies, properties, and minds of human and the environment ii. Provision contributing to proper use (of AI)	i. Proper use (of AI) that considers safety
		3) Fairness	i. Consideration for bias in data ii. Consideration for bias in algorithms, etc., of AI models	i. Consideration for bias in configurations and data of AI systems and services	i. Consideration for bias in input data or prompt
		8) Education/literacy			
	Black-boxing and inadequate explanations of decisions	6) Transparency	i. Ensuring verifiability ii. Providing relevant stakeholders with information	i. Documentation of system architectures and the like ii. Providing relevant stakeholders with information	i. Providing relevant stakeholders with information
		7) Accountability	i. Explanation to AI providers of conformity to common guiding principles ii. Documentation of development-related information	i. Explanation to AI business users of conformity to common guiding principles ii. Documentation of service agreements or the like	i. Explanation to relevant stakeholders ii. Effective use of provided documents and conformity to agreements
Societal Risks	Inappropriate use of personal information	1) Human-centric (1) Human dignity and autonomy of individuals (2) Paying attention to manipulations by AI on decision-makings and emotions			
		4) Privacy protection	i. Proper data training	i. Deployment of mechanisms and measures for protecting privacy ii. Countermeasures against privacy violation	i. Countermeasures against inappropriate input of personal data and privacy violation
	Occurrence of accidents related to lives, etc.	2) Safety	i. Proper data training ii. Development that takes into consideration the lives, bodies, properties and minds of humans and the environment iii. Development contributing to proper use (of AI)	i. Actions against risks that consider the lives, bodies, properties, and minds of human and the environment ii. Provision contributing to proper use (of AI)	i. Proper use (of AI) that considers safety
	Discrimination in triage	1) Human-centric (1) Human dignity and autonomy of individuals (2) Paying attention to manipulations by AI on decision-makings and emotions			
	Excessive dependence	2) Safety	i. Proper data training ii. Development that takes into consideration the lives, bodies, properties and minds of humans and the environment iii. Development contributing to proper use (of AI)	i. Actions against risks that consider the lives, bodies, properties, and minds of human and the environment ii. Provision contributing to proper use (of AI)	i. Proper use (of AI) that considers safety
		3) Fairness	i. Consideration for bias in data ii. Consideration for bias in algorithms, etc., of AI models	i. Consideration for bias in configurations and data of AI systems and services	i. Consideration for bias in input data or prompt

Appendix 1. Relevant to Part 1
AI's risks



As described above, the benefits from AI use have been increasing through technological advancements, whereas the risks that have been posed by conventional AI have also been increasing further due to emergence of generative AI. Some new risks have been realized by generative AI as well. In addition, because a lot of generative AI services are easy to use, they might be used in a way that incurs unexpected risks.[16]

Generative AI is rapidly advancing, whereas the technologies and ideas for eliminating the risks are progressing on a daily basis. However, the intrinsic risks in generative AI greatly depend on its technological characteristics. To prevent continuing abstract discussions, it is important to devise effective AI governance as ideas for better use when planning measures.

The risks in generative AI change in accordance with the external environment and technological trends. Because there is no reproducibility, it is difficult to identify the cause of an error. Therefore, the social and technological standardization, validity of tests, establishment of the feedback loop, and redefinition of legal risks and human right risks are necessary. It is important to preserve proper evidences suited for the context as well.

Necessity to build AI governance has become greater, to derive benefits from AI, reduce risks, and strengthen competitiveness through the business use of AI.

Note that being concerned about risks too much is also a kind of risk, because doing so will keep AI business actors stopping without doing anything, choosing to stop using AI until every risk is eliminated, or using a perfect safeguard.

[16] The emergence of AI agents necessitates attention to the potential for increased complexity and severity of risks related to safety, such as accidents, excessive dependence, and unemployment of workers.

Appendix 2. “Section 2. E. Building AI Governance”

As described in Appendix 1. B. “AI’s benefits and risks,” it is important to build AI governance for managing risks of AI at the level acceptable to stakeholders and maximizing benefits offered by AI in order to receive benefits and control risks of AI. In doing so, AI business actors are expected to apply appropriate solutions in light of the continuously changing environment and goals, and evaluate and review if they are working appropriately on a continuing basis.[17]

The behavioral goals, practice guidelines and practical examples, which AI business actors should pay attention to when building AI governance, are as follows:

The behavioral goals described above are the general and objective goals. It is important for all AI business actors involved in the development, provision and use of AI systems and services that may cause certain risks to the society to meet such behavioral goals (see “Table 5. List of Behavioral Goals” for the complete picture). On the other hand, as for the practice guidelines and practical examples intended for virtual companies, useful elements will vary depending on the unique and specific situations in which AI business actors are placed, and purpose, method and evaluation target of an AI system and services developed, provided and used by AI business actors. For this reason, whether or not the practice guidelines and practical examples are adopted will be left up to each AI business actor. Even when they are adopted, it is expected to consider modification and selection based on the circumstances of AI business actors.

It is expected to build and operate AI governance regimes according to the requests from stakeholders through collaboration between IT, privacy and security governance, etc. in AI business actors and collaboration between AI business actors across the entire value chain. It is also important to review the systems, rules and organizations to minimize the management workloads and speed up the decision-making process and operation based on the agile governance concept when building AI governance. In addition, it is expected to optimize AI governance and management and use limited resources efficiently.

[17] As a basic concept for evaluating AI safety, the AI Safety Institute (AISI) has published the “Guide on Evaluation Perspectives for AI Safety,” which can be referenced by AI system developers and providers when conducting AI safety evaluations. Additionally, the institute has introduced the “Red Teaming Method” as one approach for AI safety evaluation.
AISI, “Guide on Evaluation Perspectives for AI Safety” (September 2024)
https://aisi.go.jp/effort/effort_framework/guide_to_evaluation_perspective_on_ai_safety/
AISI, “Guide on Red Teaming Method for AI Safety” (September 2024)
https://aisi.go.jp/effort/effort_framework/guide_to_red_teaming_methodology_on_ai_safety/

Table 5. List of Behavioral Goals

Classification	Behavioral Goals
1. Environment and risk analyses	1-1 Understanding benefits/risks 1-2 Understanding social acceptance of AI 1-3 Understanding Company’s AI proficiency
2. Goal setting	2-1 Setting AI governance goals
3. System design	3-1 Requiring evaluation of deviation from goals and measures to minimize the deviation 3-2 Improving literacy of those in charge of AI management system 3-3 Enhancing AI management through cooperation with each other between AI business actors and divisions 3-4 Reducing burden related to incidents involving users through preventive and prompt action
4. Operation	4-1 Ensuring that the operation of AI management system is explainable 4-2 Ensuring that the operation of each AI system is explainable 4-3 Considering proactive disclosure of AI governance practices
5. Evaluation	5-1 Verifying AI management system functions 5-2 Considering opinions of outside stakeholders
6. Environment and risk reanalysis	6-1 Reimplementing behavioral goals 1-1 to 1-3 at an appropriate time

A. Building of AI governance and monitoring by management

1. Environment and risk analyses

Behavioral goal 1-1 [Understanding benefits/risks] AI business actors will, under the leadership of the management team, clarify the purpose of development, provision and use of AI, specifically understand that there are not only benefits offered by AI but also unintended risks, report them to the management, share with the management and update the understanding at an appropriate time.

[Practice Guidelines]

AI business actors will, under the leadership of the management team, take the following[18] measures:

- Define the purpose of development, provision and use of AI such as creation of value in business and solution of social issues.
- Specifically understand the benefits and risks including unintended risks in a way that is relevant to a company’s own business.
- In doing so, pay attention to risks that should be avoided and topics span across several AI business actors, secure the benefits and reduce risks in the entire value chain / risk chain.
- Establish a system for promptly reporting to / sharing with the management.

“Risks” include the following examples of risks, and loss may be incurred due to a fine or liability for damages because of deterioration of reputation and violation of laws and regulations resulting from these risks. See Appendix 1. “B. AI’s benefits and risks” for details about risks.

- Common risks of general AI
  - Output of biased results and discriminative results, filter bubbles, echo chambers, disinformation, handling of inappropriate personal data, data pollution attacks, obscuring, leak of confidential information, abuse of AI system services, energy consumption and environmental burden, reproduction of bias, etc.
- Risks presented by generative AI
  - Hallucinations, misinformation acceptance, relation with copyrights and other rights and eligibility, etc.
- Risks arising from organization and management
  - Not recognizing products or services include AI, lack of consideration regarding AI in governance, inappropriate or ubiquitous use of AI due to lack of environmental awareness or planning, etc., lack of organization of relationships between humans and AI such as separation of work duties, etc.

Important topics span across several AI business actors to secure benefits in the entire value chain / risk chain and reduce risks include the following examples:

- Distribution of responsibilities between AI business actors or within an AI business actor
- Improvement of quality of AI systems and services in general
- Possibility of new value created by mutual connection between AI systems and services (System of Systems)
- Improvement of literacy of AI Business Users and non-business users

[18] It should be noted that the executives may be held responsible for their management and supervision of AI governance obligations.

It is expected to design the most appropriate system for reporting to and sharing with the management based on the characteristics of a company/organization, for example, the following measures can be taken.

- Establishment of an internal organization for AI governance which assumes responsibilities to the board of directors (AI ethics committee, AI ethics examination committee, etc.)
- Reporting of measures related to AI governance at a board of directors meeting
- Documenting data that organize benefits/risks of use of AI for a company/organization and passing them internally
- Reflection of data in governance framework used internally, etc.

[Practical Examples]
[Practical Example i: Understanding benefits and risks]

It is important that AI business actors to examine not only benefits but also risks under the leadership of the management (including the examination the management itself implements by taking the lead instead of leaving it up to an officer in charge or staff in charge), share the results of examination and update the understanding at an appropriate time.

It is considered that benefits are already known, but we reorganized the benefits which may be offered by AI technology by using the comprehensive and detailed instruction manual, etc. such as “AI White Paper”[19] issued by the Information-technology Promotion Agency, Japan.

We also investigated if there was any incident in the past related to the same or similar function and area as that of an AI system/service we intend to develop, provide or use, or if there was any specific indication of a possible incident even if no incident has occurred in the past. Information on incidents can be obtained from various documents and the Internet. As we are planning to develop, provide and use AI only in Japan, we first collected information shared in Japan. In doing so, we referred to the “AI Utilization Handbook - Using AI wisely -” by the Consumer Affairs Agency.[20] For example, the Checkpoints include “AI can misrecognize speech, give incorrect instructions, and collect information about normal conversations.” This is a remark that expressed potential incidents from the viewpoint of non-business users.

There are also substantial number of books on AI that mention incidents and what may happen in the future. The Deep Learning for GENERAL: JDLA Certificate Examination offered by the Japan Deep Learning Association (JDLA) assesses ethical issues and the participants can obtain information on incidents as part of the examination. The “Final Recommendations on Profiling” clearly explains some cases.[21] In addition, we used the incident database described in “Column 1: Sharing information on incidents” below as a reference while recognizing that the social acceptance of AI systems and services vary from country to country and region to region. Based on the results of analysis conducted until now, it was discovered that many of the incidents are related to the handling of personal data, fairness and safety. An analysis of benefits/risks of specific individual AI systems and services will be conducted at the time of deviation evaluation for the Behavioral Goal 3-1.

[Practical Example ii: Understanding risks using framework when the scope of use of AI is extensive]

Since we develop, provide and use the various areas of AI systems and services, we roughly organize the incidents that had social impacts and issues indicated as having social impacts in the future in addition to the Practical Example i in order to grasp their overall picture in light of the general framework. We have created and use our own framework by referring to the OECD

[19] “AI White Paper 2023” by AI White Paper Editorial Committee (May 2023)
[20] “AI Utilization Handbook - Using AI wisely -” by the Consumer Affairs Agency (July 2020), https://www.caa.go.jp/policies/policy/consumer_policy/meeting_materials/review_meeting_004/assets/ai_handbook_200804_0001.pdf
[21] The “Final Recommendations on Profiling” by the Personal Data + α Study Group (April 2022), https://wp.shojihomu.co.jp/wp-content/uploads/2022/04/ef8280a7d908b3686f23842831dfa659.pdf

framework for the classification.[22] The chapter for the Economic Context roughly corresponding to the environment and risk analysis included in the OECD framework for the classification shows the general framework from the viewpoint of the relationship between the OECD AI principles and industrial sector, intended business use, stakeholders and scope of impacts, etc. We are currently considering reflecting such classification in our own framework while taking into account that this classification is just a tool that help understand the risks in general. An analysis of benefits/risks of specific individual AI systems and services will be conducted at the time of deviation evaluation for the Behavioral Goal 3-1.

[Practical Example iii: Understanding benefits/risks of collaboration between several AI business actors when the scope of use of AI is extensive]

We are aware that the scope of AI systems and services we develop, provide or use is extensive, and occurrence of an incident will have a significant impact on the society. For this reason, we conduct a cross-sectional analysis of benefits and risks of AI by considering that combination of information obtained from experiences we were directly involved in and experiences of our competitors or, in some cases, other industries will enable a more effective analysis. We continue this analysis at a regular frequency so we can consider reviewing the AI governance goals at an appropriate time even before the occurrence of an incident.

[Practical Example iv: Internal sharing of benefits/risks discovered]

We understand that the scope of AI systems and services we develop, provide or use is extensive, and such development, provision and use will have a significant impact on the society. Therefore, we consider that it is important to share the benefits and risks of AI internally and follow the following procedures:

First, we summarize information obtained and results of analysis, document data and pass them on to internal members involved. Members involved can provide comments and feedback about this document, and they exchange opinions actively. Specifically, we have a discussion with staff in charge at the relevant division and interested members through internal study groups and workshops to collect opinions from different perspectives.

We also appoint a full-time employee responsible for measures and progress related to AI Governance internally and such employee makes a report at a board of directors meeting. This promotes effective communication with the management.

These internal sharing systems secure the transparency and establish environment in which the entire organization can receive benefits of AI as much as possible and at the same time control risks appropriately.

[Practical Example v: Measures for generative AI]

Most recently, generative AI are starting to emerge, and we consider that this is a good opportunity for us. When using it for our business, we establish internal user guidelines according to the “Generative AI User Guidelines”[23] issued by JDLA. We also review information provided by the governments such as “Cautionary warning about use of generative AI, etc.” of the Personal Information Protection Commission Japan. It is also necessary to collect information on generative AI through news reports and social media, etc. We make efforts to understand the latest trends through the above measures.

[22] OECD, “OECD Framework for the Classification of AI Systems: a tool for effective AI policies”, https://oecd.ai/en/classification
[23] The “Generative AI User Guidelines Ver. 1.1” of the Japan Deep Learning Association (October 2023), https://www.jdla.org/document/#ai-guideline

Column 1: Sharing information on incidents

Regarding the risks associated with the development and operation of AI systems, a lot can be learned from the past incidents. Since AI systems are built inductively based on the data set, and risks associated with AI systems include many unintended risks, understanding the past incidents is effective to reduce the risks. In general, information on cases of incidents is obtained from public information such as news reports and theses, but it is not easy to access necessary information.

In order to solve this accessibility issue, the Partnership on AI released the AI Incident Database (AIID)[24] in November 2020. The AIID’s list includes not less than 2,000 cases of incidents with URL links and also provides a search application. In addition to the Partnership on AI, the AI Incident Tracker is available on GitHub.[25] OECD released the OECD AI Incidents Monitor (AIM)[26] as well. The AIM monitors news reports in the world as incidents and analyzes more than 150,000 pieces of English articles provided by the Event Registry which is the news intelligence platform everyday to include them in the list.

On the other hand, it is a challenge to establish such database. Most of the cases of incidents in the AIID are those in the initial list provided by scholars. It is also a challenge to accumulate information focusing on important information as the amount of information increases with the spread of AI. Some analyses point out that it is not easy to collect cases of incidents actively and treat them as shared assets because each company’s near-miss incidents not available to the public themselves are important experience and may become intellectual properties of each company.

[24] Partnership on AI, “AI Incident Database”, https://incidentdatabase.ai/
[25] jphall663, “awesome-machine-learning-interpretability”, https://github.com/jphall663/awesome-machine-learning-interpretability/blob/master/README.md#ai-incident-tracker
[26] OECD, “OECD AI Incidents Monitor (AIM)”, https://oecd.ai/en/incidents

Behavioral Goal 1-2 [Understanding social acceptance of AI]:
AI business actors are expected to understand the current level of social acceptance based on the opinions of stakeholders under the leadership of the management before starting the serious development, provision and use of AI. It is also expected to reconfirm the opinions of stakeholders in light of the changes in external environment even after the start of the serious development, provision and use of AI systems and services.

[Practice Guidelines]

AI business actors will, under the leadership of the management team, take the following measures:

- Identify stakeholders.
- Make efforts to understand the social acceptance after the identification for the development, provision or use of AI.
- Reconfirm the opinions of stakeholders at an appropriate time as necessary in light of the rapidly changing external environment even after starting the provision.

Identify stakeholders by considering the benefits and risks, which the provided AI offers to individuals, organizations, local communities and environment through a lifecycle. It is expected to pay attention to the possibility that the scope of stakeholders may be broader than expected. For example, the OECD framework for classification lists the following persons as a stakeholder:

- Persons who belong to each AI business actor
- Non-business users
- Corporations (business)
- Governmental agencies
- Research institutes
- Scientists/researchers
- Citizens’ groups
- Children and other socially vulnerable individuals and groups, etc.

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

- Official documents, scientific investigations, etc.
  - Surveys published by the governments and think tanks
  - Research papers
  - Opinions from citizens’ groups about AI systems/services
  - Seminars and conferences about AI ethics and quality
- Latest news reports
  - Investigation of cases of incidents
  - Reactions of stakeholders including non-business users on social media, blogs, bulletin boards and news coverage, etc.

The “ISO/IEC 23894:2023”[27] lists the following examples as external environment of an organization.

- Social, cultural, political, legal, registrational, financial, technical, economic and environmental factors
  - Relevant laws and regulations including those related to AI
  - Guidelines related to AI issued by the governments, civil society, academic conferences and industry groups, etc.
  - Guidelines and frameworks for each area, etc.
- Factors and trends which affect the goals of an organization
  - Technical trends and progress of each area of AI
  - Social and political meaning of introduction of AI systems including organization in the social and scientific guiding principles
- Relationship, recognition and values of stakeholders, etc.

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

- Contractual relationship and commitment to it
- Complexity of alignment and dependency between AI systems, etc.

To reconfirm the opinions of stakeholders, the following methods are effective:

- Direct feedback from stakeholders
- Evaluation of in-house AI management system and operation by experts

[Practical Examples]
[Practical Example i: Understanding social acceptance]

First, we used the results of surveys sent to non-business users released by the governments, public institutions and think tanks, etc. to understand the social acceptance. For example, the Consumer Affairs Agency has conducted a survey and released its results regarding the level of knowledge and understanding of AI users about “(i) the understanding of AI by consumers, (ii) expectations for AI by consumers, issues and usage trends, (iii) AI in services used by consumers (what kind of risks can exist), and (iv) risks related to AI services” at the “Study Group on Consumers’ Response to Digitalization AI Working Group.” Since we are considering to expand the AI field internationally, we also referred to the surveys sent to overseas non-business users. In addition, we referred to the opinions from citizens’ groups about AI systems/services.

As the information about social acceptance obtained from these sources will be used for the general design of AI governance, it is expected to cut off branches and extract mainstream information so the management can make decisions. We use information obtained through the Behavioral Goal 1-1 and at the same time use the risk-based approach to organize information on the social acceptance such as by classifying various AI systems/services into categories based on the degree of their risk including the intended use for which social understanding is unlikely to be fostered even if any kind of explanation is given, intended use for which social understanding is likely to be fostered if active and sufficient explanation is given, intended use for which social understanding is likely to be fostered if explanation is given as necessary, and intended use which is unlikely to cause risks to non-business users, etc.

[Practical Example ii: Understanding social acceptance using external seminars, etc.]

In addition to the Practical Example i, we take initiatives to send staff in charge to seminars and conferences on AI ethics and quality held by universities and industry groups. Recently, many of these seminars, etc. are held in the webinar format, and we can obtain information more efficiently than before. If we access overseas webinars, we can understand the international trends of AI ethics and quality.

[Practical Example iii: Understanding social acceptance through stakeholders]

We have used the methods used in the Practical Examples i and ii until now, and we understand that expectations from stakeholders about the appropriate use of AI by us are relatively high because we develop, provide and use AI systems/services seriously and extensively. For this reason, instead of understanding the opinions of stakeholders indirectly and passively, under the leadership of the management, we changed our policy to understand them directly and actively.

Under this new policy, we invite experts who are familiar with the social acceptance of AI and regularly hold meetings of an AI governance committee which includes external experts, etc. We use this committee not only to receive results of evaluation of our AI management system and operation but also to improve the understanding of environment in which we are placed such as general social acceptance of AI. We understand that information obtained by the committee is more thorough for us compared to general information obtained in the Practical Examples i and ii and is often not known to the public. We use the risk-based approach to analyze the social acceptance in detail by combining information obtained by this committee and general information obtained in the Practical Examples i and ii. The results of analysis are organized by members in the leadership positions, and they are reported by the members in the leadership positions to those in the management positions (executive members).

Behavioral Goal 1-3 [Understanding company’s AI proficiency]:
AI business actors will, under the leadership of the management, evaluate its in-house AI proficiency and reevaluate it at an appropriate time to implement the Behavioral Goals 1-1 and 1-2 based on the experience of development, provision and use of its AI systems/services, the number and experience of employees including engineers involved in the development, provision and use of AI systems/services and the level of literacy of such employees about AI technologies and ethics, etc. unless it deems that the level of risk is low in light of the intended use of AI to be used, its business field and size, etc. If possible, it is expected to disclose the results of the evaluation to the extent reasonably possible. If the risk is deemed low, and the AI proficiency evaluation is not conducted, it is expected to disclose the fact that the evaluation is not conducted to stakeholders along with the reasons.

[Practice Guidelines]

AI business actors will, under the leadership of the management team, take the following measures:

- Consider the necessity of AI proficiency evaluation in light of the business field and size, etc. of AI business actors.
- If the evaluation is deemed necessary, visualize the responsiveness to AI risks and evaluate the AI proficiency (how much preparation necessary for the development, provision and use of AI systems/services is done).
  - If possible, disclose the results of the evaluation to stakeholders to the extent reasonably possible.
- If the evaluation is deemed unnecessary, disclose the fact that the evaluation is not conducted to stakeholders along with the reasons if possible.

Improvement of efficiency, etc. by AI systems/services may bring benefits to business such as mitigation of labor shortages, productivity gains and development of high value-added business. On the other hand, unregulated provision of AI systems/services for business involves AI’s unique risks such as unintended loss of fairness and security issues. Therefore, it is important for AI business actors to start introducing AI after understanding these risks which are considered as the disadvantages of the introduction of AI, and it is important to evaluate the AI proficiency.

Use of the following guidelines is effective to evaluate the AI proficiency. It is expected to check the latest version of each of the guidelines because it may be reviewed in light of the changes in environment including advancement in use of generative AI.

- Guidelines published in the “Using AI to Realize Society 5.0 for SDGs” by the Japan Business Federation (June 2023)[28]
- Certification Examination administered by the Japan Deep Leaning Association
  - Generative AI Test[29]
  - JDLA Deep Learning For GENERAL (“G-kentei”)[30]
- NIST, “Artificial Intelligence Risk Management Framework” (AI RMF 1.0)[31]

[Practical Examples]

[Practical Example i: Evaluation of proficiency using the guidelines published in the “Using AI to Realize Society 5.0 for SDGs” (June 2023)]

We evaluate our AI proficiency and reevaluate it at an appropriate time under the leadership of the management when developing, providing or using AI systems/services so that we will not

[28] “Using AI to Realize Society 5.0 for SDGs” by the Japan Business Federation (June 2023), https://www.keidanren.or.jp/policy/2023/041.html
[29] “Generative AI Test” by the Japan Deep Learning Association”, https://www.jdla.org/document/#ai-guideline JDLA https://www.jdla.org/certificate/generativea
[30] “What is G-kentei” by the Japan Deep Learning Association, https://www.jdla.org/certificate/general/
[31] NIST, “Artificial Intelligence Risk Management Framework (AI RMF 1.0)”, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

focus only on benefits, lack consideration for risk and cause a significant damage to stakeholder as a result of introduction of AI systems/services.

We use the guidelines posted in the “Using AI to Realize Society 5.0 for SDGs” by the Japan Business Federation (June 2023) to evaluate the AI proficiency. This is for evaluating if the range of benefits/risks offered by our AI systems/services to the society and expansion of relevant stakeholders[32] is relative to our AI proficiency. In addition, we utilize the AI proficiency for examining the overall AI governance including examination of AI governance goals.

[Practical Example ii: Evaluation of proficiency using the unique index while referring to guidelines published in the “Using AI to Realize Society 5.0 for SDGs” (June 2023)]
We just started taking measures to examine the AI governance seriously. Therefore, we have adopted some provisions of the guidelines published in the “Using AI to Realize Society 5.0 for SDGs” (June 2023) by referring to them and created our own index that is appropriate for our AI governance. We plan to adopt more provision to check the AI proficiency in the future by conducting evaluation using the index and infiltrating the current AI governance and system internally based on the results of the evaluation.

[Practical Example iii: Evaluation of generative AI proficiency]

Since generative AI is starting to show its value these days, we consider the elements of generative AI when conducting proficiency evaluation to include its impacts by utilizing the “Generative AI User Guidelines”[33] issued by the JDLA. As we’ve heard that the “Guidelines for AI Ready Society” will be updated in light of generative AI, we plan to conduct reevaluation based on the updated guidelines.

[32] The AI business actors who are directly or indirectly involved in AI utilization through AI utilization, including AI developers, AI providers, AI Business Users, and non-business users.
[33] The “Generative AI User Guidelines Ver. 1.1” of the Japan Deep Learning Association (October 2023), https://www.jdla.org/document/#ai-guideline

2. Goal setting

Behavioral Goal 2-1 [Setting AI governance goals]: AI business actors will, under the leadership of the management, examine the necessity of and set its own AI governance goals (such as AI policies) while considering the benefits/risks that may be offered by AI systems/services, considering the social acceptance and its own AI proficiency regarding the development, provision and use of AI systems/services and paying attention to the importance of process for setting AI governance goals. In addition, it is expected to disclose the set goals to stakeholders. If the AI governance goals are not set on the grounds that potential risks are minor, it is expected to disclose the fact that the goals will not be set to stakeholders along with the reasons. If it is deemed that the “common guiding principles” in these guidelines will function sufficiently, the “common guiding principles” may be used as the governance goals in lieu of a company’s own AI governance goals. Even if the goals are not set, it is expected to understand the importance of these guidelines and take measures to achieve the Behavioral Goals 3 to 5 as appropriate.

[Practice Guidelines]

AI business actors will, under the leadership of the management team, take the following measures:

- Consider if setting “AI governance goals” of AI business actors is necessary.
  - Set the goals flexibly based on the size of each AI business actor and risks associated with AI handled.
- Set the goals if deemed necessary.
  - If possible, disclose the goals to stakeholders to the extent reasonably possible.
- If setting the goals is deemed unnecessary, disclose the fact that the goals will not be set to stakeholders along with the reasons if possible.

The following is considered as the elements of the “AI governance goals,” and typical examples are introduced in various books.

- A company’s policy consisting of measures taken to achieve the “common guiding principles” described in these guidelines (the name of policy can vary for each company, such as “AI policy”)
- Privacy policy summarizing the guiding principles for use of data related to privacy in addition to the measures taken to achieve the “common guiding principles,” etc.
- Policy for receiving more benefits such as improvement of inclusion, etc. by AI utilization
- Degree of risk tolerance

It is also effective to establish the code of conduct not disclosed to the public for employees and announce it internally (especially to staff members in charge) to improve awareness of employees when creating the “AI governance goals” to be disclosed externally.

The “common guiding principles” included in these guidelines can be used as the “AI governance goals,” and even when an AI business actor’s own AI governance goals are set, it is expected to refer to the details of the “common guiding principles.” Developing the “AI governance goals” based on the “common guiding principles” will allow each AI business actor to conduct risk evaluation based on the “common guiding principles” by linking potential risks to the “common guiding principles.”

Pay close attention to the following when setting AI governance goals:

- Ensure that there is no inconsistency or contradiction when setting goals related to AI such as “AI governance goals” and purpose of use of AI in line with the management goals such as the meaning of existence, corporate philosophy and vision of AI business actors.

- Convey the management goals such as the meaning of existence, corporate philosophy and vision of AI business actors, and goals related to AI that are consistent with the management goals when applying the PDCA cycle based on the “AI governance goals” to the organizational management.
- Consider the impacts expected by stakeholders and risks that stakeholders can face by identifying the stakeholders and avoid discrepancies.

[Practical Examples]
[Practical Example i: In the case of not setting AI governance goals]
We plan to handle AI systems/services only for the intended use with minor potential risks to the society soon after starting the development of AI systems. For this reason, we have not set AI governance goals, however, if we expand the scope of business and it cannot be said that potential risks are minor, we will consider setting AI governance goals. As a matter of course, we can explain the reasons for not setting AI governance goals, etc. to stakeholders by recording the details of consideration.

[Practical Example ii: Setting AI governance goals at small business operators]
Although we started the development of AI systems not so long ago, we decided to take AI governance measures because the risk of AI systems we develop is not minor. However, as we don’t have many employees, it is difficult to appoint a particular person in charge of AI governance, etc. Therefore, the top management prepared “Our AI development policy” which is linked to our management philosophy and shared it with the employees in charge. Thereafter, the management and employees improved “Our AI development policy” and ensured that all members have understating at the same level while sharing information on incidents occurred. Although the content fits on one page of A4 size paper, we decided to use it as our AI governance goals.

[Practical Example iii: Setting AI governance goals involving each division]
Our business portfolio includes a wide range of services, and each division is involved in AI technologies differently. In addition, since each division uses an independent company system, it is not easy for each division to agree upon a single AI governance goal. For this reason, we will respect the “Common guiding principles” of these guidelines at the moment, and at the same time, we plan to improve our understanding about AI ethics and quality by adding AI ethics and quality to the content of some of the company-wide AI training programs. Furthermore, we have established an AI help desk to collect examples of cases from each division. Although others may consider that we are making progress slowly, we think the processes to gain agreement upon AI governance goals have values. We may examine the necessity and details of AI governance goals of each division engaging in the development, provision or use of AI systems/services before setting AI governance goals of the entire AI business actor.

[Practical Example iv: Setting AI governance goals involving stakeholders]
We have abundant experience in supporting other AI business actors in addition to the development, provision and use of AI systems/services and also develop, provide and use AI systems/services for the intended use with potential risks that are not considered to be minor. Although no serious incident has occurred in relation to the AI systems/services developed by us and AI systems/services provided to other companies, we understand that the social acceptance has not been established for many of the intended use for which we provide AI systems/services. For this reason, we have set and announced our AI governance goals to enhance communication with stakeholders. It has been evaluated that understanding of our policy by stakeholders has enabled persons in charge of developing AI systems/services and stakeholders to share the basic viewpoint about AI technologies, resulting in smooth communication among them.

3. System Design (Building AI management system)

Behavioral Goal 3-1 [Requiring evaluation of deviation from goals and measures to minimize the deviation]:
Under the leadership of the management, AI business actors are expected to incorporate the process to identify a deviation of AI from each AI business actor’s AI governance goal, evaluate impacts caused by the deviation, and if a risk is realized, evaluate if it is reasonable to accept the risk by considering its seriousness, scope and frequency, etc., and if it is not reasonable to accept the risk, reconsider the development, provision or use of AI, into the appropriate stage such as the design stage, development stage, before the start of use and after the start of use of AI systems/services. It is important for the management to develop basic policies, etc. regarding the reconsideration process and for those in leadership positions to shape the process. It is also expected to include persons who are not directly involved in the development, provision or use of AI when evaluating a deviation from AI governance goals. It is not appropriate to arbitrarily reject the development, provision or use of AI only on the ground that there is a deviation. For this reason, the deviation evaluation is a step for evaluating risks, and it is just an opportunity to make improvements.

[Practice Guidelines]

AI business actors will, under the leadership of the management team, take the following measures:

- Identify difference between “AI governance goals” and current situation, use a risk-based approach to select controls for risks, and implementing appropriate levels of management for each use case, service, or product.
- Identify and evaluate a deviation of the current state of AI systems/services and “AI governance goals.”
- If a risk is realized, determine if it is reasonable to accept the risk.
- If it is not reasonable to accept the risk, reconsider/incorporate the process to reconsider the way an AI system/service is developed,[34] provided or used into the appropriate stage of development, provision or use and process of making decisions in organizations of AI business actors.
- The management will take initiatives and be responsible for the decision making, and those in leadership positions will shape the above process to follow it on a continuing basis.
  - Understand that the responsibility to build AI governance, organizational management and project management systems is as important as the operational responsibility.
- Share the deviation evaluation items in AI business actors to build internal understanding.
  - Collaborate between AI business actors to evaluate a deviation depending on the details of AI provided.

In some cases, it is expected that processes will be established for a deviation evaluation (to measure if an AI system functions as it’s designed and performs tasks such as prediction and reasoning/inference accurately) by using the knowledge of external experts and referring to the following materials based on the degree of risk of each company’s own AI systems/services.

- Standard deviation evaluation process of each industry described in the Behavioral Goal 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

[34] Commonly used development terminology for it is “Continuous Integration (CI)” or “Continuous Delivery (CD)”, etc.

The “ISO/IEC 42001”[35] states that it is important to ensure that the AI management system is integrated into the business process of an organization under the leadership of the management, and the following specific steps are expected to be taken:

- To secure sufficient resources for AI management system.
- To announce the importance of effective AI management and importance of conformity with the requirements of AI management system in AI business actors.
- To allow an AI management system to achieve the intended AI governance goals.
- To direct and support staff members who will contribute to the effectiveness of the AI management system.
- To promote continuous improvement.
- Otherwise to support relevant AI business actors and take initiatives in each area of responsibilities.

[Practical Examples]
[Practical Example i: Deviation evaluation process of small business operator]

We are a small company, our officer in charge of technology and staff members in charge of development feel close to each other, we do not have many projects, and the officer in charge of technology understand all the projects well. The officer in charge of technology determines the viewpoints for evaluating a deviation from the “common guiding principles” of these guidelines, identify a deviation from each viewpoint regarding all AI system development projects for the staff members in charge of development at a stage as early as practically possible, evaluate impacts caused by the deviation and instruct them to report to the officer in charge of technology. The officer in charge of technology evaluates impacts caused by the deviation again based on the reports from the staff members in charge of development at a meeting attended by staff members other than those in charge of development, determines if it is reasonable to accept a risk if any and reconsider the way we provide AI if it is considered unreasonable to accept the risk.

When operating this process, we refer to the standard deviation evaluation in the industry to which we belong and these guidelines according to the Behavioral Goal 3-1-1.

[Practical Example ii: Deviation evaluation process for each project of business operator with many divisions]

We are a company with many divisions and we have appointed an officer in charge of AI governance and established an AI governance committee under the direction of the officer. This committee consists of employees other than those in charge of development, provision or use of specific AI systems/services and performs duties to evaluate a deviation from the AI policy established by us for each project based on the “common guiding principles” of these guidelines. Specifically, we create an evaluation list according to the AI policy, identify a deviation by using the evaluation list regarding the development, provision or use of AI systems/services, evaluate impacts caused by the deviation, determine if it is reasonable to accept a risk if any and encourage the person in charge of the project in question to reconsider the way AI is developed, provided or used if it is considered unreasonable to accept the risk. We create a list for deviation evaluation according to the Behavioral Goal 3-1-1 by referring to the standard deviation evaluation in the industry to which we belong and these guidelines. In addition, we select an actual project and have the management accompany the person in charge of a project to elaborate the list and conduct deviation evaluation on a regular basis. The AI governance committee requires the person in charge of a project to report the results of reconsideration and requires the officer in charge of AI governance to notify the officer supervising the project if it is suspected that the reports may not be reasonable in order to make adjustments as appropriate.

[35] In addition to principles and laws of nations or international bodies, it is also useful to refer to standards set by non-governmental organizations such as the International Organization for Standardization (ISO). As an international standard for AI risk management, the “AI Management System (ISO/IEC 42001)” was issued on December 18, 2023, by a joint technical committee of ISO and the International Electrotechnical Commission (IEC).
https://www.meti.go.jp/press/2023/01/20240115001/20240115001.html

Since risks associated with AI systems/services vary substantially depending on the intended use, scope and form of use, and it is considered that the person in charge of carrying out a project knows the nature and degree of a risk more than anybody else, it may be possible to conduct only a simple deviation evaluation at a meeting attended by the management without requiring a strict deviation evaluation if it is obvious that a potential risk is minor. However, since we don’t have enough knowhow to evaluate a deviation and risk at the moment, we require a deviation evaluation by the AI governance committee for all the projects and will wait and see how things develop.

[Practical Example iii: Deviation evaluation process by collaboration between AI business actors]

In some cases, it is necessary for several AI business actors to be responsible for the deviation evaluation process. For example, if an AI provider who provides services to others appoints an AI developer to perform the development of an AI system, in some cases, it is reasonable for the AI developer and AI provider to share the responsibilities for the deviation evaluation process. In this case, it is important that the AI developer and AI provider share information on the flow of processes from the development to operation of the AI system as well as the method and standards for the deviation evaluation. Such actions are important because if the AI provider ignores a risk associated with the provision of services using an AI system, the AI developer will be put in a difficult position.

When we provide services to develop an AI system for others, we enter into an agreement that requires an AI provider providing services to be liable for any accident occurring in connection with the operation of an AI system except in cases where there are reasons attributable to us, however, we may be forced to be involved in a dispute if this type of accident occurs. For this reason, we cannot be indifferent to the way of operation of an AI system delivered by us. In the past, we noticed an operational risk at the last stage of a project and had to advise an AI provider to redesign the project and bear part of the resigning costs. For this reason, we decided to share information with AI providers who provide services to others without engaging in the development of AI systems after fully understanding each evaluation item and establishing the deviation evaluation process by referring to the standard deviation evaluation in the industry to which we belong and these guidelines. Use of the deviation evaluation process covering all items of concern and early deviation evaluation help negotiation with customers go smoother.

In some cases, it is necessary to discuss various topics in addition to the ordinary deviation evaluation process as follows:

[Practical Example iv: Additional measures taken by small business operator in light of AI risks]

We are a small company developing AI systems as our core business. The officer in charge of technology receives reports on progress for all the projects, and the reports include matters related to AI ethics such as fairness. Although some AI ethics issues can be handled with technical considerations such as preparation of sufficient data sets to enable output of reasonable results, in some cases, such measures are not adequate to handle socially sensitive topics.

Therefore, we have a meeting attended by the officer in charge of legal affairs, etc. for an AI system project that involves such sensitive topics. We refer to the ideas of leading companies already developing, providing or using the wide range of AI systems/services to identify sensitive topics. Practical magazines are useful for collecting such information.[36] Summary articles are often posted on such magazines, and it is efficient and effective to collect information on the internet, etc. by using the summary articles as a clue.

We are aware that some companies invite external experts or specialists for project to exchange opinions. We are considering to provide such opportunities for exchanging opinions as our business expands.

[36] The examples of measures for handling sensitive topics include the “Company’s Efforts for AI Ethics (1)” NBL No. 1170 (May 2020) by Satoshi Funayama, etc.

[Practical Example v: Deviation evaluation by collaboration with external experts as necessary in addition to collaboration between AI business actors]

We are a large-scale company having both divisions for developing AI systems/services and divisions for operating them. We have already established our AI policy and conducted deviation evaluation according to the policy for all the projects. Although the management can handle risks if they are identified at an early stage of a project that is a type of project we dealt with in the past, if we develop or use an AI system/services that involves a sensitive topic we have never handled before, we require employees to have consultation on a case by case basis instead of following ordinary processes. When we have such consultation, we hold a cross-sectional meeting consisting of responsible persons of the development section, operation section and legal affairs section, etc. to discuss the issue. The same procedures are taken when the management discovers such project during the ordinary deviation evaluation processes.

We regularly invite external experts and specialists to catch information related to the latest AI incidents and sensitive topics at an early stage. This allows us to handle risks for now by discussing at cross-sectional meetings based on information collected from experts and specialists or general advices. On the other hand, we consider that it may be necessary to ask the opinions of external experts, etc. for individual projects in the future because the intended use and customers of our AI systems/services have expanded.

[Practical Example vi : Activities Related to the Provision and Use of AI Using a Risk-Based Approach]

Our company has established six examination items as part of the risk-based approach to the provision and use of AI: ensuring transparency, ensuring fairness, ensuring reliability, public disclosure of AI use, protection of intellectual property, and others. For each examination item, we define potential risks and control methods (minimum measures) for those potential risks. For example, in ensuring transparency, a potential risk is “the risk of not being able to perform post-verification of AI when events occur due to AI decisions or when verification is necessary, due to not saving the model version.” As a control method for this, we stipulate “storing the training data used for AI model development.” The actual AI provision and use departments evaluate the presence of risks and the specific control methods for them, and the risk evaluation department further assesses these results to comprehensively determine the appropriateness of responses to risks.

Consideration items		Potential risks	Other specific risks ①	Control method	Other control method ②	Validity assessment	Judgment
Major items	Medium item		Entry field Responsible department		Entry field Responsible department	Entry field Risk assessment office	Entry field Risk assessment office
Ensuring transparency	Management of data and trained models	・The risk of not saving the versions of training/inference data and trained models, which makes it impossible to retrospectively verify the basis of decisions when events occur due to AI decisions or when verification is necessary	No risks other than those listed on the left	・Storage of learning data used in AI model development ・Storage of inference data used and inference results … (Exception: Data that is... is not subject to storage)	Respond as shown on the left	In ①, is there a risk that XXX could be considered ?	**Remand** Further consideration of the points to be considered in ①
…	…	…	…	…	…	…	…

Figure 8. Example of Risk-Based Approach

Behavioral Goal 3-1-1 [Ensuring consistency with the industry’s standard deviation evaluation processes]: Under the leadership of the management, AI business actors are expected to confirm if there are standard deviation evaluation processes in the industry and incorporate such processes into their own processes if any.

[Practice Guidelines]

AI business actors are expected to take the following measures under the leadership of management:

- To actively incorporate the best practices of external entities such as the standard deviation evaluation processes in the industry and measures taken by other companies and organizations, etc. in addition to its own knowledge and experience.

Since there may be guidelines that can be used as reference in each industry and guidelines for the AI reliability assessment released by each ministry, agency and industry group, it is also useful to review information provided by the relevant ministries, agencies and groups.

For example, the following guidelines are included:

- Common across industries
  International Organization for Standardization, “ISO/IEC42001” (April 2024)[37]
  National Institute of Advanced Industrial Science and Technology, “Machine Learning Quality Management Guideline” (December 2023)[38]
  NIST, “AI Risk Management Framework Playbook” (January 2023)[39]
- Government agencies and local governments
  Digital Agency, “Risk Countermeasure Guidebook for the Utilization of Text Generation AI (α version)” (June 2024)[40]
  Tokyo Metropolitan Government, “Guideline for the Utilization of Text Generation AI” (April 2024)[41]
- Agriculture
  Ministry of Agriculture, Forestry and Fisheries, “Contract Guidelines on AI and Data in the Agricultural Field” (March 2020)[42]
- Manufacturing
  Ministry of Economy, Trade and Industry, “AI Introduction Guidebook” (April 2022)[43]
- Healthcare
  Japan Digital Health Alliance (JaDHA), “Guide for the Utilization of Generative AI for Healthcare Providers version 2.0” (February 2025)[44]
  Japan Primary Care Association (JPCA), “Guideline for AI Utilization in Primary Care” (December 2023)[45]
- Finance
  Financial Services Agency, “Principles for Model Risk Management” (November 2021)[46]
  Financial Data Utilization Promotion Association (FDUA), “Practical Handbook for Financial Generative AI” (May 2024)[47]
  Financial Data Utilization Promotion Association (FDUA), “Guideline for Financial Generative AI” (August 2024)[48]
- Education
  Ministry of Education, Culture, Sports, Science and Technology, “Guidelines for the Use of Generative AI in Primary and Secondary Education” (December 2024)[49]
  Ministry of Education, Culture, Sports, Science and Technology, “On the Handling of Generative AI in University and Technical College Education” (July 2023)[50]
- Defense

[37] https://www.meti.go.jp/press/2023/01/20240115001/20240115001.html
[38] https://www.aist.go.jp/aist_j/press_release/pr2020/pr20200630_2/pr20200630_2.html
[39] https://www.nist.gov/itl/ai-risk-management-framework/nist-ai-rmf-playbook
[40] https://www.digital.go.jp/resources/generalitve-ai-guidebook
[41] https://www.metro.tokyo.lg.jp/tosei/hodohappyo/press/2023/08/23/14.html
[42] https://www.maff.go.jp/j/kanbo/tizai/brand/keiyaku.html
[43] https://www.meti.go.jp/policy/it_policy/jinzai/AIutilization.html
[44] https://jadha.jp/news/news20250207.html
[45] https://www.primarycare-japan.com/news-detail.php?nid=625
[46] https://www.fsa.go.jp/news/r3/ginkou/20211112/pdf_02.pdf
[47] https://www.fdua.org/activities/generativeai
[48] https://www.fdua.org/activities/generativeai
[49] https://www.mext.go.jp/content/20241226-mxt_shuukyo02-000030823_001.pdf
[50] https://www.mext.go.jp/b_menu/houdou/2023/mext_01260.html

Ministry of Defense, “Basic Policy for Promoting AI Utilization” (July 2024)[51]

- Cloud Services
  Ministry of Internal Affairs and Communications, “Guidebook on Cloud Services Utilizing AI” (February 2022)[52]

[Practical Examples]
[Practical Example i: Incorporating deviation evaluation processes under the guidelines of other companies and organizations]

As various viewpoints are necessary to practice AI governance, and sharing understanding with other companies is also necessary, we should not only think on our own but also refer to the measures taken by other companies and organizations, etc. Since we consider as described above, our management instructed the person in charge of governance to investigate the measures taken by external entities to build our own deviation evaluation processes.

We mainly investigated the industrial use because our core business is the development of AI systems for industrial use. During the investigation, we discovered that for example, the Ministry of Economy, Trade and Industry, the Ministry of Health, Labour and Welfare and the Fire and Disaster Management Agency had released the “Practical Examples of Reliability Assessment Records” [53] containing examples, “Format for Recording Details of Implementation” [54] to implement the “Guidelines on Assessment of AI Reliability in the Field of Plant Safety.” We also discovered that the “AI Product Quality Assurance Guidelines”[55] released by the Consortium of Quality Assurance for Artificial-Intelligence-based Products and Services contain examples of Voice User Interface, processes for industrial use, automatic operation and OCR. In addition, we discovered that the National Institute of Advanced Industrial Science and Technology had released the “Guidelines for Machine Learning Quality Management” and prepared the reference guide as specific examples in which real applications are applied for each intended industrial use. The guidelines also contain the quality management processes in line with the guidelines, form of machine learning quality guideline assessment sheet suitable for the plan and records, and also instruction manuals.[56] Part of these specific measures are reflected on our current deviation evaluation processes.

[Practical Example ii: Incorporating points to note into deviation evaluation processes when handling personal data]

We develop, provide and use AI systems/services based on the data collected from non-business users. We understand that we need to pay attention when handling data input to AI models in addition to paying attention to the AI models building and output for implementing AI governance, in particular, to protect privacy. Although we have abundant experience in handling personal data, we consider that it is important to actively turn our attention to measures taken by external entities. Therefore, our management instructed the person in charge of privacy to investigate the measures taken by external entities to build our own deviation evaluation processes.

We discovered that consideration given to AI model building and output include “Checklist for Voluntary Efforts”[57] in profiling provided by the Personal Data + α Study Group. The “Guidebook

---

[51] https://www.mod.go.jp/j/press/news/2024/07/02a.html
[52] https://www.soumu.go.jp/menu_news/s-news/01ryutsu06_02000305.html
[53] 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.fdma.go.jp/relocation/neuter/topics/fieldList4_16/pdf/r03/jisyuhoan_shiryo_03_04.pdf
[54] Ministry of Economy, Trade and Industry, Ministry of Health, Labour and Welfare, and Ministry of Internal Affairs and Communications “Format for Reliability Assessment Implementation Records”, https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fwww.fdma.go.jp%2Frelocation%2Fneuter%2Ftopics%2FfieldList4_16%2Fpdf%2Fr03%2Fjisyuhoan_shiryo_03_03.xlsx&wdOrigin=BROWSELINK
[55] Consortium of Quality Assurance for Artificial-Intelligence-based Products and Services, “AI Product Quality Assurance Guidelines ver. 2023.06” (June 2023), https://www.qa4AI.jp/
[56] National Institute of Advanced Industrial Science and Technology, “Guidelines for Machine Learning Quality Management” (July 2022), https://www.digiarc.aist.go.jp/publication/aiqm/referenceguide.html
[57] Personal Data + α Study Group “Final Recommendations on Profiling” p. 10–21 (April 2022)

on Corporate Governance for Privacy in Digital Transformation (DX) ver. 1.3”[58] was useful because it has descriptions about AI from both the input and output viewpoints. Part of these specific measures are reflected on our current deviation evaluation processes.

Behavioral Goal 3-1-2 [Provision of sufficient information about deviation possibility/countermeasures to AI Business Users and non-business users]: If a certain deviation may occur in the AI systems/services provided, AI business actors are expected to provide sufficient information about the deviation and countermeasures and also clearly indicate contact information to stakeholders under the leadership of the management.

[Practice Guidelines]

AI business actors are expected to take the following measures under the leadership of management:

- If a deviation is likely to occur between AI systems/services and the “AI governance goals,” provide information about the deviation and countermeasures to stakeholders and communicate with them by responding to inquiries, etc.
- Collaborate with AI developers and industry groups, etc. to improve the effectiveness of provision of information and contribute to the improvement of literacy of AI Business Users and non-business users by sending various information.
- Consider the level of provision of information based on the nature and probability of risks caused by a deviation.

The specific example of provision of information tailored to the literacy of stakeholders include selection of terminology.

- If there are significant differences in the literacy, explain the basic structure of AI systems/services and provide clear explanation so that all stakeholders can understand.
- If the level of literacy of stakeholders is high, provide balanced explanation using some technical terms.

Communication through receipt of inquiries includes the following:

- Contact information must be clearly indicated for receiving inquiries.
- Make the indication that AI is used in the system on the website, etc. easy to find.

[Practical Examples]

[Practical Example i: Provision of information by referring to the “Guidebook on Corporate Governance for Privacy in Digital Transformation (DX) ver. 1.3”]

We operate AI systems/services and provide AI services to a lot of non-business users. Since it is expected that there are significant differences in the AI literacy of recipients of our services, we organize and provide information related to risks such as appropriate risk management we conduct for operating AI systems/services, measures for mitigating risks and strict security management of information so that non-business users who are not familiar with AI can understand and also clearly indicate our contact information for inquiries. As described above, it is expected that there are significant differences in the AI literacy of recipients of our services, we clearly indicate that AI is in use and also indicate advantages and disadvantages of using AI in addition to the above information unless it is obvious that output via AI systems/services is used for provided information, etc. We also indicate that we can provide substitute services for non-business users who do not like AI. As we handle personal data, we communicate with non-business users on a continuing basis in accordance with the Act on the Protection of Personal Information and the guidelines of the Personal Information Protection Commission and by referring to the “Guidebook on Corporate Governance for Privacy in Digital Transformation (DX) ver. 1.3.”

[58] Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry “Guidebook on Corporate Governance for Privacy in Digital Transformation (DX) ver. 1.3” (April 2023), https://www.soumu.go.jp/mAIn_content/000877678.pdf

[Practical Example ii: Provision of information tailored to AI literacy of recipients (if the level of AI literacy of recipients is high)]

As in the Practical Example i, we operate AI systems/services and provide AI services to others, however, we provide AI systems/services to companies that use them for business, which is different from the Practical Example i. Since the level of AI literacy of recipients of our services is relatively high, we provide balanced explanation using some technical terms about the possibility of a certain deviation in the AI systems/services we provide and countermeasures for the deviation and also clearly indicate our contact information for inquiries.

We may provide AI services using AI systems for non-business users in the future, and we want to provide sufficient information tailored to the AI literacy of recipients of our AI services.

[Practical Example iii: Provision of information tailored to AI literacy of recipients (if there are differences in the level of AI literacy of recipients)]

We take the same measures as those in the Practical Example i, and we make efforts to provide information in a way that is different from other companies because we consider that there is added value in allowing AI Business Users and non-business users to select AI services using AI systems at their own discretion. We also make efforts to receive feedback not only for AI systems/services but also for the way of provision of information.

[Practical Example iv: Collaboration with AI developer, etc.]

We take the same measures as those in the Practical Example i, and we clearly specify in a contract that AI developers must provide information necessary to respond to inquires from AI Business Users and non-business users. We require AI developers to respond to “feedback” from AI Business Users and non-business users, which can be valuable information for AI Developers.

Column 2: Provision of sufficient information on deviation evaluation by data providers to AI developers

Data providers are expected to provide information on data sources, data collection policy and criteria, annotation criteria, terms of use and other information on dataset so that AI developers and AI providers can conduct a deviation evaluation appropriately, and AI developers are expected to obtain dataset from data providers who provide sufficient information. In the case of generative AI LLM, receive information from AI service providers as much as possible and share the fact that the information has been received with relevant stakeholders because provision of information on dataset may be limited.

[Points to note]

Quality related to the fairness, etc. of AI systems/services rely heavily on the data used for them. Therefore, sufficient information must be received from data providers regarding the data used for AI systems/services for AI developers and AI providers to conduct deviation evaluation.

Examples of information related to dataset includes the following:

- Data collection policy: Approaches to collection of data, etc.
- Data sources: Scope of providers/collectors of original data, scope of data collected, etc.
- Data collection policy: Data collected, items of data, collection method, collection period, etc.
- Data collection criteria: Conditions of data collected, method of cleansing, bias in data, etc.
- Data annotation criteria: Rules of annotation of images/sounds/texts, etc.
- Restriction on data use: Restriction arising from other rights, etc.
- Purpose of use of data: Specific purpose indicated by the data subject, etc. especially if data include personal data

[Practical Examples]

We are a data provider providing data to AI developers/AI providers and provide information on data sources, data collection policy and criteria, annotation criteria, terms of use and other information on dataset to allow companies developing AI systems to conduct deviation evaluation appropriately. We also provide sufficient basic information on data sources, etc. necessary for deviation evaluation even when we provide unorganized dataset.

Behavioral Goal 3-2 [Improving literacy of those in charge of AI management system]:
AI business actors are expected to consider using education materials of external sources and improve AI literacy strategically in order to appropriately operate the AI management system under the leadership of the management. For example, the following education and training can be provided: education for improving general literacy related to AI ethics and AI reliability to officers, management team and persons in charge, who are responsible for the legal and ethical aspects of AI systems/services, training related to AI technology including generative AI in addition to AI ethics to persons in charge of projects for developing, providing and using AI systems/services, and education regarding the positioning and importance of AI management system to all employees.

[Practice Guidelines]

AI business actors are expected to take the following measures under the leadership of management:

- Make efforts to improve AI literacy by using training and education materials suitable for the job titles and job duties including those offered by external lecturers.
- Use training and educational materials suitable for each member’s roles.
- Make efforts to require all employees to receive education about AI ethics which is especially important.

- Make efforts to provide training related to the reliability of generative AI technology and output results based on the recent trends of generative AI.
- Consider providing training for employees to acquire expertise when evaluation is conducted internally by a person with relevant expertise, who is independent from the design and operation of AI management system for the Behavioral Goal 5-1.
- Define the necessary personnel and skills for the proper operation of AI management systems to foster a common understanding among businesses and specify educational content.
- Utilize case studies and best practices collected through AI-related associations and organizations for internal training.

Pay close attention to a mismatch between employee development and speed of technological change because necessary AI literacy changes as the AI technology progresses.

[Practical Examples]
[Practical Example i: Education using external education materials, etc.]

As we are a small company with a small number of employees accountable for training, we decided to use external education materials instead of creating a training program for improving AI literacy. Various education programs including online courses and textbooks offered by Coursera, the U.S. for-profit education technology organization and the Japan Deep Learning Association (JDLA), etc. as well as “MANABI-DELUXE”[59] and “MANABI-DELUXE QUEST”[60], etc. are available in and outside Japan.

We use programs based on the syllabus of the JDLA’s certification examination for measuring the achievement level of employees accountable for training. JDLA’s Deep Learning for GENERAL covers a wide range of topics from the basics of AI technology to AI ethics. We confirmed that it does not place an excessive burden on employees accountable for training because 30% of the successful applicants, which is the highest number,[61] answered that they studied for 15 to 30 hours for it in the survey of G2023#3 (administered on July 7, 2023) organized by JDLA.

We recommend all employees obtain “IT Passport”[62] because we consider that improving their digital literacy is necessary for using digital technology including AI.

We think these efforts were effective as we expected. For example, someone who fragmentarily heard about incidents of AI systems/services on the news came to feel responsible for AI risks after learning topics from the basics to ethical aspects of AI technology.

[Practical Example ii: Education using a company’s own education materials]

We are a big company engaging in the development, provision and use of AI systems/services as one of our core businesses. Although we know that there are external education materials about AI technology and ethics, since the number of recipients of our AI systems/services is high, and there are numerous benefits/risks to the society, we use our own education materials which include a substantial number of examples of cases, assuming the intended use of our AI systems/services instead of external materials for general purposes. We also use the case study materials with data[63] in the “Manabi-DELUXE-QUEST” of the Ministry of Economy, Trade and Industry for practical internal education using AI.

Although we used to have the section for AI ethics at the end of a lecture about AI technology when we first created an AI training program, we are now requiring all employees to participate in an e-learning program created only for AI ethics because the management showed more interest in AI ethics after receiving advice from a committee to which external experts were invited. This e-learning program includes lectures and review tests so that even those who are

[59] Ministry of Economy, Trade and Industry “MANABI-DELUXE Website”, https://manabi-dx.ipa.go.jp/
[60] Ministry of Economy, Trade and Industry “MANABI-DELUXE-QUEST Website”, https://dxq.manabi-dx.ipa.go.jp/
[61] Japan Deep Learning Association “Answers from successful applicants of Deep Learning for GENERAL” https://www.jdla.org/certificate/general/start/
[62] Information-technology Promotion Agency, Japan “IT Passport examination website”, https://www3.jitec.ipa.go.jp/JitesCbt/index.html
(It will include questions about generative AI starting in 2024.)
[63] Ministry of Economy, Trade and Industry “Manabi-DELUXE-QUEST, “Provision of case study materials with data”, https://www.meti.go.jp/policy/it_policy/jinzai/manabi-dx-quest.html

not familiar with AI ethics can finish the program in about one hour. We believe that association with the intended use of our AI systems/services enables effective learning in a short period of time.

[Practical Example iii: Education related to generative AI]

Most recently, we think it is necessary to train employees who handles generative AI. We are encouraging employees to take an e-learning course related to generative AI on the “Manabi-DELUXE” and also considering to use the “JDLA Generative AI Test” which is a certificate examination of JDLA for testing the skills and knowledge necessary for using generative AI appropriately by referring to the “Approaches to human resources and skills required for DX promotion in the age of generative AI”[64] and “Digital Skill Standards”[65] of the Ministry of Economy, Trade and Industry.

Behavioral Goal 3-3 [Enhancing AI management through cooperation with each other between AI business actors and divisions]:
Except when the entire processes from preparation of dataset used for learning, etc. to development, provision and use of AI systems/services are performed by a division, AI business actors are expected to pay attention to trade secrets, etc., clarify issues in the operation of AI systems/services which cannot be handled by a company or a division alone and information necessary for solving such issues and also share the information to the extent possible and reasonable while ensuring fair competition under the leadership of the management. In doing so, AI business actors are expected to agree upon the scope of information to be disclosed and enter into a non-disclosure agreement, etc. for the smooth exchange of necessary information.

[Practice Guidelines]

AI business actors are expected to take the following measures under the leadership of management:

- Identify issues in the operation of AI systems/services which cannot be solved by AI business actors alone and information necessary for solving such issues.
- Share information between AI business actors to the extent possible and reasonable while paying attention to the intellectual property rights and privacy, etc.
- Fair competition must be ensured by compliance with laws, regulations and restrictions, and AI policies of AI business actors, protection of trade secrets and limited provision of data, etc. before taking the above measures.

Since the relevant laws, regulations and restrictions are likely to include the Unfair Competition Prevention Act and Act on the Protection of Personal Information, and contracts of AI business actors are related as a matter of course, it is necessary to confirm with the person in charge of legal affairs or person in charge of risk compliance. (See “Appendix 6. “Key points to note when refereeing to the “Contract Guidelines on Utilization of AI and Data.”

If AI business actors span across multiple countries, clarify the risk chain such as data distribution and implement risk management and AI governance suitable for each phase of the development, provision and use by paying attention to the consideration given by the international community related to appropriate AI governance and interoperability (consisting of two aspects: “standard” and “interoperability between frameworks”) necessary to ensure the Data Free Flow with Trust (“DFFT”).

[Practical Examples]
[Practical Example i: Sharing information carefully with customers who are not familiar with AI]

[64] Ministry of Economy, Trade and Industry “Approaches to human resources and skills required for DX promotion in the age of generative AI” (August 2023), https://www.meti.go.jp/press/2023/08/20230807001/20230807001-b-1.pdf
[65] 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

We deliver AI systems we developed, and our customers operate AI services. The accuracy of these AI systems/serveries may deteriorate as a result of operational environment and may possibly lead to destruction of a facility or other damage. For this reason, we request our customers to conduct monitoring of output of AI systems and also teach them how to find quality deterioration.

Just requesting customers who are not familiar with AI to conduct monitoring, etc. does not work. It is necessary to take time to explain and be understood about the reasons and causes of maintenance of AI systems/services (such as change in distribution of training data and input data in operation), trends in charges of output resulting from the causes. In some cases, provision of general information is sufficient, however, even if the developers of AI systems/services consider in such way, it is important for them to encourage the customers to ask questions actively so that the developers and customers will be on the same page as much as possible. It is also important to enter into a maintenance services agreement, etc. as necessary to establish a system for receiving questions actively even after the delivery of AI systems. If relearning of AI systems/services takes place, it is important to carefully explain how the output has changed as a result of the relearning. For example, we explain about the “model failure” which is a deterioration issue that occurs when data obtained as output are used as input by AI for relearning (phenomenon in which AI repeatedly learns its own mistake, and the gradually accumulated mistake results in gradual deterioration of performance of AI systems/services), etc. as points to note during relearning.

[Practical Example ii: Execution of non-disclosure agreement for smooth information sharing]

We enter into a non-disclosure agreement based on the agreement between AI Developers and AI Providers upon the scope of information to be disclosed for the smooth information sharing described above.

[Practical Example iii: Ensuring information sharing through additional oral explanation]

AI systems developed by us are created based on specific dataset, and applying them to data not included in the dataset may lead to unfavorable results. For this reason, we not only explain about data used for learning, etc., overview and accuracy, etc. of the model used but also provide information on the circumstances and data for which our AI systems should not be used to AI Providers who intend to provide the AI systems to AI Business Users. In order to ensure that information is provided, we take time separately to explain orally and obtain a signature for the acknowledgement of receipt of explanation, in addition to notice by letter or electronic document.

[Practical Example iv: Information sharing that spans across multiple countries]

We are a company developing and providing AI having our registered office in Japan. We provide AI systems/services globally, and we consider that careful collaboration is indispensable for risk management if AI Business Users or non-business users are outside Japan. In particular, it is important to pay close attention to social differences such as culture, climate and level of acceptance of AI in each country if any.

In addition, we research the laws equivalent to the Act on the Protection of Personal Information of countries in which AI Business Users and non-business users are located, and restrictions related to data security, etc. to establish security measures, based on the research.

We also collect information about international discussions related to the DFFT which can affect our business and various frameworks related to data distribution with assistance of experts.

Column 3: Example of cases in which consideration is given to social differences such as culture, climate and level of acceptance of AI in each country

Examples of response include the Hub&Spoke model used by Microsoft (in the Spoke part, AI Champ is appointed from among provider countries and regions to include the perspectives of such countries and regions for response).[66] [67]

As an example of its multi-stakeholder engagement, these is the Global Perspectives Responsible AI Fellowship which was built in partnership with Stimson Center's Strategic Foresight Hub. The purpose of this fellowship is to invite relevant parties of the Global South Countries to various discussions related to AI.[68] [69]

Behavioral Goal 3-3-1 [Understanding current status by sharing information among AI business actors]:
AI business actors are expected to pay attention to the trade secrets, understand the current state of information sharing between AI business actors and update their understanding at an appropriate time under the leadership of the management unless they conduct all the processes from preparation of dataset for learning, etc. to use of AI systems/services by themselves.

[Practice Guidelines]

AI business actors are expected to take the following measures under the leadership of management:

- Share information such as sources of data / volume, quality and distribution of data, overview of each category of data used for development of AI systems.
- When sharing information, refer to the "Guidelines for Machine Learning Quality Management" of the National Institute of Advanced Industrial Science and Technology and other efforts for standardizing information sharing.

In doing so, AI business actors are expected to understand the following matters accurately. In order to share information between AI business actors smoothly and promote social implementation of AI technology, standardization of shared information is expected.

- Develop AI systems by taking the provision and use into consideration.
- Provide AI systems after accurately understanding under which restrictions the AI systems were developed and how they will be used as services.
- Use AI services after understanding the method of use intended by AI providers and within the scope of such intention.

Methods of information sharing/collection include the following:

- Reviewing the guidelines developed by the relevant ministries, agencies and industry groups, etc.
- Becoming a member of a group related to AI ethics and quality.
- Referring to the examples of past cases occurring in and outside Japan.
  - Referring to reports of professional institutions.
  - Participating in seminars, etc.

[66] Microsoft "The building of Microsoft's responsible AI program: Governance as a foundation for compliance" (February 2023) https://news.microsoft.com/ja-jp/2021/02/02/210202-microsoft-responsible-ai-program/
[67] 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/
[68] Microsoft "Advancing AI governance in Japan: Governing AI within Microsoft" (October 2023), https://news.microsoft.com/ja-jp/2023/10/06/231006-about-the-potential-of-ai-in-japan/
[69] 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/

- Interviewing experts, etc.

[Practical Examples]
[Practical Example i: Efforts for standardizing information sharing between AI business actors]

To determine our approaches to provision of information, we decided to pay attention to the trade secrets, understand the current state of information sharing between AI business actors and regularly update our understanding under the leadership of the management.

We discovered through information collection that various efforts are made to standardize information sharing between AI business actors. For example, the National Institute of Advanced Industrial Science and Technology released the “Guidelines for Machine Learning Quality Management” aiming to use them as criteria for social consensus for the quality of machine-learning-based systems, and we learned that the Ministry of Economy, Trade and Industry, the Ministry of Health, Labour and Welfare and the Fire and Disaster Management Agency had created a format for recording reliability assessment results in the field of plant security using these guidelines as the basis. We also learned that there is a suggestion that model cards should be introduced based on the idea that it is important to indicate the performance of AI models as the ingredient lists of foods, etc. contribute to the responsible decision-making of people.[70]

There is no standard documentation procedure for sharing information on the performance and quality of the learnt models of machine learning models, etc. at the moment, however, we will refer to various efforts instead of creating our own criteria when developing an internal system.

[Practical Example ii: Understanding current status of information sharing among AI business actors through groups related to AI ethics and quality]

We belong to a group related to AI ethics and quality and actively exchange information with other member companies regarding the appropriate approaches to provision of information on the performance of AI systems/services. Although it is important to provide sufficient information on AI systems/services to AI Business Users and non-business users, it is not appropriate to consider that providing information that is not easy to understand for people other than experts or a large volume of detailed information is appropriate because they are not always familiar with the nature and limit of AI. We also consider that it is important to communicate with many stakeholders indirectly by exchanging opinions with other companies in addition to our direct experience in order to find appropriate approaches to provision of information.

Information that should be provided by AI Developers to AI Providers includes information on data used for the development of AI systems. This information includes data sources (or open data), volume and distribution of data and overview of data in each category, etc. It is important to explain the overview of algorithm selected (or not selected) at the time of development, generated models and particularly conditions tests were conducted and what level of accuracy was achieved as a result of the tests, etc.

Although these viewpoints are not new to companies with abundant experience in the development and provision of AI systems/services, we consider that the way information is given is important. What it means is that what type of information, in how much detail is explained. Understanding the current state of information sharing between AI business actors is important for considering the overall design of AI governance, and it is meaningful to participate in a group related to AI ethics and quality.

[Practical Example iii: Collaboration between AI business actors that span across multiple countries]

We are a company engaging in AI development and provision. Since we often collaborate with other AI business actors, we place a high priority on information sharing and actively conduct a trend survey of past cases occurring in and outside Japan.

[70] Google, “Vertex AI”, https://cloud.google.com/vertex-ai

First, we refer to reports of universities and professional institutions, etc. In addition, we refer to the examples of efforts posted on the websites of leading companies introduced on these reports, etc.

We feel that most recently, the importance of information on social media, etc. is increasing. We watch the posts on social media and other online platforms, see information about seminars and encourage our employees to actively participate in the seminars, etc. closely related to us.

Furthermore, we regularly invite experts such as AI consultants who know about the recent trends and cases to receive advice on how to include the information in our strategies and what kind of actions we should take based on the information.

Behavioral Goal 3-3-2 [Encouraging daily information collection / opinion exchange for environmental/risk analysis]:
AI business actors are expected to establish rules for the development and operation of AI systems/services, collect information on the best practices and incidents, etc. and encourage internal opinion exchange on a daily basis under the leadership of the management.

[Practice Guidelines]

AI business actors are expected to take the following measures under the leadership of management:

- Establish rules and collect information on the best practices and incidents, etc. on a daily basis.
- Hold discussions with other divisions internally and involve in group activities participated by other companies even if an internal AI management team is established.

[Practical Examples]

[Practical Example i: Encouraging discussions led by the management]

Although guiding principles for AI ethics are being developed, as there is no answer for how to respect the guiding principles, we have no choice but to seek the answer, and since other companies are working in the same manner, our management encourages the person in charge at each division to collect information and exchange opinions regarding the appropriate development, provision and use of AI and instructs them to share information and opinions with other divisions at the internal discussions and study group.

Although we haven’t found a perfect solution, we came to understand the major trends by continuing these activities, and we reflect the achievements from the activities in environmental and risk analysis conducted at an appropriate time.

[Practical Example ii: Encouraging discussions at small business operators]

We are a small company engaging in the development of AI systems. Since some of us consider that the priority should be given to the business growth rather than the respect for AI ethics, we decided to start with internal discussions and study groups regarding the AI ethics, which involve the legal affairs division and technical division. Since the definitions or how to use words may vary division to division, we appointed a facilitator. This allowed the discussions to go forward, and we found out that engineers who insisted that the priority should be given to the growth had read theses about fairness, etc. and that their recognition about AI ethics is not much different from others. The development processes are changing to those in consistency with AI ethics since engineers started showing interests in realizing the respect for AI ethics with technology. We wish to exchange opinions with other companies in the future.

Behavioral Goal 3-4 [Reducing burden related to incidents involving AI Business Users and non-business users through preventive and prompt action]:
AI business actors are expected to reduce incident-related burden of AI Business Users and non-business users by taking preventive and prompt action under the leadership of the management.

[Practice Guidelines]

AI business actors are expected to take the following measures under the leadership of management:

- Take preventive and prompt action for incidents such as system failure, information leak and receipt of complaints.
- Establish a system for preventing or promptly responding to incidents throughout the entire lifecycle.

Pay close attention to the following points when establishing a system for preventing or promptly responding to incidents:

- Consider preventive measures and advance preparation measures using accumulated past cases and information collected for the Behavioral Goal 3-3.
- Distribute responsibilities among related AI business actors (distribute responsibilities to those who can reduce risks).
- Promptly respond to economic loss by using insurance that covers the intended use with certain probabilistic economic loss.

[Practical Examples]

[Practical Example i: Clarifying the distribution of responsibilities]

Since AI business actors and individuals in various positions such as AI Developers, AI Providers, AI Business Users and non-business users are often involved in the development, provision and use of AI systems/services, and because of so-called the “black box” nature of AI, the attribution of responsibilities tends to be ambiguous. To prevent incidents in advance, it is important to distribute responsibilities to those who can reduce risks. Therefore, we have established a system for promptly responding to incidents by clarifying the person responsible for incidents and giving certain rights and power to such person. It is also important to improve the ability to promptly respond to incidents by preparing for incidents in advance.

[Practical Example ii: Use of insurance for incidents and continuous research and development]

We basically take actions as presented in the Practical Example i and are also considering to use insurance for part of the intended use. For the purposes of use which provide substantial benefits to the entire society but may cause economic loss due to certain uncertainty during the operation of AI systems/services, we consider that it is important to reduce burden of AI Business Users and non-business users by using insurance to promptly respond to economic loss that may be caused by incidents. As a matter of course, we recognize the importance of reducing the uncertainty of AI systems/services to continuously improve the trust from AI Business Users and non-business users and continue to engage in research and development to achieve it.

Behavioral Goal 3-4-1 [Distribution of burden of responding to uncertainty between AI business actors]:
AI business actors are expected to clarify the attribution of responsibilities to respond to the uncertainty of AI systems/services and minimize risks as a whole under the leadership of the management.

[Practice Guidelines]

AI business actors are expected to take the following measures under the leadership of management:

- Understand that although it is technically possible to respond the uncertainty of AI systems/services to some extent,[71] it is difficult to completely remove it.
- Clarify the attribution of responsibilities between AI business actors to the extent possible and reasonable.

[71] Approaches to reduce the uncertainty include taking measures at the time of development such as preparation of appropriate dataset, selection of appropriate models, verification before the start of use of AI systems and conducting examinations.

In order to clarity the attribution of responsibilities between AI business actors, execution of a contract may be effective.

Although circumstances vary for each business system/service, there is disagreement in opinions about whether any of AI business actors needs to assure the quality of AI systems/services, and the “Major precautions for referring to “Contract Guidelines on Utilization of AI and Data” attached as Appendix 6 is useful.

In addition, if the value chain/risk chain related to the development of AI and the provision of services using AI span across multiple countries, carefully consider appropriate AI governance for cross-border data flow and data localization, etc.

[Practical Examples]
[Practical Example i: Response to uncertainty through communication of information to other AI business actors]

We engage in the development of AI systems and we consider that use of AI by the relevant stakeholders will benefit the improvement in social trust. As we collect information, we discovered that there are AI Provider companies considering that AI systems/services are on an extended line of traditional software and that AI Developers should assume all responsibilities for the quality of AI systems/services. On the other hand, we found out that there are cases where AI providers themselves can determine the timing of relearning if careful explanation is given to AI Providers until they fully understand the expectation for AI systems/services. Moreover, as described in the “AI System Quality Assurance Guidelines,” we understood that the idea that “engineers, teams and organizations responsible for quality assurance need to engage in activities to deepen the understanding of customers about AI systems at the same time as development and marketing” started to gradually spread. However, since it is still difficult to change the idea that AI Developers should assure the quality, we will regularly conduct investigation about the burden of responding to uncertainty while expecting that positive impacts caused by activities such as the “AI System Quality Assurance Guidelines” will spread.

[Practical Example ii: Preparing explanation for pursuit of responsibilities]

We are an AI Provider providing AI services using AI systems developed by other companies. AI developers have executed a contract by referring to the model contract in the “Contract Guidelines on Utilization of AI and Data.”[72] According to this, AI Developers of AI systems/services (learnt model) do not warrant the completion of work, and performance and quality, etc. of achievements, on the other hand, they must perform services with care at the level higher than certain threshold. We believed that we just operate AI systems/services developed by other companies and didn’t seriously think about the importance of what kind of accountability we should meet as an AI provider if there is any inappropriate case in connection with the operation of AI systems/services or if we are requested by non-business users to provide explanation in other circumstances.

However, we changed our policy and decided to help reduce risks as an AI Provider and provide explanation as necessary with the cooperation of AI Developers after we realized that, aside from the attribution of the final legal responsibilities, since we provide services to AI Business Users directly, we cannot be relieved of any and all responsibilities to at least primarily respond to requests from AI Business Users regarding the AI systems/services operated by us and that a reputation risk will be caused to us if we cannot provide sufficient explanation.

[Practical Example iii: Response to uncertainty related to data handling]

We are planning to appoint another company to develop AI systems using data retained by us, and we wished to leave the quality assurance of data in addition to cleansing and other preprocessing of data up to another company because we don’t have much knowhow regarding

[72] Ministry of Economy, Trade and Industry “Contract Guidelines on Utilization of AI and Data” (June 2018)
https://www.meti.go.jp/policy/mono_info_service/connected_industries/sharing_and_utilization/20180615001-1.pdf

data handling. We misunderstood that if we collect and provide data currently retained by us to AI Developers of AI systems, the AI Developers who are specialized in handling data will process the data as necessary to develop AI systems/services we desire.

However, as we collect information before appointing a contractor to develop AI systems/services for us, we found out that there are points to note of AI Developers in the “Practical Guidebook on Data Provision for Fostering Human Resources of Experts in AI and Data Science”[73] which summarizes the information we can refer to when providing data between general AI business actors. The guidebook introduces the idea that “only clients who appoint others to provide services can control the quality of entrusted data before the provision” and the idea that “the benefits from the use of deliverables belong only to clients, and as a rule, responsibilities for damage caused by the use/implementation, etc. of deliverables must be assumed by clients based on the idea of responsibility for loss and responsibility for liability.”

We now understand that the details of data necessary for the development of AI systems will be determined based on the details of AI systems/services we plan to develop and that AI Developers can only respond in limited ways. We stop here and think about the burden of response to uncertainty among AI business actors, considering that it is very important part of lifecycle of development, provision and use of AI systems/services even before providing data.

[Practical Example iv “Response to uncertainty by generative AI”

We are a company developing and providing AI systems/services using generative AI to AI business Users including other countries.

First, we focus on executing a clear and fair contract regarding copyrights and other rights related to data for learning and generative model by noticing that issues are more likely to arise in generative AI in connection with copyrights and other rights. Since the data used during the development process are multi-national, we understand that there is a possibility that rights will arise based on different legal frameworks. For this reason, we listen to the opinions of experts to take an inventory of the relevant laws and regulations and risks. Moreover, we clarify the scope of responsibilities to be assumed together with AI Business Users. In doing so, we pay attention so we can act to resolve any legal issue successfully by recording the consideration process in documents to secure the transparency.

[Practical Example v: Response to uncertainty in the case of spanning across multiple countries]

In addition, we are giving consideration to AI governance related to cross-border data flow and data localization in order to handle issues when AI’s value chain/risk chain span across multiple countries. In doing so, we receive advice from experts, review the relevant laws and regulation of each country and take necessary measures depending on the details of AI services provided and possible risks caused.

We started to give consideration to a different method of data storage as a risk hedge so that we can flexibly react to changes in international restrictions. Specifically, we will develop a data center for each region to respond to legal requirements for data handling in a specific country, secure the flexibility to apply to legal changes in each country by using cloud and give consideration to distributed data processing to smoothly respond to legal environment which is different from country to country by separating data transfer and processing.

Behavioral Goal 3-4-2 [Preliminary consideration of response to incidents/disputes]:
AI business actors are expected to consider determining policies and developing plans to promptly provide explanation to AI Business Users and non-business users, identify scope of impacts and damage, organize legal relationship, and take victim relief measures, damage spread prevention measures and recurrence prevention measures, etc. promptly under the leadership of management. They are also expected to perform a practical dry run of such policies or plans as appropriate.

[73] Ministry of Economy, Trade and Industry “Practical Guidebook on Data Provision for Fostering Human Resources of Experts in AI and Data Science” (March 2021)

[Practice Guidelines]

AI business actors are expected to take the following measures under the leadership of management:

- Develop policies and plans for responding to AI incidents.
- Perform a practical dry run of the above as appropriate.

It is expected that the following systems are developed in advance in preparation for AI incidents.

- Establish a liaison office.
- Assign an officer in charge.
- Assign roles to each person in charge.
- Approaches and processes for responding to AI incidents.
- Communication system to contact related parties in companies such as the risk management division.
- Communication system to contact related parties outside companies and experts such as legal counsels.
- Processes to notify stakeholders, etc.

If AI incidents involving AI systems/services have substantial impacts to business, consider including AI incidents as a key factor for implementing the Business Continuity Plan (BCP).

[Practical Examples]

[Practical Example i: Developing a system at a small business operation in preparation of incidents]

We are a small to medium-side company providing AI systems/services. It is important to lower the possibility of occurrence of AI incidents as much as possible, however, since it is difficult to eliminate the possibility of occurrence of AI incidents, we understand that it is important to develop and implement a plan for minimizing damage incurred as a result of AI incidents.

Specifically, we have established a liaison desk, assigned an officer in charge, developed not only an internal communication system but also a communication system to contact related parties and experts outside our company in preparation of AI incidents. Although it is difficult to take all possible measures to respond to all kinds of incidents, we have categorized and organized major AI incidents that may occur in light of our AI services to some extent and developed general policies to respond to the incidents. We also perform a dry run regularly to check if the developed policies can be implemented.

[Practical Example ii: Developing a system in preparation of AI incidents through involvement of external experts]

We are a big company developing and providing AI systems/services. We have established a liaison desk, assigned an officer in charge and developed not only a system for communicating and collaborating with the risk management division, legal affairs division, public relations divisions and crisis management division but also a communication system to contact related parties and experts outside our company.

We consult with experts in advance and organize the legal responsibilities that may arise in the several patterns of possible AI incidents and conduct risk evaluation. It is useful to organize legal responsibilities and relationship of AI business actors and non-business users by category in advance because of various types of damage such as personal injury and property damage accidents, infringement of privacy, financial damage, etc. We also keep in mind that the causes for outputting abnormal results vary (abnormal algorithm, authenticity of training data and bias in training data, etc.) and that unexpected impacts are likely to occur, which should be taken into account as matters for consideration unique to AI systems/services. We make efforts to regularly update the technical and operational system to reduce impacts to business even if any unexpected event occurs.

[Practical Example iii: Developing a system in preparation of AI incidents through inclusion of AI incidents in BCP]

We have developed a company-wide Business Continuity Plan (BCP), but the business continuity may be interrupted if the AI system we operate stops. Therefore, we decided to include AI incidents in one of the triggers for BCP activation and have developed an initial response and business continuity plan in preparation of the suspension of all or part of the AI system. We also recognize that just developing a plan is not enough and that not being able to implement the plan in the event of an emergency will result in significant risk, therefore, we practice implementing the plan at least once every year.

4. Operation

Behavioral Goal 4-1 [Ensuring that the operation of AI management system is explainable]: AI business actors are expected to meet the transparency and accountability for relevant stakeholders regarding the status of operation of an AI management system, for example, by recording the deviation evaluation processes of the Behavioral Goal 3-1, etc. under the leadership of management.

[Practice Guidelines]

AI business actors are expected to take the following measures under the leadership of management:

- To keep the status of operation of an AI management system explainable to relevant stakeholders to the extent appropriate and reasonable.

The following measures are useful for improving the accountability of status of operation of AI management systems:

- Recording the status of implementation of the deviation evaluation processes of the Behavioral Goal 3-1.
- Retain records of internal/external meetings related to the development, provision and use of AI systems/services (keep them accessible to related parties other than persons in charge).
- Provide internal training related to AI.

It is effective to independently prepare a checklist for deviation evaluation processes and review and record the status of implementation of deviation evaluation processes based on the checklist.

- It is also useful to refer to the checklist attached as Appendix 7 (Separately attached) and customize it when considering.

For the purpose of providing explanation to other divisions and external parties, the descriptions in overseas documents, etc. can be helpful to make the explanation accurate and understandable as much as possible.

For example, the “Four Principles of Explainable Artificial Intelligence”[74] by NIST explains the four principles of AI and five categories of explanation.

[Practical Examples]

[Practical Example i: Ensuring that records are kept and making them available for access for explanation]

Since obtaining data and information during the “operation” leads to decision-making for improvement, we consider that the “operation” is the key for improvement through environmental and risk evaluation, etc.

We understand that not only the implementation of AI governance but also keeping records is important for further improvement and consider that it is a requirement to keep records of system design. For example, we keep records of deviation evaluation in each AI system development project, prepare overview of AI training if it’s provided, retain the minutes of internal meetings and meetings with other AI business actors regarding the development and operation of AI systems/services and make them available for access by related parties in addition to persons in charge.

---

[74] NIST, “Four Principles of Explainable Artificial Intelligence (Draft)” (August 2020), https://www.nist.gov/publications/four-principles-explainable-artificial-intelligence
As of September 2021, the five categories of explanation are 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

We are a relatively large company, and we don’t have difficulty achieving the behavioral goals related to general corporate governance. However, we are concerned that the internal organizational differentiation creates a gap in the expertise or level of understanding between the divisions and may adversely affect the collaboration between the organizations. For example, as for the service desk for inquiries established to achieve the Behavioral Goal 3-1-2, we are afraid that failure to understand technical details by the person responding to inquiries may result in delay in discovering a serious incident. Although we make efforts to improve the literacy of our employees to achieve the Behavioral Goal 3-2, for the time being, we will actively report to the management about not only the summary but also the details of inquiries from external parties.

We keep records of the deviation evaluation processes under the Behavioral Goal 3-1 accurately and in a manner that is understandable to others as much as possible for the purpose of providing explanation to other divisions and external parties and make efforts to be aware of the limits of explanation.

[Practical Example ii: Keeping records by using a checklist at a small business operator]

We are a small company engaging in the development of AI systems. The officer in charge of technology knows about all the projects, works on the programming by himself, reads and understands theses, has a lot of knowledge about AI and also has a strong interest in the issue of AI ethics. For this reason, we believe that no issue will arise due to a gap in expertise between the divisions. On the other hand, the persons involved in the projects with high expertise tend to think the behavioral goals can be achieved without checking every single time. For resolving this issue, a deviation evaluation checklist is attached to the progress report of the projects to allow the officer in charge of technology to conduct interviews as necessary.

In addition, we analyze that a gap is likely to develop between public understanding and our understanding because we are highly specialized. For resolving this issue, we try to be aware of the social acceptance by checking the status of operation and regularly sharing the circumstances learned through daily collection of information and opinion exchange, which we perform to achieve the Behavioral Goal 3-3-2.

[Practical Example iii: Using checklist across the entire AI lifecycle]

We are a company developing and providing AI systems/services. We make efforts to prevent risks in advance by using a checklist across the entire AI lifecycle.

Instead of creating our own original checklist, we use the “Appendix 7 (separately attached) Checklist” of these guidelines by customizing it to meet our unique needs. A check list created has items to be handled by each AI business actor and items that need to be handled through collaboration between AI business actors. We customized the checklist by collaborating with other divisions in our company, had discussion with AI business users who are our customers and took the entire AI lifecycle into consideration to create our own checklist.

If the checklist is expanded at random, it tends to be reduced to a formality, therefore, we pay attention to the number of items in the checklist, remove the items already well taken care of in our company and replace them with new items to refine the check items from time to time.

Behavioral Goal 4-2 [Ensuring that the operation of each AI system is explainable]:
AI business actors are expected to record the results while monitoring the pilot and full-scale operation and implementing the PDCA cycle to conduct deviation evaluation during the pilot and full-scale operation of each AI system/service under the leadership of the management. AI business actors developing AI systems are expected to support the monitoring by AI business actors providing and using AI systems.

[Practice Guidelines]

AI business actors are expected to take the following measures under the leadership of management:

- Record the results while monitoring the operation by AI business actors while implementing the PDCA cycle.

- Collaborate between AI business actors if it is difficult for each AI business actor to handle by itself.

Specifically, it is useful for AI business actors to collaborate in the following circumstances:

- AI Developer uses settings to automatically obtain input/output logs which have a substantial impact on the performance.
- AI Developer explains the specific method of monitoring to AI Provider.
- Discuss the necessity of relearning based on the output from AI systems/services.
- Exchange views between AI Developer and AI provider about the expectorations for AI systems/services.

[Practical Examples]
[Practical Example i: Recording logs of collaboration between AI business actors]

We operate AI systems/services and provide the AI systems/services to AI Business Users. We requested AI Developer to develop AI systems/services and received explanation from the person in charge of AI development including the details of dataset and behaviors of AI models to ensure not only accuracy but also fairness. This person in charge of development told us that the maintenance of AI systems/services is necessary to ensure accuracy and fairness if any difference arises between the users assumed during the development and actual users.

Since there is no employee is knowledgeable enough to interpret the code of AI systems/services, we requested AI Developer to develop AI systems/services which automatically record input and output logs that substantially affect the performance and also requested the AI Developer to teach us how to monitor the AI systems/services. After that, we created a checklist and established the management procedures to maintain the performance to achieve part of the Behavioral Goal 3-1. Currently, we conduct monitoring on a continuing bases and keep records by following these management procedures.

[Practical Example ii: Notifying the timing of relearning by collaboration between AI business actors]

We are a company developing AI systems/services provided by other companies. Although we don’t legally own AI systems/services, we assume certain responsibilities for the operation of AI systems/services under a maintenance services agreement and have the profile of AI Provider. In these circumstances, cooperation of a company operating AI systems/services on a daily basis (AI Provider) is necessary to maintain the performance of AI systems/services. As a matter of fact, the AI Provider records output from AI systems/services, determines if the quality is deteriorating based on the output, examines the actual status and reports to us. The AI Provider also participates in the meeting for discussing the necessity of relearning.

The reason why the AI Provider can determine the timing of relearning is because the AI Provider thoroughly understands what it specifically expects from AI systems/services and what they can specifically do. It is important for AI Developer to understand the expectations of AI Provider for AI systems/services and carefully explain why they can do until AI Provider fully understands. As described in the “AI System Quality Assurance Guidelines,” it is important that “engineers, teems and organizations in charge of quality assurance engage in activities to deepen the understanding of customers about AI products together with the development and marketing teams.”[75]

Behavioral Goal 4-3 [Considering proactive disclosure of AI governance practices]:
AI business actors are expected to classify information on setting AI governance goals, and establishment and operation of AI management systems, etc. as non-financial information of the Corporate Governance Code and disclose such information. Non-listed companies are also expected to consider disclosing information on activities related to AI governance. If they determine not to disclose the information as a result of the consideration, they are expected

[75]Consortium of Quality Assurance for Artificial-Intelligence-based Products and Services “AI Product Quality Assurance Guidelines ver. 2023.06” (June 2023)

to announce the fact that the information will not be disclosed along with the reasons to stakeholders.

[Practice Guidelines]

AI business actors are expected to take the following measures under the leadership of management:

- Consider ensuring the transparency of information related to AI governance including the basic approaches to AI, and establishment and operation of their AI management systems.
- Consider classifying the information as non-financial information of the Corporate Governance Code when disclosed.
- If the information is not disclosed, announce the fact that it will not be disclosed along with the

Specifically, information on AI that is expected to be disclosed includes the following.

It is considered that releasing such information externally will lead to an increase in trust, brand recognition and improvement of awareness, etc.

- Basic approaches/polices for AI
- Efforts related to AI ethics
- AI governance

[Practical Examples]

[Practical Example i: Disclosing AI governance goals on website, etc.]

We are a small company engaging in the development of AI systems. As we consider that the development of AI systems is not merely a technical activity and that it has to be supported by deep understanding of the society, we put priority on spreading this thinking internally rather than expressly setting AI governance goals. Our customers and shareholders support this perspective. Although we believe that it is important to respect the “common guiding principles” of these guidelines, the most important thing is to understand the philosophy behind them, etc.

Since we are an unlisted company, the Corporate Governance Code does not apply to us, but we actively share our thinking about AI described above on our website, etc. Our potential customers and non-business users of our AI systems/services consider that AI systems/services are not technical tools but they are sociotechnical tools, which differentiates us from other companies.

[Practical Example ii: Considering to include in non-financial information]

The Company is a listed company developing AI systems. Appropriate AI development is our important mission, and we have already established our AI policy and developed a system for implementing the policy. We have also announced these activities on our website and to the press. On the other hand, although we considered sending messages about these activities from the management, we haven’t been able to send such messages because our AI-related business does not directly affect mid to long-term revenue at this moment.

In these circumstances, we received a survey about corporate governance from an institutional investor, which includes a question about the handling of AI ethics. If the intention of investors to invest in mid to long-term development is reflected in this kind of survey, we can guess that information related to AI ethics is necessary to determine whether or not a company will make sound development. We are planning to include information on our efforts related to AI ethics in our annual report and actively send information from the management.

5. Evaluation

Behavioral Goal 5-1 [Verifying AI management system functions]: AI business actors, under the leadership of management, are expected to request individuals with relevant expertise independent from the design and operation of AI management systems to evaluate whether or not the AI management system, such as the deviation evaluation process, is appropriately designed and operated in light of the AI governance goals, that is, whether or not the AI management system is functioning properly to achieve the AI governance goals through the implementation of Behavioral Goals 3 and 4.

[Practice Guidelines]

AI business actors are expected to take the following measures under the leadership of management:

- Management clarifies the key points of evaluation for continuous improvement in their own words.
- Assign individuals with relevant expertise independent from the design and operation of AI management systems.
- Since the factors causing risks change, it is necessary to regularly review risk controls and risk management methods in the risk-based approach.
- The above individuals monitor whether or not the AI management system is functioning properly.
- Based on the results of monitoring, make continuous improvements.

Specifically, individuals with relevant expertise independent from the design and operation of AI management systems are assumed to be the following persons:

- In the case of in-house internal audit
  - Internal audit department
  - Self-audits, etc. by adding AI Developers who are not involved in the audited work to the AI management system
- In the case of using external resources
  - External auditors and international organizations, etc.[76]
    - Those who are able to utilize and apply high level of expertise and audit experience of other companies

In each case, it is important to pay close attention to the following points:

- In the case of in-house internal audit
  - Measures should be taken to enhance effectiveness, such as requiring reporting to the department in charge of risk management or the officer in charge of AI governance (the person in charge of auditing who is immediately above the officer).
  - It should be ensured that the evaluation does not become superficial because the internal audit department is not familiar with AI, such as assigning individuals who can understand the technical aspects of AI in the internal audit department, and making each department cooperate in auditing by the internal audit department.
    - For example, those pointed out in the audit are biased toward the operational processes that are easy to see, and the design and development processes are few.
- In the case of using external resources
  - External auditors, etc. do not necessarily have detailed information on the issues specific to each AI business actor or the specific circumstances, etc. of each AI business actor. Therefore, it is important for each business operator to collect information on social acceptance and engage in dialog, etc. with stakeholders voluntarily, rather than entrusting it to external auditors, etc.

[76] World Economic Forum, “The Presidio Recommendations on Responsible Generative AI” (June 2023)

- There is a high need to use external resources when there is a need to explain to relevant stakeholders whether the AI management system is functioning properly. In doing so, it is necessary to clarify the scope of assessment and reporting required depending on what management criteria, evaluation criteria in which country. Then, it is important to select external resources that have the expertise to conduct the evaluation.

The standards[77] for management and auditing organizations are currently being discussed internationally, and it is expected that trends will be monitored.

[Practical Examples]
[Practical Example i: Monitoring through the internal audit department]

The Company has an independent internal audit department that audits the operation, etc. of internal regulations before the introduction of the AI management system. When the AI management system was introduced, the scope of operations of the internal audit department was expanded to include the AI management system. With the cooperation of each department, the Company’s person in charge of internal audits investigates and confirms whether the organization, regulations, etc. are properly operated and functioning effectively, and if inappropriate operation or dysfunction is observed, requests improvement from the relevant department, and shares best practices from other departments, if any.

Social acceptance of AI systems and services has been changing. The Company believes that it is important to make improvements in line with social acceptance, and conducts internal audits mainly in areas with high expectations from society and areas with a large number of incidents reported, while referring to environmental and risk analysis. In order to obtain cooperation from each department for improvement, we select high-risk areas instead of conducting strict conformity assessments of all areas in accordance with internal rules, etc. It is easy to obtain cooperation from each department if you explain the reason for selection.

[Practical Example ii: Monitoring using self-audit]

We are a small company engaging in the development of AI systems. The Company does not have an internal audit department to evaluate the AI management system, and self-audits are conducted by members of the development department who are not directly involved in the AI management system. Since the first line of audit, which is self-audit, tends to be generous to members of its own organization, the results of self-audit are reported to the person in charge of auditing who is immediately above the officer in charge of AI governance, the content of the report is organized and reported to the officer in charge of AI governance. The officer in charge of AI governance is well versed in AI technology and ethics, so we believe that it is functioning well despite being self-audited. Currently, we are considering holding cross-departmental feedback meetings to share audit results and exchange opinions in order to reinforce third-party perspectives and communicate that internal audits are for the improvement of AI systems.

[Practical Example iii: Monitoring combining internal and external audits]

The Company has an internal audit department, but decided to use external audits for the AI management system. In external audits, we expect a high level of expertise and the horizontal development of audit experience of other companies. Social acceptance of AI systems has been changing, and market sentiment has not been formed. Even if we are proud that we are responding adequately in our own way, there may be blind spots.

External audit services are mainly provided by consulting firms, etc. By undergoing audits by external experts, it is possible to receive advice that utilizes expert information from both inside and outside the company. In addition, we expect the third party’s view and objectivity of the advice of external experts to have the effect of facilitating feedback within the company.

[77] Information technology — Artificial intelligence — Management system as the management standard, and ISO/IEC42006 Information technology — Artificial intelligence - Requirements for bodies providing audit and certification of artificial intelligence management systems as the standard for auditing organizations, are under discussion.

On the other hand, we are worried about the possibility of being passive. External experts are not always well versed in issues, etc. unique to each business. In order to make the most of the advice of external experts, it is important to actively understand the social acceptance of AI even if we rely on external audits.

Behavioral Goal 5-2 [Considering opinions of outside stakeholders]: Under the leadership of management, AI business actors are expected to consider seeking opinions from stakeholders regarding AI management systems and their operation. If, as a result of the consideration, it is determined that the content of the opinion will not be implemented, it is expected that the reason therefor will be explained to stakeholders.

[Practice Guidelines]

AI business actors are expected to take the following measures under the leadership of management:

- Consider seeking opinions from stakeholders on the AI management system and its operation.
- If the content of the said opinion is not carried out, the reason therefor will be explained to stakeholders.

In addition, in order to collaborate with stakeholders, it is important to build a network through the following initiatives, and to be able to obtain advice on a daily basis according to the company’s circumstances.

- Hold in-house training with external lecturers.
- Form a loose network and exchange information outside of work with people who are highly interested in AI ethics and quality
- Actively utilize opportunities such as conferences and exchanges of opinions on AI ethics and quality.
- Establish an organization consisting of experts in AI and other fields, including external experts on AI governance.

[Practical Examples]

[Practical Example i: Consideration through an organization including external experts on AI governance]

The “Appropriate Cooperation with Stakeholders Other Than Shareholders” chapter of the Corporate Governance Code states that companies should endeavor to appropriately cooperate with stakeholders, including employees, customers, business partners, creditors and local communities. In particular, it is important for the board of directors and the management to exercise their leadership in establishing a corporate culture where the rights and positions of stakeholders are respected and sound business ethics are ensured. In addition, due to growing interest in the appropriate development, provision, and use of AI systems and services, not only listed companies but also unlisted companies may be required to cooperate with stakeholders when evaluating and reviewing AI governance and AI management systems.

The Company believes that the initial setting including the setting of the AI policy and the creation of a system for achieving the AI policy should be done by the company itself, and that subsequent improvements should also be made by the company itself. However, the Company also places importance on cooperating with stakeholders in order to understand “how society sees it.” The Company has already set forth the AI policy and announced the meaning of the AI policy and activities to achieve the AI policy. However, we believe that it is necessary to know “how society sees it” and to ensure objective ethics. For the purpose of repeated dialog with stakeholders, we will establish an organization that includes external experts on AI governance, composed of experts in AI and other fields. In addition to AI technology experts, experts in legal, environmental and consumer issues are also invited. Since it is not enough to receive general comments, the Company is devising ways to present our specific issues and gain deep insights.

[Practical Example ii: Consideration utilizing opportunities for exchanging opinions]

There is a tendency to focus on “visible measures” such as the establishment of an external expert committee as in Practical Example i, but we believe that such a place is not the only option. What is important is to connect gently to the network of people who are highly interested in AI ethics and quality, and to be part of the information network. The Company’s management is encouraged to actively speak out at a place to exchange opinions on AI ethics and quality, and to actively take on the role of speakers at conferences and other events. Of course, such activities are included in the performance evaluation.

There is a concern that such an approach does not gather opinions. The reason for this concern is thought to be that Japanese people do not speak honestly at places for exchanging opinions and conferences. But the so-called “active sonar” people, who draw others’ opinions by expressing their opinions, know that there are people who give personal opinions after an exchange of opinions or conference. We believe that it is important to hear such opinions.

Following this management network, we held in-house training with an external lecturer. In this training, in addition to explaining the Company’s AI governance efforts to employees engaged in AI-related work, we also asked this external lecturer to evaluate our efforts. Since this external lecturer exchanges opinions with the Company’s management on a daily basis, we were able to obtain advice tailored to the Company’s circumstances, and the participants highly evaluated the training.

Under these circumstances, we are considering the establishment of an external expert committee, but we do not feel the need at this time.

6. Environment and risk reanalysis

Behavioral Goal 6-1 [Reimplementing Behavioral Goals 1-1 to 1-3 at an appropriate time]:
Under the leadership of management, AI business actors are expected to promptly grasp changes in the external environment such as the emergence of new technologies and changes in social systems such as regulations with regard to Behavioral Goals 1-1 to 1-3, and reevaluate the AI system, update understanding, and acquire new perspectives in a timely manner, thereby improving, reconstructing or making operational improvements in the AI system. When implementing Behavioral Goal 5-2, AI business actors are expected to consider obtaining external opinions for the review of AI governance as a whole in line with agile governance, which is emphasized in these guiding principles, including environment and risk analysis, in addition to the existing AI management system and its operation.

[Practice Guidelines]

AI business actors are expected to take the following measures under the leadership of management:

- Grasp changes in the external environment, such as the emergence of new technologies, technological innovations related to AI, and changes in social systems such as regulations.
- Reevaluate, update understanding, acquire new perspectives, etc. in a timely manner, and improve, reconstruct or change the operation of the AI system accordingly.
- Make the concept of AI governance take root as an organizational culture.

With regard to social trends, it is also important to obtain external information through such means as holding regular meetings with external experts.

Although timely reanalysis varies depending on AI business actors, apart from regular (quarterly, semi-annual, annual, etc.) implementation, the following timing can be considered as candidates:

- When a serious “near-miss” case occurs
- When a serious AI incident occurs at other companies
- When there is increased social attention to a specific AI technology or AI incident
- When the regulatory environment changes socially

For example, the following means are useful in establishing a system to recognize the occurrence of serious “near-miss” cases:

- Establishing a framework that makes it easy for employees to report a “near-miss” case
  - Introduction of an anonymous reporting system, introduction of a reward system for near-miss case reporters, educational activities, etc.
- Regular risk assessment and establishing a monitoring system

In order for the system and operation of AI governance to function, the concept of AI governance should be permeated throughout the organization and taken root as a culture. For this purpose, it is important for those who belong to AI business actors to recognize their role in AI governance and have a sense of involvement that is optimal overall so as not to fall into partial optimization. Examples of initiatives for cultivating culture in AI business actors include the following:

- Introduction of a personnel evaluation system that evaluates steady daily AI governance transmission activities such as cross-organizational consortiums and community activities
- Education at the time of assignment and transfer of new employees
- Mentioning the attitude toward AI governance in the code of conduct and booklets, etc., which are essential for employees
- Regular e-learning and training, etc.

[Practical Examples]
[Practical Example i: Reanalysis in line with the opportunity to report to management]

The Company regularly analyzes the environment and risks, and reports to the management, except in the event of a serious “near-miss” case, a significant increase in public attention to a specific AI incident, or a change in the regulatory environment. The discussion on the appropriate development, provision, and use of AI systems and services is very active, but it is important to prevent AI governance fatigue through agile reanalysis and to grasp major trends in an agile manner. The opportunity to report to management is a good opportunity to look at major trends.

[Practical Example ii: Reanalysis in line with the holding of a meeting including external experts on AI governance]

The Company regularly analyzes the environment and risks as shown in the Practical Example i. However, since there are overlapping elements in the verification of AI governance and AI management systems, we include the benefits/risks that AI systems and services can bring, as well as social acceptance of the development and provision of AI systems and services to the agenda of a regularly-held organizational meeting including external experts, etc. on AI governance to which external experts are invited, so as to hear about big trends on these issues from external experts.

B. Examples of business operator’s efforts at AI governance

Examples of business operators that are promoting AI governance are introduced here.[78]

Column 4: ABEJA’s efforts for AI governance

ABEJA, a startup with approximately 100 employees that develops digital platform businesses, established the Ethical Approach to AI (EAA) as a meeting of external experts in 2019 to identify and resolve ethical issues in specific projects. We regularly consult on matters such as direction. We are also working on internal systems by appointing a person with knowledge of AI-related legal affairs, ethics, and governance to handle legal affairs as well. The company is building it to create an agile system commensurate with the scale of the startup under the CEO's leadership.

The company has established AI policies by identifying important values for each business content, rather than citing abstract fairness and transparency.

Specifically, regarding the contracted development business, the challenge was deciding what to protect as a business that receives AI development contracts from customers in a wide range of industries. Since this depends on the customer’s domain, the content should emphasize “dialogue with customers” and “exchange of opinions,” and should not focus on the commonly cited “transparency” and “fairness”. On the other hand, with regard to facial recognition services, we have identified important values according to the business content, such as “privacy”, and are taking steps to ensure that the descriptions are appropriate to the business content.

In addition, when conducting AI ethics checks for individual projects, the appointed person mentioned above conducts general ethics checks for all cases when checking non-disclosure agreements and outsourced development. And when there are issues, feedback regarding ethical issues is provided to customers through project managers or others. For cases in which it is particularly difficult to make a decision, the EAA mentioned above will be consulted.

Additionary, as an AI development contractor, the company believes that it is important to realize ethics in customers’ AI development, so provides a number of services related to AI ethics consulting, such as establishing ethics and risk management systems for customers’ AI and creating ethics checklists.

As a result of these efforts, an increasing number of customers are choosing the company because it “deals with AI ethics adequately”’ or “provides AI ethics consulting services”. It is also attracting attention as a start-up company working on.

Although many days have not passed since the publication of (the Draft of) AI Guidelines for business, we refer to them from time to time if there are any items related to the company. We also plan to refer to it when we work on building the Japanese LLM adopted by NEDO in the future.

[78] For practical examples of building AI governance, please refer to the AI Governance Association’s “Working Paper on the Implementation Status of AI Governance” (August 2024). https://uploads-ssl.webflow.com/65322a024d0afe70af851cc5/66b2b65a0a1d6c6834291501_240805-aiga-implement-wp.pd

Column 5: NEC Group’s efforts for AI governance

In 2018, NEC established the Digital Trust Business Strategy Department to create and promote Companywide strategies to incorporate the notion of respect for human rights into business operations in relation to AI utilization. In 2019, NEC formulated the NEC Group AI and Human Rights Principles (the “Companywide principles”). As a part of governance structure, a Chief Digital Officer (CDO) has been appointed as the AI Governance Officer and the relationship between the AI Governance Officer, the Risk Control and Compliance Committee and the Board of Directors has been clarified in terms of corporate governance. The company has also established the Digital Trust Advisory Council, an External Expert Council, and is actively collaborating with external parties to address AI governance as part of its management agenda (see Figure 9: Implementation Framework for AI Governance).

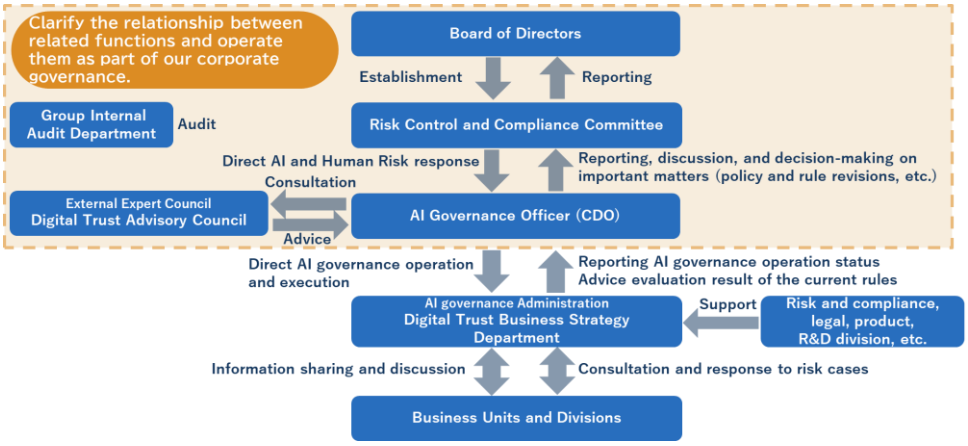


Figure 9. Implementation framework for AI Governance

The Companywide principles were developed by the Digital Trust Business Strategy Department based on domestic and international principles as well as the company’s vision, values, and business activities. The principles were formulated in April 2019, after engaging in dialogues with various internal and external stakeholders, including relevant departments within the company such as R&D, sustainability, risk management, marketing, business divisions, and external experts, NPOs, and consumers. The Companywide principles have been formulated to guide our employees to recognize respect for privacy and human rights as the highest priority in our business operations in relation to social implementation of AI utilization and consist of seven items: Fairness, Privacy, Transparency, Responsibility to Explain, Proper Utilization, AI and Talent Development, and Dialogue with Multiple Stakeholders. To incorporate the Companywide principles into business operations, the Digital Trust Business Strategy Department is taking the lead in developing internal systems and conducting employee training. Specifically, they have established Companywide rules outlining the governance structure and essential matters to be observed. They have also developed guidelines and manuals that stipulate responsive matters and operational flow, and risk check sheets. A Risk Mitigation Process has been established for conducting risk assessments and implementing countermeasures for AI utilization at each phase, starting from the planning and proposal phase.

Web-based training for all employees and internal lectures for those involved in AI business and for management teams are also conducted. In these internal lectures, external experts are invited as speakers to promote understanding, incorporating the latest market trends and case studies (see Figure 10. Overall picture of AI governance initiatives).

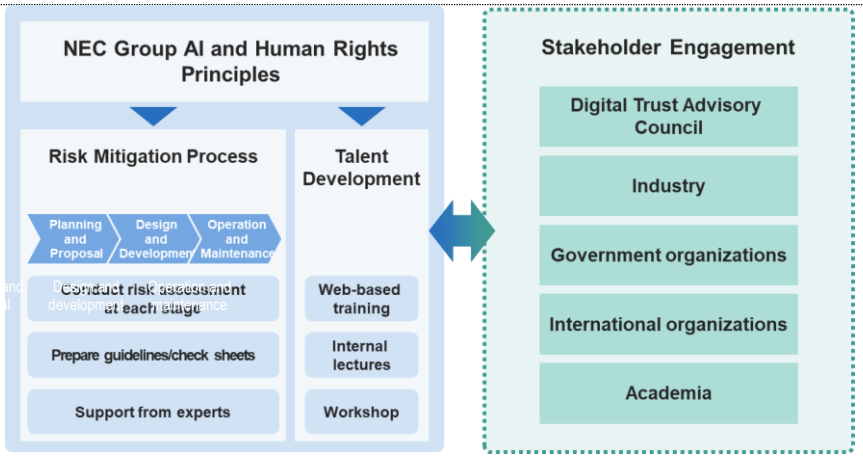


Figure 10. Overall picture of AI governance initiatives

In these initiatives, five levels of “maturity” are defined for each of the 21 action targets listed in the Ministry of Economy, Trade and Industry’s “Governance Guidelines for Implementation of AI Principles” (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. (See Figure 11. Use of 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 revising internal rules and operations. In 2023, the company established rules for internal use of generative AI (large-scale language models) and is actively utilizing generative AI.

Maturity / Action Targets	Lv.1 Performed Sporadic implementation by individuals	Lv.2 Managed Repetitive implementation following the policy	Lv.3 Defined Establishment of a unified standard process	Lv.4 Measured Execution of quantitative evaluations	Lv.5 Optimized Continuous optimization based on feedback
1-1. understand not only positive impacts but also negative impacts, including…	guidelines on AI usage for domestic and peer companies…				
1-2. understand the current state of social acceptance based on opinions of not only direct stakeholders, but…	consumer surveys and AI usage published by governments, civil groups, …				
…	…				
6-1. take other relevant actions with respect to Action Targets 1-1 through 1.3, …	response to specific incidents when significant "near-miss" events occur…				

Figure 11. Use of the former “AI Governance Guidelines”

Column 6: Toshiba Group’s efforts for AI governance

In 2022, the Company launched the AI-CoE Project Team to lead the group-wide AI measures, and formulated the AI Governance Statement that embodies the Group’s management philosophy system from the perspective of AI utilization. Referring to the former AI Governance Guidelines, the Company constructs AI Governance based on this statement. Within this framework, a working group consisting of experts in various fields such as privacy, security, and legal affairs centering on the AI-CoE Project Team, and representatives from the business side, has been formed to promote AI Governance.

Specifically, in addition to visualizing and promoting the utilization of AI technology assets owned by the Group through the creation of an AI Technology Catalog and developing AI human resources through its own training program, the Group is working to build a mechanism to maintain the quality of its AI systems through the development, etc. of MLOps (a mechanism for managing the lifecycle of machine learning models) and AI Quality Assurance System (see Figure 12. Overview of Group’s AI governance).

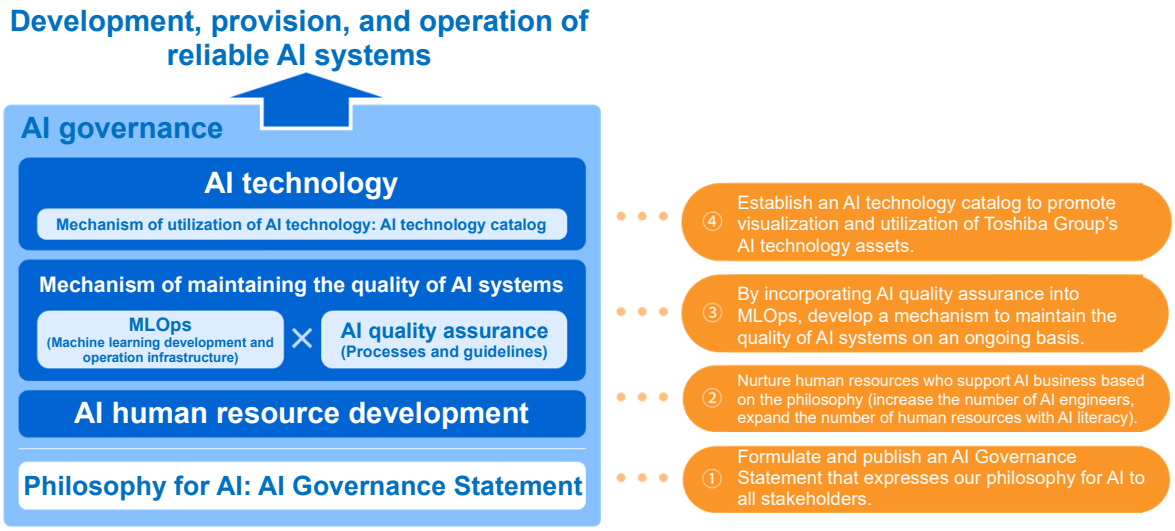


Figure 12. Overview of Group’s AI governance

This AI Governance Statement reflects the management philosophy system of Toshiba Group, and for the purpose of articulating the philosophy regarding AI, consists of seven elements: “Respect for human dignity,” “Ensuring safety and security,” “Commitment to compliance,” “Developing AI and cultivating talent,” “Realizing a sustainable society,” “Emphasis on fairness,” and “Emphasis on transparency and accountability.”

Based on this statement, a mechanism to maintain the quality of AI systems has been constructed based on the two axes of AI Quality Assurance and MLOps. In terms of AI Quality Assurance, the AI Quality Assurance Guidelines have been formulated to organize ideas and issues to be addressed in the development of AI systems, and complete processes are organized to identify necessary work and deliverables to be made in the AI Quality Assurance Process based on these guidelines. In addition, the Group is working to visualize AI quality by evaluating AI quality assurance from the user’s perspective, which tends to be seen from the developer’s perspective through Quality Cards.

MLOps brings together a team of business, machine learning experts, system developers, and systems operators. The Group is working on continuous improvement of AI systems to prevent

performance deterioration due to environmental changes after the start of operation. By linking these, the Group develops, provides, and operates reliable AI systems.

As a result of taking these AI governance initiatives, not only AI experts (engineers) but also the entire Toshiba Group has been able to improve literacy (not only opportunities for AI use, but also increased risk awareness) necessary for the development, provision, and operation of AI systems.

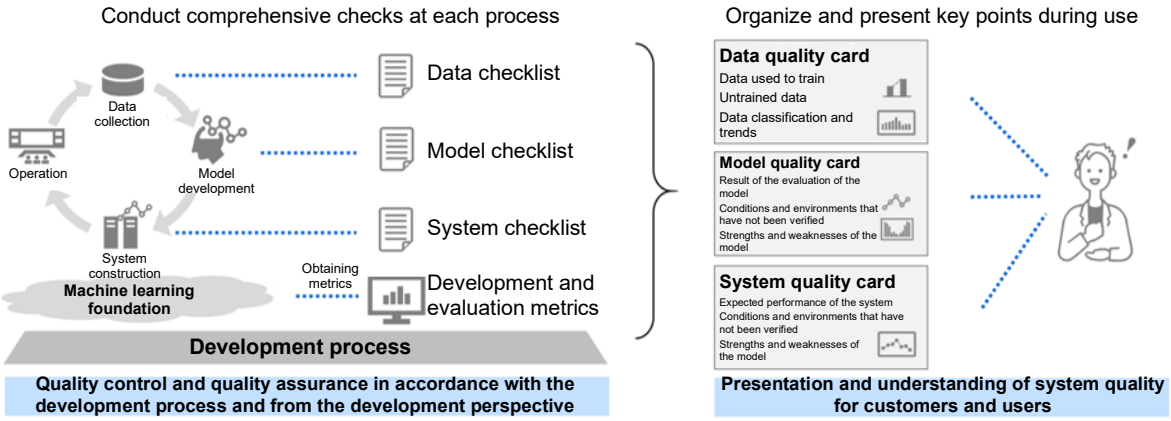


Figure 13. AI Quality Assurance Guidelines and flow of use of Quality Cards

Column 7: Panasonic Group’s efforts for AI governance

In 2019, the Company established an AI Ethics Committee within the former Panasonic Corporation and formulated the AI Ethics Principles to be observed within the company. In 2022, the organization was reorganized into the Panasonic Group AI Ethics Committee as an organization to implement the Group-wide AI Ethics Principles, and in the same year, the Group announced the Panasonic Group AI Ethics Principles. Currently, this AI Ethics Committee plays a central role in developing and utilizing the AI Ethics Check System, which will be in operation from 2022, and providing AI ethics education for all employees (see Figure 14. System and overview of AI governance).

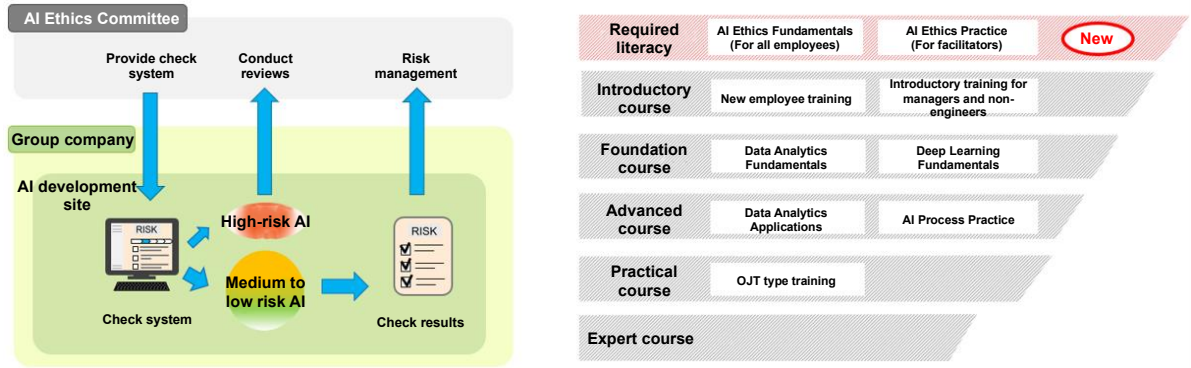


Figure 14. System and overview of AI governance

The AI Ethics Committee was established within Panasonic Holdings Corporation, and in addition to publishing the AI Ethics Principles, it engages in activities that earn the trust of users and society in a wide range of business areas. Specifically, one or more AI Ethics Officers are selected from all the Group’s operating companies, and they work together with the legal, intellectual property, information systems/security, and quality departments to establish a group-wide AI ethics promotion system (see Figure 15. Structure of the AI Ethics Committee). In order to respond to the Panasonic Group’s wide-ranging business fields, each AI Ethics Officer promotes AI ethics activities within the operating company group, and the AI Ethics Committee supports them.

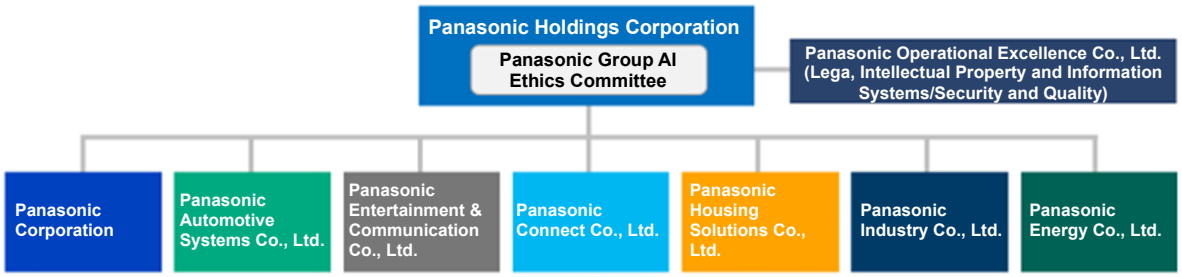


Figure 15. Structure of the AI Ethics Committee

As one of the efforts of the AI Ethics Committee, the AI Ethics Check System has been developed. This aims to efficiently and effectively conduct AI ethical risk checks while preventing the increase of on-site burden and impediment of innovation in highly-diversified and a wide range of AI utilization within the Group. It is a system that can generate necessary and sufficient checklists according to the characteristics of products and services, and it is possible to check whether the AI under development deviates from the AI Ethical Principles. In addition, for each check item, thorough explanations and information, technology, and tools

regarding countermeasures are provided, and the site is able to independently check and improve AI ethics. The results of self-checks are aggregated, analyzed by the AI Ethics Committee, and reflected in the activities. The first version of the check items was prepared in view of domestic and overseas guidelines based on the former AI Governance Guidelines of the Ministry of Economy, Trade and Industry. After actual operation, revisions are made as needed to reflect opinions from the field (see Figure 16. AI ethics check system).

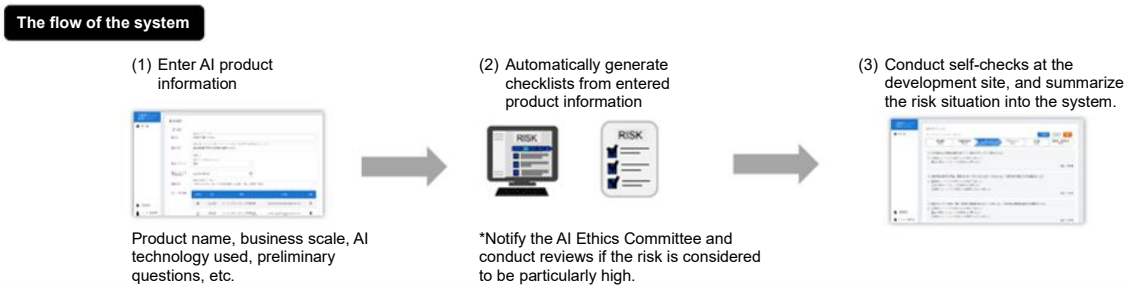


Figure 16. AI ethics check system

Column 8: Fujitsu Group’s efforts for AI governance

As a developer and provider of AI, the Company is responsible for the creation of a sustainable society by resolving concerns and unexpected inconveniences related to AI and using appropriate technologies. In addition to actively participating in international discussions on AI ethics, the Company is promoting advanced internal governance initiatives such as the AI Ethics Check described below and the appointment of AI Ethics Supervisors in overseas regions. The Company also focuses on efforts to disseminate AI ethics outside the company, such as introducing AI governance initiatives and publishing Generative AI Utilization Guidelines.

Referring to the five principles proposed by the European Consortium AI4People, which the Group joined in 2018, Fujitsu Group formulated the Fujitsu Group AI Commitment in 2019, and further developed specific criteria and procedures for making decisions according to how AI is utilized to implement the commitment (see Figure 17. Fujitsu Group Commitment). In addition, in order to obtain objective evaluations of AI governance efforts, the Fujitsu Group AI Ethics External Committee has been established. The committee invites outside experts, emphasizing diversity such as life medicine, ecology, law, SDGs, and consumer issues in addition to AI technology. The committee, in which the president and other management participate as observers, summarizes active discussions as proposals and shares them with the Board of Directors, thereby incorporating AI ethics into corporate governance as an “important issue for corporate management.”

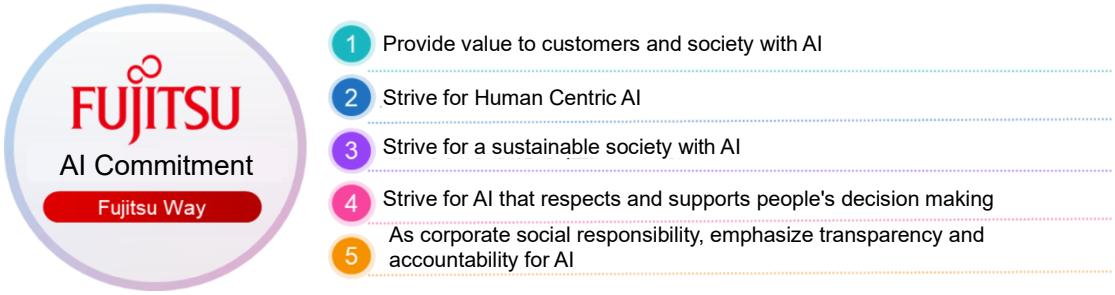


Figure 17. Fujitsu Group Commitment

As a result of institutionalizing education on AI ethics in 2020, the level of awareness of employees has dramatically improved, and it has become possible to provide advice from the ethical viewpoint to AI user companies as a consulting service.

In 2022, recognizing that AI ethics is a management issue for the entire Group, the AI Ethics Governance Office[79] was established directly under the company (Corporate Division) as an organization to lead the AI ethics strategy. At that time, individuals with experience in a wide range of occupations, such as those who have worked in development and sales, are appointed from various parts of the Group, and the Group creates a place where individual opinions are respected so that the digital native generation can play an active role. In the Office, frank exchanges of opinions and proposals take place on a daily basis, and various measures to penetrate AI ethics created from this are promoted throughout the Group (see “Figure 18. AI ethics governance system”).

[79] For details, see the white paper “Recommendations from the Fujitsu Group AI Ethics External Committee and Examples of Fujitsu Practices” on the Fujitsu AI Ethics Governance specialized website.
https://global.fujitsu/ja-jp/technology/key-technologies/ai/aiethics/governance

Figure 18. AI ethics governance system

Furthermore, the scope of the mandatory AI Ethics Review has been expanded to include all business negotiations in Fujitsu Group, and the promotion and improvement of projects with ethical issues will be determined through consultations among legal affairs, research and development, DE&I, business divisions, etc., thereby aiming at thorough governance that goes beyond the perspective of quality assurance and security related to AI. Overseas, in addition to ethical reviews at the headquarters, AI Ethics Supervisors are assigned in each region to conduct ethical reviews at the time of implementation of AI locally.

In addition, the Company is working to promote the development and provision of safe, secure and reliable AI both internally and externally through the AI Ethics Impact Assessment free of charge. In addition to the guidelines published by the Cabinet Office, the Ministry of Internal Affairs and Communications, and the Ministry of Economy, Trade and Industry, in complying with various AI ethical guidelines in Japan and overseas including guiding principles, etc. of the OECD, the EU, and the U.S., the AI Ethics Impact Assessment was formulated to extract items related to AI system developers and operators, and to evaluate the ethical impact of AI on people and society. In addition to this publication, it promotes the penetration of AI ethics initiatives throughout society through study groups with user companies, industry-academia collaboration, and standardization activities (see “Figure 19. Overview of AI ethics impact assessment”).

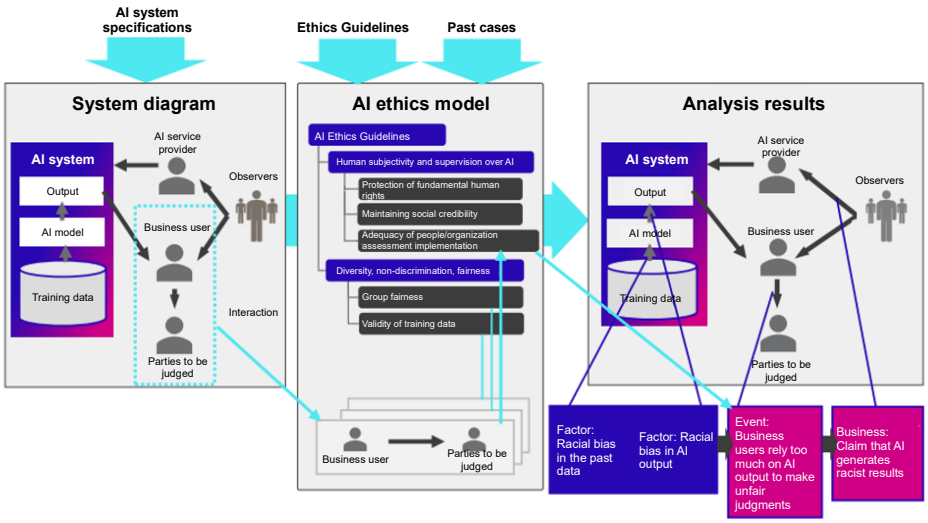


Figure 19. Overview of AI ethics impact assessment

Column 9: SoftBank's AI Governance Initiatives

Under the strategy of “Beyond Carrier”, SoftBank is working on the provision of innovative services and promotion of digital transformation (DX), utilizing cutting-edge technologies such as AI and IoT. While the use of AI is expanding, ethical considerations are required. Therefore, in July 2022, SoftBank formulated the “SoftBank AI Ethics Policy” to ensure the proper use of AI and the provision of safe and secure services. Specifically, the policy sets guidelines in six areas: “Human-Centered Principles”, “Respect for Fairness”, “Pursuit of Transparency and Accountability”, “Safety Assurance”, “Privacy Protection and Security Assurance”, and “Education for AI talent and Literacy”, and has declared to conduct business operations and service development in accordance with these guidelines. The company has also established a system to apply this policy to its group companies, and as of July 2024, 74 companies have decided to apply it. Based on the AI Ethics Policy, the company has formulated various internal rules such as regulations, standards, guidelines, and check sheets. In formulating these, the company has considered the compliance with principles such as the Cabinet Office’s “Social Principles of Human-Centric AI”, which form the foundation for the “AI Guidelines for Business Ver1.0” the Ministry of Internal Affairs and Communications’ “AI R&D GUIDELINES”, “AI Utilization Guidelines”, and the Ministry of Economy, Trade and Industry’s “Governance Guidelines for Implementation of AI Principles” (See “Figure 20: Overview of SoftBank’s AI Ethics Policy”).

Figure 20: Overview of SoftBank’s AI Ethics Policy

In promoting AI governance at SoftBank, the AI Strategy Office, the strategic arm of the AI business, is responsible for the mission, and the AI Governance Promotion Office has been established as an independent and specialized unit within the AI Strategy Office to promote the governance of the internal AI application divisions. The steering committee includes our CIO (Chief Information Officer), CDO (Chief Data Officer), CISO (Chief Information Security Officer), and CCO (Chief Compliance Officer) provide management and oversight support. The AI Ethics Committee, consisting of internal and external members, serves as an advisory board to provide advice and promote governance (See “Figure 21: AI Governance Promotion Structure”).

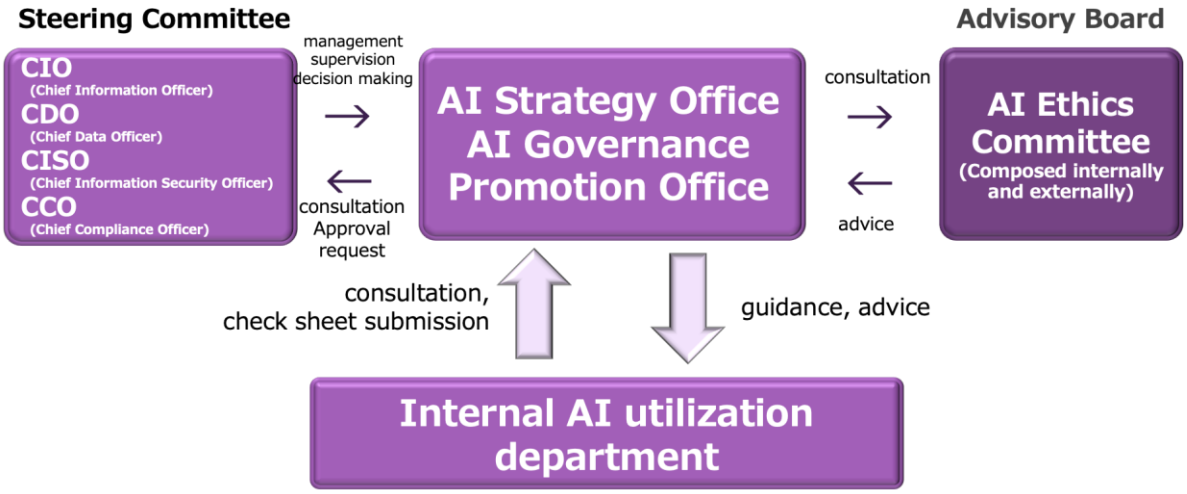


Figure 21: AI Governance Promotion Structure

The most important part of SoftBank’s strategy to promote AI governance is “promotion of AI ethics and governance education”. Specifically, SoftBank promotes education for all employees, including executives, through e-learning training once a year, online study sessions twice a year, and a monthly e-mail newsletter delivery (See “Figure 22: SoftBank’s Promotion Activities for AI Ethics and Governance Education”). The agenda of the educational content includes case studies of AI incidents in Japan and overseas, precautions to be taken when using AI including generative AI (bias, information leakage, copyright infringement, hallucination, etc.), and social trends in AI ethics, etc., to improve the literacy of all employees.

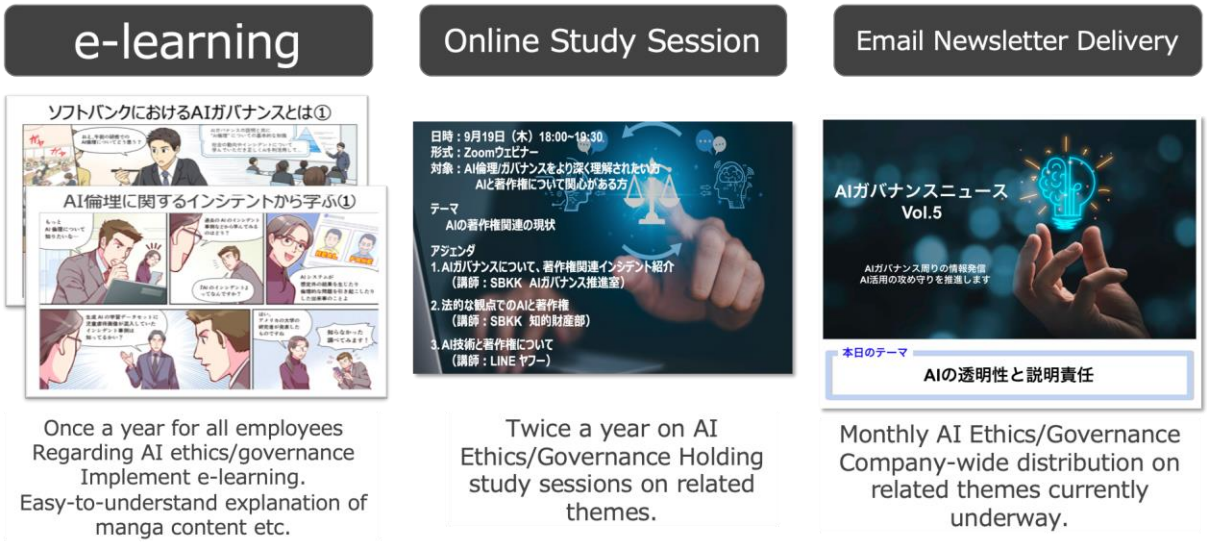


Figure 22: SoftBank’s Promotion Activities for AI Ethics and Governance Education

Column 10: NTT DATA’s AI Governance Initiatives

NTT DATA started AI governance development in 2019, with the aim of creating value and developing a sustainable society through fair and sound use of AI. The development and operation of domestic and international AI governance is promoted by the AI Governance Office, whose overall activities consist of six areas (Figure 23). This paper introduces the specific efforts related to AI Governance implementation in areas (1), (2), (3), and (6) in Figure 23.

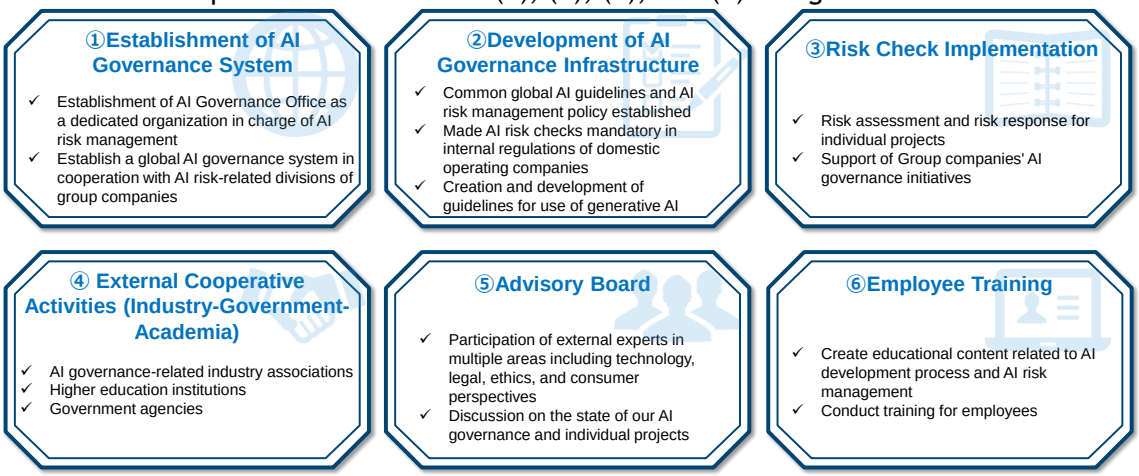


Figure 23: Activities of AI Governance

① Establishment of AI Governance System

The AI Governance Office has been established as a dedicated organization in charge of AI risk management, considering that risks arising from the use of AI (AI risks) have a significant impact regardless of the size of the business. Therefore, we have gathered experts in technology, legal affairs, intellectual property, and information security.
https://www.nttdata.com/global/ja/news/release/2023/032301/

In addition, AI risk-related contact points have been set up at each Group company to establish a cooperative system for 200,000 employees at approximately 600 companies in Japan and overseas, to develop a system that can control AI risks on a global basis.

② Development of AI Governance Infrastructure

In implementing AI governance, the following documents have been developed.

・NTT Data Group AI Guideline: A common approach for NTTDATA to use AI.
https://www.nttdata.com/global/en/news/press-release/2019/may/ntt-data-introduces-ai-guidelines

・AI Risk Management Policy: A global common policy document that defines the AI risks to be managed and the management framework (roles of each company, scope of responsibility, and implementation requirements).

・Internal rules (implementation in the project decision-making process): Internal rules that make AI risk checks mandatory when implementing development projects involving AI (AI projects).

・Generative AI Usage Guidelines: A guideline document that defines considerations and response policies for each position of developers, providers, and users with respect to generative AI. This classification of positions is the same as in the Guidelines for AI Providers.

③ Risk Check Implementation

The risk response consists of two steps. In STEP 1, all AI projects are assessed for risk (self-check) to determine whether they fall under “prohibited level”, “high risk”, or “no risk”. In STEP 2, the AI Governance Office supports risk responses for AI projects that are judged as “prohibited”

or “high risk” in STEP 1. If an unknown AI risk that is difficult to judge is detected, the AI Advisory Board, consisting of external experts, is asked to assess the AI risk and provide advice on how to respond to it.

⑥ Employee Training

Since it is important to improve employees’ AI risk literacy to ensure detection and handling of AI risks, educational contents related to AI risk management and AI development process have been created to provide training for employees. These educational contents are created for each role of business and corporate divisions that provide AI and explain AI risks that should be addressed and how to deal with them, with specific examples.

Through the above activities we were able to mitigate the incidents for about 200 AI projects from April 2023 to the end of 2024. In addition, the sensitivity of employees to AI risks has increased through training, resulting in an increase in active inquiries from employees in the early stages of AI projects.

With the knowledge gained from these practical activities, we are actively participating in discussions in private organizations on AI governance and in the formulation of voluntary guidelines. We will continue to improve our AI governance activities by self-assessing the comprehensiveness of our activities and the level of achievement with reference to the guideline for AI providers.

In addition, by leveraging this expertise in AI governance development, implementation, and risk response, we have started offering a consulting service (Figure 24) that provides total support for the establishment, operation, and improvement of AI risk management systems, as well as risk assessment and response for individual AI systems. We have already started to provide support upon receiving requests from companies in various industries, such as finance, telecommunications, and information services. By feeding back the knowledge gained through support to its own activities, we will pursue even more advanced AI governance.
https://www.nttdata.com/global/ja/news/topics/2024/073100/

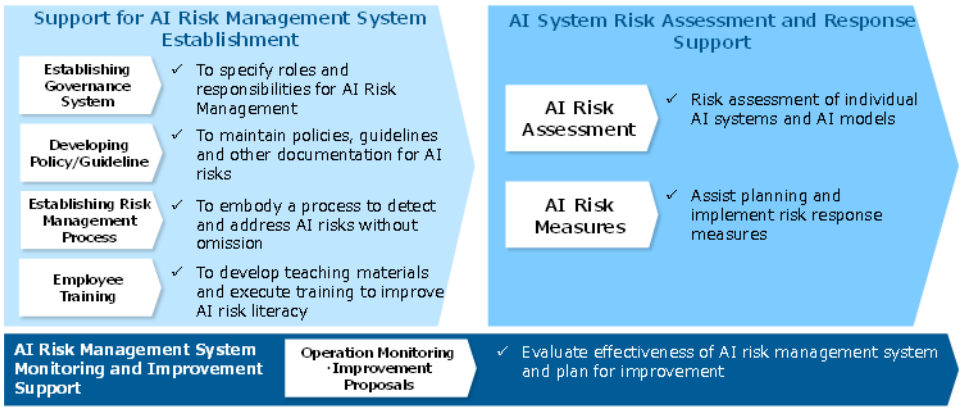


Figure 24: AI Governance Support Details

Column 11: Ubie, Inc.’s AI Governance Initiatives

Ubie, Inc., an AI healthtech startup founded by a physician and an engineer in 2017, aims to guide people to appropriate medical care by leveraging technology. Ubie, Inc. offers AI-based medical questionnaire and generative AI services for healthcare providers and AI symptom checker for consumers. As part of their AI governance framework, Ubie, Inc. has established an internal “Generative AI Utilization Promotion Team” and a “Risk and Compliance Committee.” These teams work on maximizing the value of products using generative AI and improving internal productivity, while also addressing legal and security issues related to AI utilization. (Figure 25. AI Governance Framework at Ubie, Inc.) Despite challenges in securing specialized personnel due to being a startup, Ubie, Inc. adopts a non-hierarchical organizational structure to facilitate regular information sharing and discussions between these teams. This approach helps them to respond swiftly and appropriately to rapid changes in technology and regulations surrounding AI, even with limited human resources.

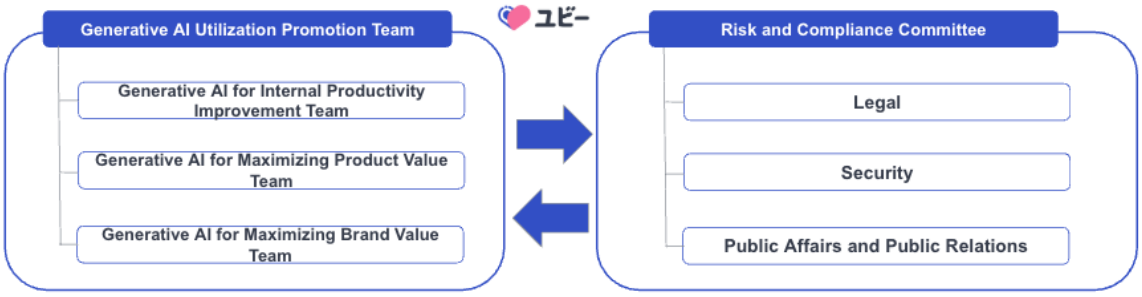


Figure 25. AI Governance Framework at Ubie, Inc.

Ubie, Inc. considers vendor risk, cloud security risk, data security risk, and internal communication guidelines in its AI risk management efforts. They conduct risk assessments and reviews from legal, security, and reputation perspectives, evaluating the likelihood and impact of risks to prioritize and implement risk countermeasures. These assessments are carried out objectively by specialists in each risk area, independent from the generative AI development and utilization teams. If there are risk concerns, they are discussed within the Risk and Compliance Committee. The company is currently working on standardizing a risk assessment framework tailored to the agile development process.

Furthermore, regardless of whether the systems are for customers or for internal use, Ubie, Inc. defines acceptable risks based on use cases and data confidentiality and implements risk countermeasures. For example, when providing generative AI services, there is a risk that highly confidential data held by client healthcare providers could be accessed or utilized by AI system developers. To mitigate this, Ubie, Inc. adopts a policy of using generative AI models that offer options to ensure that only healthcare providers have access rights to its own data and prevent its use in model training.

Additionally, the company communicates these risk management policies externally as part of its corporate stance. They have published a “For Safety and Security” page on their website, summarizing efforts to protect privacy and ensure security, positioning customer privacy protection as one of the most important management issues, and declaring a company-wide commitment to addressing privacy challenges.

Figure 26. Basic Policy on Initiatives for Safety and Security

In addition to risk management, the company actively promotes employee education to advance the use of generative AI internally. This includes onboarding sessions on generative AI basics for all employees, risk and compliance training, internal generative AI ideathons and hackathons, and the establishment of an “AI Utilization Award” in the internal recognition system to enhance employees’ understanding of generative AI. They also conduct regular surveys on the usage rate of generative AI tools available internally, with a utilization rate of 85% as of January 2024, indicating that nearly all employees are using generative AI to improve their business processes.

Furthermore, to keep up with the rapidly changing trends in AI, including generative AI, the company participates in “The Japan Digital Health Alliance,” an industry group in the healthcare sector, and leads a working group on generative AI. Through this industry group, they regularly share the latest information on generative AI and policy trends with a diverse range of member companies. In January 2024, they pioneered the formulation of the “Generative AI Utilization Guide for Healthcare Providers,” an industry guideline, and plan to release a revised version 2.0 in February 2025, contributing to rule-making in the generative AI field.

Column 12: Establishing Rules for AI Utilization in Kobe City

Kobe City, a designated city with a population of approximately 1.5 million (7th among designated cities) and about 20,000 city employees (including teachers), has established the “Ordinance on the Utilization of AI in Kobe City” (hereinafter referred to as the “AI Ordinance”) to effectively and safely utilize AI under certain rules. The ordinance applies to Kobe City and contractors or businesses undertaking city operations. In drafting the AI Ordinance, the city incorporated responsibilities that it should fulfill as an AI user, based on the AI Business Guidelines and the EU’s “Artificial Intelligence Act.” Efforts were made to replace terms from guidelines and regulations with words that are easier for city employees to understand.

The AI Ordinance stipulates that a risk assessment must be conducted when the city intends to use AI for administrative actions. It is not realistic to eliminate all risks associated with AI, so it is important for employees to correctly recognize AI risks and establish mechanisms to address them. By referencing the “risk-based approach,” the ordinance imposes careful procedures for decisions that could significantly impact citizens’ rights and interests, while simpler checks are applied to other cases, balancing AI utilization promotion with safety assurance.

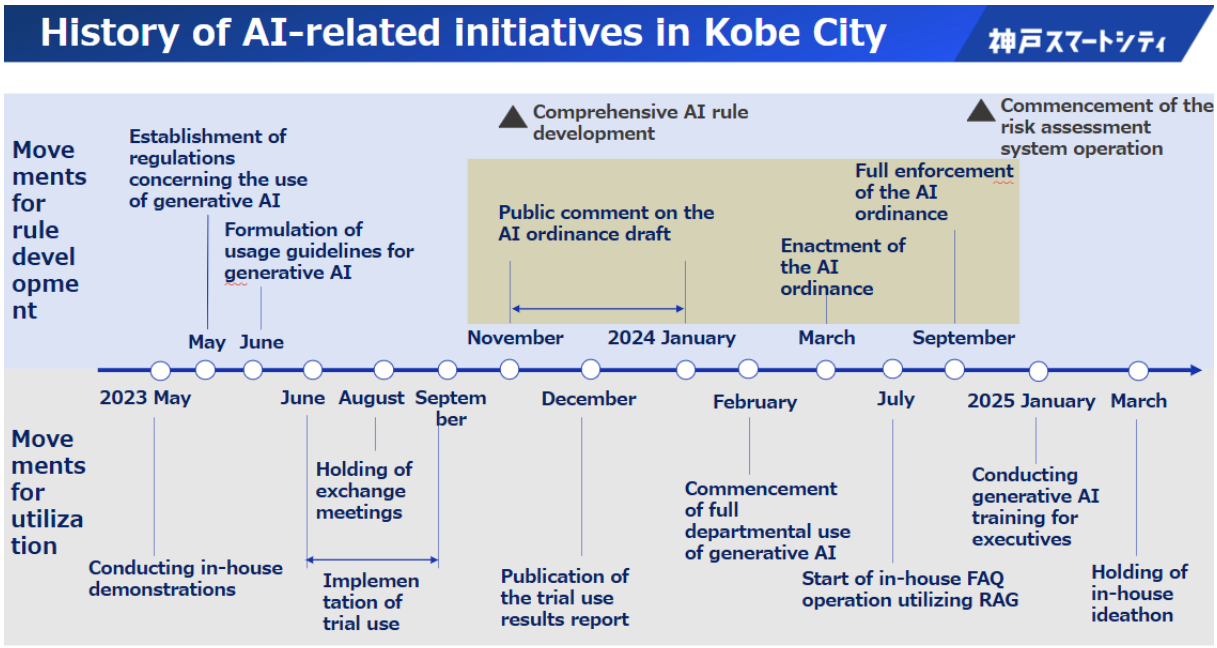


Figure 27. History of AI-Related Initiatives in Kobe City

Risk assessment items include: 1) Human-centric approach, 2) Identification of impact scope, 3) Privacy protection, 4) Safety assurance, 5) Transparency assurance, 6) Fairness assurance (bias countermeasures), 7) Security assurance, 8) Accountability assurance, 9) Employee education, and 10) Responsibility for decisions. The risk assessment method varies based on the magnitude of the risk. For matters that could significantly affect citizens’ rights and interests, the information security policy department conducts reviews based on a 48-item worksheet, while other matters are confirmed by the department head based on a checklist. The operation of evaluation criteria also involves advice from AI utilization advisors consisting of experts.

In parallel with these rule-making efforts, AI utilization by city employees in Kobe City is advancing. This includes the use of Microsoft Copilot by all employees, in-house development of various applications using generative AI, and the use of RAG for internal FAQs, as well as specific

cases like image judgment (e.g., drawing review[80] using AI technology for image analysis). As a prerequisite for these utilizations, the “Kobe City Generative AI Utilization Guidelines” have been formulated to explain compliance matters when using generative AI, and education is provided to establish rules such as the AI Ordinance and guidelines. City-wide initiatives include beginner training on “What is AI” and training for department heads on the risk assessment system based on the AI Ordinance.

[80] For details, refer to the case of “Water Supply Equipment Construction Drawing Review Using AI Review Application (Kobe City Waterworks Bureau)” in the “FY2024 Water Innovation Award Application Case Studies” by the Japan Water Works Association Water Technology Research Institute.
http://www.jwwa.or.jp/info/pdf/innovation/innovation_r6_apply.pdf

Appendix 3. For AI Developers

In this chapter, “points” and “specific methods” are explained for the contents described in the Main Part “Part 3: Matters Related to AI Developers.” After that, in “C. Common guiding principles” in the Main Part “Part 2: The Society to Aim for with AI and Matters to be Tackled by AI Business Actors,” specific methods that should be especially considered regarding AI Developers will be explained.

The “specific methods” described here is just an example. Some of them are written on both traditional AI and generative AI, or some are only applicable to either one of them. When considering specific responses, it is important to take into consideration the extent and probability of the risks posed by the AI to be developed, the technical characteristics, and the resource constraints, etc. of AI business actors.

In addition, the AI business actors developing advanced AI systems should also observe the “Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems”[81] established by the Hiroshima AI Process (description of “D. Common guiding principles for business operators involved in advanced AI systems” in the Main Part “Part 2: The Society to Aim for with AI and Matters to be Tackled by AI Business Actors”) and “Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems”[82] (see below C. Matters to be observed in developing advanced AI systems).

[81] Ministry of Foreign Affairs of Japan “Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems” (October 2023) https://www.mofa.go.jp/mofaj/files/100573469.pdf
[82] 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. Descriptions of Part 3 “Matters Related to AI Developers”

[Contents of the main part (repeat)]
<u>During data preprocessing and training</u>

D-2) i. Proper data training

- ✧ Properly collect training data through privacy-by-design, etc., and if it contains third-parties’ personal data, data requiring attention to intellectual property rights, etc., ensure that such data is properly handled in compliance with laws and regulations throughout the lifecycle of AI (“2) Safety,” “4) Privacy protection,” “5) Ensuring security”).
- ✧ Implement proper protective measures before and across training by, for example, considering the deployment of any data management and restriction function that controls access to data (“2) Safety,” “5) Ensuring security”).

[Points]

In order to improve the quality of AI models, it is important for AI Developers to pay close attention to the quality of data used for AI training, etc.

- Pay close attention to the quality (accuracy, integrity, etc.) of the data used for AI training, etc., in view of the characteristics and applications of the AI to be used.[83]
- In addition, the accuracy of decisions made by AI is assumed to be impaired or decrease after the fact. Thus, it is expected that standards for accuracy should be established in advance, taking into account such factors as the assumed scale of infringement of rights, the frequency of infringement of rights, the applicable technical level, and the cost of maintaining accuracy. If the accuracy falls below such standards, the data is trained again, paying close attention to the quality of the data.
- The term “accuracy” as used herein also includes whether the AI is making ethically correct decisions (for example, whether the AI is using violent expressions or making hate speech, etc.).

[Specific methods]

- Verify that the data does not contain personal data, confidential information, rights including copyrights or legally protected interests.
  - ➢ Extraction of unique expressions
    - ✧ Names of persons, credit card numbers, etc.
- Handle personal data, confidential information, copyrights, etc. appropriately if it contains information related to rights or legally protected interests.
  - ➢ Differential privacy
    - ✧ To add noise to data so that AI Developers do not know the actual data.
  - ➢ Data management console
    - ✧ To provide tools and consoles that allow the person who provided personal data to decide whether or not to provide personal data, withdraw consent, etc., and to easily grasp the current situation.
  - ➢ Data encryption
    - ✧ To use strong encryption algorithms to protect information when transferring and storing data.
- Implement measures to ensure that data is appropriate (quality such as accuracy and integrity is ensured) and safe.
  - ➢ Check timestamps, etc.

[83] As multimodal generative AI integrates and processes different types of data, the quality and quantity of data from each modality directly impacts AI performance. Therefore, it is crucial to apply appropriate preprocessing to the data and ensure a balanced preparation of data.

- Implement means to understand the source of the data, to the extent technically feasible and reasonable.
  - Data lineage (building provenance mechanisms)
    - To know where the data originally came from, how it was collected, managed, and moved within AI business actors over time.
    - Such data includes the identifier of the service or AI model that created the content, but it is not required to include user information.

[References]

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

[Contents of the main part (repeat)]
<u>During data preprocessing and training</u>

D-3) i. Consideration for biases included in data

- ✧ Take reasonable measures to control the quality of the data, noting that depending on the learning process of training data and AI models, there may be biases (including potential biases that do not appear in the training data) (“3) Fairness”).
- ✧ Based on the fact that biases cannot be completely eliminated from the process of training data, make sure AI models are trained with properly represented data sets and check AI systems assume no unfair bias (Bias that causes unreasonable disadvantages to specific individuals or groups without a rational explanation) (“3) Fairness”).

[Points]

AI Developers should pay close attention to the fact that AI system judgments may include biases. In addition, it is expected not to discriminate individuals and groups depending on judgements generated by AI systems. AI development based on various methods, not on sole method, is expected for it is difficult to eliminate all biases from AI models.

- Data is nothing more than a fragment of an event or phenomenon, and does not fully reflect the real world. Therefore, it should be noted that there is a risk of bias in the data and certain communities appearing under or over-represented on the data[84]. In addition, it should be checked whether there are biases and underrepresentation or overrepresentation in the underlying data.
- Because there is a possibility that prejudices and biases in the real world are latent in data, and as a result, existing discrimination may be inherited and reproduced, close attention should be paid to this in relation to “fairness.”

[Specific methods]

- Before training
  - ➢ Determination of features not to be used[85]
    - ✧ Do not allow AI models to learn about attributes that may cause prejudice or discrimination, such as race, ethnicity or gender, except in limited cases, such as checking whether an unfair bias is occurring in the AI system.
    - ✧ When deciding which attributes should not be trained by AI models, consideration should be paid to the reasons listed in Article 14, paragraph (1) of the Constitution of Japan (all of the people are equal under the law and there shall be no discrimination in political, economic or social relations because of race, creed, sex, social status or family origin) and the attributes referred to in international rules related to human rights.
    - ✧ To prevent bias from occurring because the amount of data the AI learns is too small, it should also take into account the approximate level of data required to perform the intended behavior.
  - ➢ Management and improvement of data quality
    - ✧ Reconstructing data
      - For example, delete some data and adjust the annotations so that the gender ratio of the data is appropriate for the purpose of AI development.
    - ✧ Label review
      - In the data preprocessing, it should be noted that in many cases, the labeling of the training data is created and assigned by humans, so there is a bias of the labeling person (intentionally or unintentionally).

[84] The latter is said to be a problem of “underrepresentation and overrepresentation.”
[85] Features are numerical representations of data features and are used in machine learning. For example, height and weight, etc. fall under the features of a person. These numerical values are fed to algorithms and used to train and predict models.

        - ✧ Attention to data representation
        - ✧ Compliance with ISO/IEC 27001 (Information Security, Cybersecurity and Privacy Protection - Information Security Management System - Requirements)
        - ✧ Evaluation based on ISO/IEC 25012 (Software Engineering - Software Product Quality Requirements and Evaluation (SQuaRE) - Data Quality Model)
- During training
    - ➢ Regularization with the addition of a penalty term for fairness
        - ✧ To use an optimization technique with fairness constraints
    - ➢ Implementation of Reinforcement Learning from Human Feedback (RLHF)
        - ✧ A learning process to reflect human value standards and preferences in the output of an AI model
- After training
    - ➢ Monitoring of data, learning processes and results
        - ✧ Consider reconstructing the data, such as adjusting algorithms by humans as necessary, and periodically reviewing the quality and quantity of data to be trained.
    - ➢ Implement proper data storage and access control.
        - ✧ Data encryption and secure storage
        - ✧ Compliance with ISO/IEC 27002 (Information Security, Cybersecurity and Privacy Protection - Information Security Management) regarding data storage and access
- When utilizing RAG
    - ➢ Appropriate handling of the data being referenced when utilizing RAG
        - ✧ This includes properly executing tasks such as selecting information sources, preprocessing data, chunking, and constructing vector databases.

[References]

- Digital Agency “Data Quality Guidebook (β Edition)” (June 2021)
- National Institute of Advanced Industrial Science and Technology “Guidelines for Machine Learning Quality Management, 4th Edition” (December 2023)
- Consortium of Quality Assurance for Artificial-Intelligence-based Products and Services “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 13: During data preprocessing and training Examples of consideration for biases included in data[86]

[Use case name]
Loan in 7 minutes

[Scope]
A fully automated solution that analyzes customer behavior and uses AI to make credit decisions in several minutes to make the optimal loan proposal for the customer.

[Situations handling data]
The solution interacts with internal (e.g. transaction data) and external (e.g. credit information agencies) systems to collect all the details about the customer, automatically performs risk estimation by applying algorithms based on AI and machine learning techniques, and calculates the appropriate offer for the customer.

[Implementation method]
Use of Fairness by Design[87], a development method that considers fairness from the design stage

- As a result of quantifying the weight of attributes such as income, place of employment, and transaction history that become the decision criteria for loan screening, and attributes such as age, gender and nationality related to fairness by using a participatory design method that incorporates stakeholders' opinions from the design stage, it becomes possible to develop AI models that balance business requirements and fairness. Furthermore, it incorporates algorithms to reduce cross-biases that appear when attributes such as age, gender, and nationality are combined under specific conditions as a method to remove prejudice that is not acceptable in terms of differences in culture or business practices.

Use of OSS technology by the Intersectional Fairness Project under the Linux Foundation[88] as a countermeasure to potential bias[89]

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

[86] It shows examples of consideration for biases, etc. included in data. These examples cited use cases collected in the technical report ISO/IEC TR 24030 :2024 (2024), which was formulated by the subcommittee SC 42 (ISO/IEC JTC 1/SC 42) under the technical committee ISO/IEC JTC1 established jointly by the International Organization for Standardization (ISO) whose Japan side is represented by the Japanese Industrial Standards Committee (JISC), and the IEC. (https://www.iso.org/standard/84144.html)
[87] Fujitsu "Press Release: Development of Fairness by Design, an AI development method that considers fairness from the design stage, which differs depending on culture and business practices" (March 2021), https://pr.fujitsu.com/jp/news/2021/03/31-1.html
[88] The Linux Foundation "Homepage (the world's largest and most popular open source software project)" https://www.linuxfoundation.jp/
[89] The Linux Foundation Projects, "Intersectional Fairness", https://lfaidata.foundation/projects/intersectional-fairness-isf/ Fujitsu "Press Release: Fujitsu's automated machine learning technology and AI fairness technology launched as an open source project of the Linux Foundation" (September 2023), https://pr.fujitsu.com/jp/news/2023/09/15.html

[Contents of the main part (repeat)]
When developing AI

D-2) ii. Development that takes into consideration the lives, bodies, properties and minds of humans and the environment

- ✧ Set clear policy/guidance about safe use of AI to avoid danger incurred unexpected service/use of AI by developers ("2) Safety"):
  - Requirements for not only the performance under use conditions expected under various circumstances but also the performance achievable under the use in an unexpected environment
  - Requirements for methods for minimizing risks (loss of control of a linked robot, inappropriate output, etc.) (guardrail technologies, etc.)

[Points]

AI Developers should pay close attention to ensure that the AI system does not cause harm to the lives, bodies, properties and minds of humans, and the environment by taking countermeasures as necessary in view of the nature and mode of the expected damage.

AI Developers are expected to verify and confirm the validity in advance in order to evaluate the risks related to the controllability of the AI system. As a method of risk assessment, it is possible to conduct experiments in confined spaces such as laboratories and secure sandboxes before the AI developed in society is put to practical use.

Furthermore, AI Developers should be careful to organize in advance the measures to be taken in the event of causing harm.

In addition to observing the existing laws and regulations and guidelines, AI Developers are also expected to use a new technology to respond to issues caused by a new technology.

[Specific methods]

- Requirement for the ability to withstand unforeseen environments
  - ➢ Implementation of failsafe functionality
    - ✧ Design to migrate systems for safety as priority in the event of a failure
  - ➢ Fault tolerant design
    - ✧ A design policy that can maintain functions and continue operation even if a part of the component fails or stops, such as by switching to a spare system
  - ➢ Foolproof design
    - ✧ Design to operate safely, even during incorrect operation
- Minimizing risks (loss of control of a linked robot, inappropriate output, etc.)
  - ➢ Building AI governance
  - ➢ Guardrail setting
  - ➢ Fallback design
    - ✧ A design policy that makes it possible to partially stop and reduce functions when problems occur, such as operating the system based on rules and subjecting the final judgment of a human being
  - ➢ Consideration and implementation of appropriate mitigation measures to address identified risks and vulnerabilities
  - ➢ Introduction of a phased review process
    - ✧ Prepare detailed confirmation items for the AI system
    - ✧ Review based on the confirmation items throughout the entire AI lifecycle, including before deployment and placement on the market
- Adoption of transparent development strategies
  - ➢ In order to ensure development without compromising safety, identify potential risks in upstream areas such as development design and formulate strategies to mitigate risks throughout the development process.
- Consideration of measures to be taken in the event of harm
  - ➢ Initial action
    - ✧ Take action according to the necessary procedures depending on the urgency of

the system including the AI.
- ✧ Recovery by rollback of the AI system, use of alternative system, etc.
- ✧ Stop the AI system (kill switch).
- ✧ Disconnect the AI system from the network.
- ✧ Confirmation of the details of the harm
- ✧ Reporting to relevant stakeholders

➢ (In case of serious damage) Investigation of the cause, analysis, recommendation, etc. by a third party organization

- Study of new technologies to address risks
  - ➢ Development of AI to detect and defend against new cyberattacks
  - ➢ Development of AI to remove inappropriate AI-generated products, etc.

[References]

- Ministry of Internal Affairs and Communications, “Current Status and Issues of Information Distribution in the Digital Space” (November 2023)
- Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry, “Guidebook on Corporate Governance for Privacy in Digital Transformation (DX) ver. 1.3” (April 2023)
- National Institute of Advanced Industrial Science and Technology “Guidelines for Machine Learning Quality Management, 4th Edition” (December 2023)
- Information-technology Promotion Agency, Japan “Sec journal Vol. 10 No. 3 Special Feature ‘Reliability and Safety’” (September 2014)
- Information-technology Promotion Agency, Japan “White Hacker Study Group 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 14: Examples of how guardrails can be used to minimize risk

In order to minimize the risks of AI systems, it is expected that a "guardrail" as a mechanism to control those risks should be studied. There are several types of such guardrails, and it is expected that they will be utilized according to the requirements of development.

Examples include the following:

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

For example, as a specific guardrail method, rinna Co., Ltd. provides developers with Profanity Classification API[90], an API that can be used to detect inappropriate expressions related to discrimination, cruelty, politics, religion, etc., and to monitor social media and reviews, etc. In addition, when publishing a Japanese-specific image generation model or incorporating it into the service, rinna Co., Ltd. utilized a safety check tool called Safety Checker[91] to check for inappropriate images on the generated content[92].

[90] Profanity Classification API
https://developers.rinna.co.jp/api-details#api=profanity-classification-api&operation=profanity-classification-api
[91] Safety Checker
https://github.com/huggingface/diffusers/blob/main/src/diffusers/pipelines/stable_diffusion/safety_checker.py
[92] Model card when publishing a Japanese-specific image generation model (Japanese Stable Diffusion): See the Safety Module section
https://huggingface.co/rinna/japanese-stable-diffusion

[Contents of the main part (repeat)]
When developing AI

D-2) iii. Development contributing to proper use (of AI)

- ✧ Establish clear policies and guidance on how AI can be used safely in order to avoid unexpected harm caused by the provision or use of AI (“2) Safety”).
- ✧ When giving a post-training to a pre-trained AI model, select a proper pre-trained AI model (whether a license for the commercial use is granted, pre-training data, specs required for the training and execution, and so on) (“2) Safety”).

[Points]

When developing AI systems, AI Developers are expected to cooperate with relevant parties to take preventive measures and follow-up measures (information sharing, shutdown and recovery, clarification of the cause, and measures to prevent recurrence, etc.) according to the nature and mode, etc. of damage that may be caused or has been caused by incidents that may occur or have occurred when using AI, security breaches, privacy breaches, etc.

[Specific methods]

- Guardrail setting
  - ➢ Topical Rail
    - ✧ A method to avoid topics that are not relevant to a specific use case or the intentions of AI Business Users and non-business users
  - ➢ Moderation Rail
    - ✧ A method to ensure that answers do not contain ethically inappropriate language
- Adjustment of AI models in view of the objectives
  - ➢ Characteristics of the data
    - ✧ By comparing the characteristics of the data on the new task with the data on which the original AI model was trained, consider whether the characteristics learned by the original AI model can be applied to the new task.
  - ➢ Domain of the new task
    - ✧ Confirm that the domain of the new task being fine-tuned matches the domain of the original AI model. In the case of different domains, consider adjustments such as fine tuning of only some layers.
  - ➢ Language match
    - ✧ Confirm that the original AI model matches the language of the new data. In case of differences, consider adjustments such as tokenization methods and vocabulary expansion.

[References]

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

[Contents of the main part (repeat)]
When developing AI

D-3) ii. Consideration for bias in algorithms, etc. of AI models

- ✧ Consider the possibility that bias can be included by each technical element that makes up the AI model (prompts entered by AI business users or non-business users, reference information and collaborating external services used by AI models for inference, etc.) ("3) Fairness")
- ✧ Make sure AI models are trained with properly represented data sets and AI systems assume no bias based on the fact that bias cannot be completely eliminated from AI models ("3) Fairness")

[Points]

AI Developers should note that the learning algorithms used in AI may bias the output of AI. It is expected to develop based on various methods, not on sole method, for biases cannot be completely eliminated from AI models.

In addition, in order to maintain the fairness of judgments made by AI, in light of the social context in which AI is used, the rational expectations of people, and the significance of such judgments on the rights and interests of those subject to AI-based judgments, it is expected to involve human judgment as to whether or not to use such judgments, or how to use them, etc.

[Specific methods]

- Bias detection and monitoring
  - ➢ Consideration for prompts entered by AI Business Users
    - ✧ Explain to AI Providers the necessity of concluding terms of use, etc. with AI Business Users
  - ➢ Confirmation of information and external services at the time of inference, etc.
- Review of features
  - ➢ Clarification of sensitive attributes (individual attributes such as gender and race of target persons that should be excluded from the perspective of fairness) in each business operator
    - ✧ When clarifying such attributes, consider the reasons listed in Article 14, paragraph (1) of the Constitution of Japan and the attributes referred to in international rules related to human rights.
  - ➢ Clarification of the details of fairness to be secured regarding sensitive attributes
    - ✧ Group fairness
      - Remove sensitive attributes and make predictions based only on unsensitive attributes (unawareness).
      - Ensure the same predicted results across groups with different values for sensitive attributes (demographic parity).
      - Adjust the ratio of the error of the predicted result to the actual result so that it does not depend on the value of the sensitive attribute (equalized odds).
    - ✧ Individual fairness
      - Individuals with equal attribute values other than sensitive attributes are given the same predicted result.
      - Individuals with similar attribute values are given a similar predicted result (fairness through awareness).
- Use AI models that take bias into account in machine learning models
  - ➢ Use of IPW (Inverse Probability Weighting)
    - ✧ Method to ensure equality by weighting the collected data by groups, etc.
- Achieving fairness in machine learning systems (from qualitative approach to quantitative method)
  - ➢ AI Developers should consider realizing the fairness risks analyzed by the AI Provider through quantitative fairness metrics such as "uniformity of results" from the implementation stage as necessary.
- Involve human judgment based on the social context and rational expectations of people.

- When statistical forecasting is difficult (high uncertainty);
- When there is a need for a convincing reason to make a decision (judgment), such as when it has a significant impact on a specific individual or group of people;
- When discrimination to specific individuals or groups is assumed due to the fact that the training data contains social bias against minorities (Bias based on various social attributes such as race, creed, and gender).

[References]

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

[Contents of the main part (repeat)]
When developing AI

D-5) i. Deployment of mechanisms for security measures

- ✧ Throughout the development of an AI system, take security measures appropriately based on the characteristics of the adopted technologies (security by design) (“5) Ensuring security”).

[Points]

Keeping AI security in mind, in order to ensure the confidentiality, integrity, and availability of an AI system, it is expected that reasonable measures will be taken in light of the technical level at that time. In addition, it is expected that the measures to be taken in the event of a security breach should be sorted out in advance, taking into account the applications and characteristics of the AI system, the magnitude of the impact of the breach, etc.

The security of the AI system to be developed will be ensured by considering security from the early stage of the development process, referring to the security-by-design, etc. defined by the National Center of Incident Readiness and Strategy for Cybersecurity (NISC) in “Measures to Incorporate Information Security from the Planning and Design Stages.” If security functions are added later, or the security tool is run just before shipment, there is a possibility that many returns will occur, resulting in a large development cost. If security measures are taken at the early stage of the development, there will be few returns, resulting in the creation and provision of AI system software with good maintainability.

In addition to conventional information systems, machine learning systems including LLM have assets (training data, AI models, parameters, etc.), elements such as stakeholders (AI model providers, etc.), and properties such as stochastic outputs, which require further improvement of analysis methods and management measures. Therefore, it is important to develop and apply security analysis methods and countermeasures based on the technical characteristics of machine learning.

[Specific methods]

- Security by design
  - ➢ Examples of security measures implemented
    - ✧ Threat assessment
      - Clarification of the threats and possible attacks that the software is facing, and clarification of what to protect the software from
    - ✧ Security requirements
      - The secure behavior of software itself is defined. The types of requirements include system functionality requirements, availability, maintainability, performance. Security requirements are the requirements related to security among the system requirements, and the definition of the objectives necessary to safely operate the system is established. The security requirements are described as part of the system requirements definition document or as a security requirement definition document.
      - Define security requirements by selecting an appropriate standard from the standards used in your own organization or from other frameworks, combining multiple methods to the extent that is reasonable and technically possible for your organization.
    - ✧ Security architecture
      - Provide AI Providers with the architecture information required for AI systems that incorporate the developed AI.
      - Customize and use the architecture recommended by the platform provider with AI system.
    - ✧ Software Bill of Materials (SBOM: Software Bill of Materials)
      - Create SBOMs to facilitate visibility and configuration management of the software suite embedded in the product.
    - ✧ Responsible use of Open Source Software

        - To conduct screening for responsible use
        - To clarify dependencies
        - To contribute to problem solving, development, and maintenance of open source
- Strengthening of security measures
    - Risk assessment
        - Conduct information security risk assessments, identify and prioritize risks.
            - ISO/IEC 27001: Information Security, Cybersecurity and Privacy Protection - Information Security Management System - Requirements
            - SP800-30: Guidance for conducting risk assessments
    - Access control and authentication
        - Grant the minimum necessary access rights, and employ strict authentication methods for AI Developers or administrators to access AI systems.
            - ISO/IEC 27001: Information Security, Cybersecurity and Privacy Protection - Information Security Management System - Requirements
            - SP800-53: Security and privacy control measures for organizations and information systems
        - Establish a robust insider threat detection program for content that falls under the category of intellectual property and trade secrets that are important to AI business actors.
    - Raising awareness and training
        - Provide awareness raising education and training to meet cybersecurity obligations and responsibilities based on relevant policies, procedures, contracts, etc.
            - ISO/IEC 27002: Information Security, Cybersecurity and Privacy Protection - Information Security Management Measures
            - SP800-50: Build IT security awareness raising and training programs.
    - Ensure data security
        - Encryption is used when data is in transit, and security protocols are enforced when data is stored and processed.
            - ISO/IEC 27001: Information Security, Cybersecurity and Privacy Protection - Information Security Management System - Requirements
            - SP800-53: Security and privacy control measures for organizations and information systems
    - Processes and procedures to protect information
        - Organize and establish security policies, processes, and procedures to protect information systems and assets.
            - ISO/IEC 27001: Information Security, Cybersecurity and Privacy Protection - Information Security Management System - Requirements
            - SP800-37: Guide to Applying the Risk Management Framework for Federal Information Systems: A security lifecycle approach
    - Paying attention to failures, etc. when using open source
        - If there is information such as bugs included in the open source, promptly update the open source.
            - ISO/IEC 27009: Supply chain security management
            - SP800-161: Supply chain risk management
    - Maintenance
        - Perform and record maintenance work using approved and controlled tools.
            - ISO/IEC 27001: Information Security, Cybersecurity and Privacy Protection - Information Security Management System - Requirements
            - SP800-40: Develop patch and vulnerability management programs
    - Monitoring and incident response
        - Build a monitoring system and implement an incident response process when an abnormality is detected in the AI system.
        - Properly document incidents as they arise and consider mitigating identified risks and vulnerabilities.

    - ISO/IEC 27001: Information Security, Cybersecurity and Privacy Protection - Information Security Management System - Requirements
    - SP800-61: Computer Incident Response Guide
- See “Table 6. Examples of damage and threat of systems using machine learning” as examples of attack methods.

Table 6. Examples of damage and threat of systems using machine learning[93]

Details of damage			Threats that cause damage	
			Machine learning-specific threats	Other threats
Breach of integrity or availability	System malfunction	Due to the unintended behavior of machine learning elements	Data poisoning attack	Traditional attacks against software and hardware that implement machine learning elements
			Model poisoning attack	
			Misuse of pollution models	
			Evasive attack	
		Due to other factors		Traditional attacks against systems
	Waste of computational resources	Due to machine learning elements	Data poisoning attack (resource depletion type)	Traditional attacks against software and hardware that implement machine learning elements
			Model poisoning attack (resource depletion type)	
			Misuse of pollution models	
			Sponge attack	
		Due to other factors		Traditional attacks against systems
Breach of confidentiality	Leakage of information about AI models		Model extraction attack	Traditional attacks that steal AI models
	Leakage of sensitive information contained in training data		Information leakage attack on training data	Traditional attacks that steal data
			Data poisoning attack (information embedded type)	
	Leakage of other confidential information		Model poisoning attack (information embedded type)	

[References]

- Ministry of Economy, Trade and Industry “OSS Utilization and Management Methods for Securing Security” (April 2021)
- Ministry of Economy, Trade and Industry “Guidance for Introducing SBOM for Software Management” (July 2023)
- National Institute of Advanced Industrial Science and Technology “Guidelines for Machine Learning Quality Management, 4th Edition” (December 2023)
- Information-technology Promotion Agency, Japan “Security by Design Guide Instruction Book” (August 2022)
- NCSC, “Guidelines for secure AI system development” (November 2023)
- NIST, “The NIST CYBERSECURITY FRAMEWORK (CSF) 2.0” (February 2024)
- ISO/IEC 27000 series
- NIST, SP800 series

[93] Quoted from National Institute of Advanced Industrial Science and Technology “Guidelines for Machine Learning Quality Management, 4th Edition” (December 2023)

[Contents of the main part (repeat)]
When developing AI

D-6) i. Ensuring verifiability

- ✧ Note that the prediction performance and output quality of AI may significantly change or may fail to attain the expected precision after the use of AI is started. Preserve work records for follow-up verification and take measures to maintain and improve the AI quality (“2) Safety,” “6) Transparency”).

[Points]

In order to ensure the verifiability of AI input and output, etc., AI Developers are expected to record and save logs at the time of development, and to develop AI systems in such a way that AI Providers, etc. can obtain logs of input and output.

AI Developers are expected to design and develop AI systems in a way that ensures transparency so that AI Providers can understand AI systems and provide them appropriately to AI Business Users.

[Specific methods]

- Recording and storage of logs
  - ➢ Specifically, record and store the following logs:
    - ✧ What data was used when developing AI, etc.
  - ➢ When considering the necessary “logs,” record and store appropriate logs, referring to the management system to which your organization is certified and the “documents” and “records,” etc. required by contracts. Specifically, consider the following:
    - ✧ Purpose of recording and storage of logs
    - ✧ Accuracy of logs
    - ✧ Frequency of acquisition and recording of logs
    - ✧ Time, storage period, and storage method (location, amount etc.)
    - ✧ Protection of logs
      - Ensuring confidentiality, integrity, availability, etc.
    - ✧ Scope of logs to be disclosed, etc.
- Consider the methods to improve the explainability interpretability. It should be noted that there may be trade-offs with development when considering the following:
  - ➢ Use of simple AI models
    - ✧ Choose as simple an AI model as possible to meet your requirements.
      - Logistic regression, decision trees, etc.
  - ➢ Local explanation method
    - ✧ Use a local explanation method to explain AI model predictions.
    - ✧ It is a method to explain the behavior of AI models for specific data points, such as LIME (Local Interpretable Model-agnostic Explanations).
  - ➢ SHAP value (Shapley Additive exPlanations)
    - ✧ By evaluating how much each feature contributes to the prediction of AI models based on game theory, it becomes easier to understand the relative impact of each feature.
  - ➢ Visualization of feature contribution[94]
    - ✧ Use a method to visualize the features that are important to AI models.
      - This includes feature importance plots and partially dependent plots, etc.
  - ➢ Analysis of the AI model in detail
    - ✧ Employ a method to analyze the structure and behavior of the AI model in detail.
    - ✧ Frameworks such as TensorFlow and PyTorch can also visualize the output and gradient of the model’s middle layer.
  - ➢ Choosing an AI model architecture

[94] In the case of multimodal generative AI, it becomes difficult to explain how much each modality has influenced the final decision, making it important to consider means of interpreting the model’s output results and devising ways to clarify the influence of each modality.

        - ✧ Also pay attention to the choice of an AI model architecture in order to emphasize interpretability.
    - ➢ Consideration of stakeholder participatory approaches
        - ✧ Incorporate feedback from stakeholders (e.g., AI Providers and AI Business Users) and knowledge from domain experts.
    - ➢ Introduction of watermarking that clearly indicates the use of AI when technically possible
        - ✧ In order to enable AI Business Users and non-business users to recognize that they are interacting with the AI system, consider introducing labeling, disclaimers, and other mechanisms as well.
    - ➢ Improving of transparency in the basis of outputs by introducing RAG
        - ✧ When generating responses by searching external information sources, it becomes possible to indicate sources and citations.
- Analysis of AI output trends based on the combination of multiple inputs and outputs for AI
    - ➢ For example, the observation of output changes when the input pattern is changed little by little, etc.

[References]

- National Institute of Advanced Industrial Science and Technology “Guidelines for Machine Learning Quality Management, 4th Edition” (December 2023)
- Consortium of Quality Assurance for Artificial-Intelligence-based Products and Services “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)

[Contents of the main part (repeat)]
After developing AI

D-5) ii. Consideration for the latest trends

- ✧ New attack methods to AI systems are increasing on a daily basis. In order to address those risks, considerations to be noted in each step of development should be identified ("5) Ensuring security").

[Points]

AI Developers are expected to pay close attention to the latest trends in order to deepen technical insights and implement more advanced and sustainable AI development.

By doing so, AI Developers, working with AI Providers, are expected to ensure that the AI systems are utilized as intended at the appropriate time according to the degree of risk and monitor post-deployment vulnerabilities, incidents, new risks, and misuse, thereby taking appropriate measures to address them.

[Specific methods]

- Confirmation of the latest trends through the following:
  - International conferences, journals such as arXiv
  - JVN iPedia Vulnerability Countermeasure Information Database
  - Community of developers such as social media
  - Open source project reference
  - Media coverage, etc.

[References]

- Information-technology Promotion Agency, Japan "Promotion of AI(Artificial Intelligence)"[95]
- Information-technology Promotion Agency, Japan "JVN iPedia Vulnerability Countermeasure Information Database"[96]
- Cornell University, "arXiv"[97]

[95]Information-technology Promotion Agency (IPA), Japan "Promotion of AI (Artificial Intelligence)" (Japanese website) https://www.ipa.go.jp/digital/ai/index.html
[96] Information-technology Promotion Agency (IPA), Japan "JVN iPedia Vulnerability Countermeasure Information Database" https://jvndb.jvn.jp/index.html
[97] Cornell University, "arXiv," https://arxiv.org

[Contents of the main part (repeat)]
After developing AI

D-6) ii. Providing relevant stakeholders with information

- ✧ Provide information to relevant stakeholders in a timely manner (including cases where you provide the information via AI providers) about the AI systems that you develop ("6) Transparency"). This information may include, for example, the items listed below:
  - Possibility of changes in output or programs due to learning by AI systems ("1) Human-centric")
  - Information on safety, including technical characteristics of AI systems, mechanisms for ensuring safety, foreseeable risks that may arise as a result of using the AI system, and remedies against them ("2) Safety")
  - The expected scope of use set by AI developers in which the AI can be safely used in order to prevent harm by AI provision or use unexpected during development ("2) Safety")
  - Information on the operational status of AI systems, causes of failures, and status of actions against them ("2) Safety")
  - Details of an update for AI, if any, and information on reasons for the update ("2) Safety")
  - Policies on collecting data learned by AI models, how AI models learn the data, and the system for implementing the learning ("3) Fairness," "4) Privacy protection," "5) Ensuring security")

[Points]

AI Developers are expected to explain the status of observance of the Common guiding principles for the AI systems they develop for the purpose of gaining a sense of satisfaction and peace of mind among stakeholders, as well as presenting evidence of AI behavior for that purpose (including cases where they do so via the AI Provider).

It should be noted that this is not intended to disclose the algorithm or the source code itself, but is expected to be implemented within a reasonable extent in light of the characteristics and applications of the technology to be adopted, while taking into consideration privacy and trade secrets.

[Specific methods]

- An indication that AI is being used and the scope of use of AI
- Formulation and clarification of AI policies on ethics
  - ➢ Publish ethical principles and policies and articulate the commitment of AI Developers to the Code of Ethics.
    - ✧ It also includes disclosure of policies regarding personal data, user prompts, AI system output, privacy, etc. to a reasonable extent.
  - ➢ For details, please refer to Appendix 2. Behavioral Goal 2-1 [Setting AI governance goals].
- Dialogue with stakeholders
  - ➢ Engage in dialogue while providing information on ethical initiatives and improving transparency to stakeholders through a website and other means.
    - ✧ For details, please refer to Appendix 2. Behavioral Goal 5-2 [Considering opinions of outside stakeholders].
  - ➢ Consider a mechanism to encourage relevant stakeholders to report post-deployment issues or vulnerabilities discovered to AI Developers.
    - ✧ For example, establish incentives such as a reward system for reporting incidents in order to facilitate the discovery of vulnerabilities through relevant stakeholders after deployment.
    - ✧ For details, please refer to Appendix 2. Behavioral Goal 3-4-2 [Preliminary consideration of response to incidents/disputes].

[References]

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

[Contents of the main part (repeat)]
After developing AI

D-7) i. Explanation to AI Providers of conformity to Common guiding principles

- ✧ Explain to AI providers that the prediction performance or output quality of AI may significantly change or may fail to attain the expected precision after AI starts to be used and that risks may arise as a result of this characteristic. Provide AI providers with relevant information as well. Specifically, communicate the following items (“7) Accountability”):
  - Measures against bias that technological elements forming AI models may introduce. Those elements may include training data, AI model training process, prompts assumed to be entered by AI business users or non-business users, and reference information and collaborating external services used by AI models for inference (“3. Fairness”).

[Points]

AI Developers are expected to provide meaningful and useful information to AI Providers, provide explanations according to the social context and the magnitude of the risks, and to disclose to the extent possible reports, etc. in an understandable format regarding the development content and technical evaluation.

It should be noted that this is not intended to disclose the algorithm or the source code itself, but is expected to be implemented within a reasonable extent in light of the characteristics and applications of the technology to be adopted, while taking into consideration privacy and trade secrets.

[Specific methods]

- While explaining the response status to the extent that it does not violate trade secrets, etc., if a trade-off arises, the following matters will be implemented:
  - ➢ Evaluate whether non-disclosure is acceptable and to what extent it should be disclosed from the perspective of transparency and ethics, etc.
  - ➢ Document the decision process.
  - ➢ The decision-making person should be responsible for the decision.
  - ➢ Supervise the decision appropriately and on an ongoing basis.

[References]

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

[Contents of the main part (repeat)]
After developing AI

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

- ✧ In order to improve traceability and transparency, prepare documents on your AI system development processes, data collection and labeling affecting decision-makings, algorithms you have used, and the like, as far as possible in a form that third parties can use to validate the documents ("7) Accountability").

(Note) This does not require to disclose all the documents prepared.

[Points]

AI Developers, in cooperation with stakeholders as necessary, are expected to properly document, maintain and retain the AI development process and reported incidents, etc., and be mindful of ensuring third-party verifiable status and mitigating identified risks and vulnerabilities.

[Specific methods]

- Implementation of documentation
  - ➢ Documentation of the AI development process
    - ✧ Provide a reasonable explanation on how the decision-making was conducted, starting with the source of the data, and record it as a transparency report to ensure traceability.
    - ✧ When implementing the above, keep in mind that in the event of an unexpected incident in the AI system, all people in the AI value chain may be in a position to be asked to explain something.
  - ➢ Documentation of reported incidents
    - ✧ Document incidents properly and consider mitigating identified risks and vulnerabilities.
- Documentation method
  - ➢ Regularly update these documents.
  - ➢ The form and medium of documentation will be chosen by AI business actors. It does not necessarily have to be printed.
  - ➢ It shall be available to stakeholders depending on the context of utilization.

[References]

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

[Contents of the main part (repeat)]
<u>After developing AI</u>

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

- ✧ It is expected to implement the following items as far as possible and contribute to the creation of innovation opportunities (“10) Innovation”):
  - Research and develop quality, reliability, and development methodologies, and the like for AI.
  - Contribute to the maintenance of the sustainable economic growth and the provision of solutions for social challenges.
  - Promote internationalization, diversification, and collaboration among industry, academia, and government sectors, including watching trends in international arguments, such as DFFT, and joining AI developer communities and academic societies.
  - Provide all of society with information about AI.

[Points]

Since AI Developers can directly design and modify AI models, they highly influence the output of AI in AI systems and services as a whole, and they are especially expected by society to lead innovation.

[Specific methods]

- Develop and promote information sharing standards, tools, mechanisms, and best practices to ensure the safety, security and reliability of AI systems, thereby establishing mechanisms to adopt them as needed.
  - ➢ Share best practices for improving safety and ensuring security across organizations.
  - ➢ Collaborate with stakeholders such as industry, academia, government agencies, and non-profit organizations.

B. Descriptions of "Common guiding principles" in Part 2

Although not mentioned in the Main Part, "Part 3 Matters Related to AI Developers," specific methods for the Main Part, "Part 2" "Common guiding principles," which are especially important for AI Developers, are explained here.

In addition, when requested by AI Providers or AI Business Users, AI Developers will take measures such as providing necessary information.

[Contents of the main part (repeat)] *Excerpts of the main paragraph only

1) Human-centric

When developing, providing, or using an AI system or service, each AI business actor should act in a way that does not violate the human rights guaranteed by the Constitution of Japan or granted internationally, as the foundation for accomplishing all matters to be conducted, including the matters described later. In addition, it is important that each AI business actor acts so that the AI expands human abilities and enables diverse people to seek diverse well-being.

- Regarding "(1) Human dignity and autonomy of individuals"
  [Relevant contents (repeat of contents of the main part)]
  - Based on the social context of AI use, respect human dignity and the autonomy of individuals.
  - In particular, when linking AI with someone's brain or body, refer to bioethics discussions, etc. in other countries and research institutions, together with information of peripheral technologies.
  - When profiling using AI in a field where individual rights and interests can be severely affected, use AI respecting the dignity of individuals, maintaining the accuracy of the outputs as much as possible, understanding limitations of predictions, recommendations or judgments, etc. of AI, and carefully considering possible drawbacks, and do not use it for inappropriate purposes.

  [Specific methods]
  - Establish an officer in charge of AI ethics and an internal organization related to AI governance
  - Examples of bioethics discussions in other countries and research institutions that can be referenced when developing AI are as follows:
    - ✧ Reports, etc. issued by international organizations such as the United Nations (UN), and the World Health Organization (WHO)
    - ✧ Research papers published by academic institutions such as universities
  - When profiling using AI in a field where individual rights and interests can be severely affected, it is especially useful to take into account the following points in development:
    - ✧ Minimize any potential bias in the data and algorithms used for profiling to achieve fair and equal results.
      - Monitoring the output of data and algorithms
      - Ensure that the individuals concerned have the opportunity to receive human judgment, in addition to AI judgment, etc.
    - ✧ If the data used for profiling contains personal data, handle such personal information appropriately.
      - Set up of a data clean room
      - Implementation of machine learning for privacy protection, etc.
    - ✧ Ensure that the developed AI system functions properly and that potential risks to individuals are managed appropriately.

- Regarding "(2) Paying attention to manipulations by AI on decision-makings and emotions"

- [Relevant contents (repeat of contents of the main part)]
- Do not develop, provide or use AI systems and services for the purpose of unjustly manipulating or on the precondition of manipulating human emotions at the unconscious level, such as decision-making and cognition.
- When developing, providing or using an AI system or service, pay necessary attention and take necessary countermeasures against the risk of excessive reliance on AI, such as automation bias[98].
- Pay attention to AI utilization that might instigate biased information or values and unwillingly limit the options that should be originally available to people including AI Business Users, such as a filter bubble.
- Carefully handle AI outputs, especially when they can be relevant to procedures that might significantly affect the society, such as elections and decision-making in a community.

[Specific methods]
- As a countermeasure against the risk of excessive reliance on AI for automation bias, etc., it is useful to ask AI Providers to alert AI Business Users and non-business users.
- For example, it is useful to consider serendipity (accidental or unexpected discoveries) as a countermeasure to filter bubbles.
  - ✧ Specifically, the use of a variety of information sources and the review of algorithms, etc.
- In cases where it can be relevant to procedures that might significantly affect the society, such as elections and decision-making in a community, it is useful, for example, to make final decisions by humans or to evaluate AI systems from an ethical perspective rather than from a technical perspective.

- Regarding "(3) Countermeasures against disinformation"
  [Relevant contents (repeat of contents of the main part)]
  - Generative AI has enabled everyone to forge fake information that seems to be true and fair, so recognize the increasing risk of destabilizing and confusing the society through disinformation, misinformation, and biased information generated by AI, and take necessary countermeasures.[99]

  [Specific methods]
  - For example, it is useful to indicate that the product is a product of generative AI to ensure the appropriate use of AI outputs.

- Regarding "(4) Ensuring diversity/inclusion"
  [Relevant contents (repeat of contents of the main part)]
  - In addition to ensuring fairness, to prevent information poverty or digital poverty and allow more people to enjoy the benefits of AI, pay attention to make it easy for socially vulnerable people to use AI.

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

- Regarding "(5) Providing user support"
  [Relevant contents (repeat of contents of the main part)]

[98] Refers to a phenomenon in which automated systems or technologies are excessively trusted or depended on when humans make judgments and decisions.
[99] A joint study is being conducted by the National Institute of Information and Communications Technology (NICT) and KDDI Corporation to develop a high-performance LLM capable of suppressing hallucinations, which is an issue with generative AI. For example, indicating that an output is generated by AI is useful to ensure the appropriate use of AI outputs.

- To the extent reasonable, provide information on the functions of AI systems and services and peripheral technologies, and make available the function to provide information in a timely and appropriate manner to determine opportunities for selection.

[Specific methods]
- Explanation of information on data handling
  - How to use the data input into the AI model created by the AI Developer for additional training, etc.
  - Information about the source and processing of the data used for training
- Securing of transparency of algorithms and AI models
  - Disclosure of algorithm logic, if possible
  - Examples of input/output
- Notification of changes and updates

- Regarding "(6) Ensuring sustainability"
  [Relevant contents (repeat of contents of the main part)]
  - Examine the impact of the whole lifecycle on the global environment during the development, provision, and use of AI systems and services.

  [Specific methods]
  - Adoption of lightweight AI models
    - Improve energy efficiency by using lightweight, highly resource-efficient AI models instead of large-scale, high-precision AI models in line with AI requirements.
  - Optimization of the size of AI models
    - Design AI models and develop algorithms that are conscious of the efficient use of computational resources and the minimization of energy consumption.
  - Effective use of data
    - Improve data quality, eliminate redundancy, and avoid retrieving unnecessary data.

[References]
- Ministry of Internal Affairs and Communications, "The Situation and Issues Surrounding FactChecking in Japan and the World" (May 2019)
- Ministry of Internal Affairs and Communications, "Collection of Multistakeholder Initiatives on Countermeasures Against False and Misleading Information on the Internet" (May 2024)
- EU, "Ethics Guidelines for Trustworthy AI" (April 2019)
- OIS Research Conference, "AI and Citizen Science for Serendipity" (May 2022)
- Council of Europe, "Risk and Impact Assessment Method for AI Systems from the Perspective of Human Rights, Democracy, and the Rule of Law (HUDERIA)" (November 2024)

[Contents of the main part (repeat)] *Excerpts of the main paragraph only

4) Privacy protection

It is important that during the development, provision, or use of an AI system or service, each AI business actor respects and protects privacy in accordance with its importance. At this time, relevant laws should be obeyed.

- Regarding "(1) Protection of privacy across AI systems and services in general"
  [Relevant contents (repeat of contents of the main part)]
  - Observe relevant laws, including the Act on the Protection of Personal Information, and formulate and announce the privacy policy of each AI business actor, to take measures to respect and protect the privacy of stakeholders, in accordance with its importance, based on the social contexts and legitimate expectations of people.

[Specific methods]

- Strengthening[100] security measures to protect privacy (see "Table 7. Examples of machine learning-specific threats, attack interfaces, attack execution phases, attackers, and methods for attack" for machine learning-specific attack methods).
  - Introduce appropriate encryption methods and access control mechanisms.
  - Carry out tests and fine-tuning to prevent leakage of personal data.
  - Also consider introducing privacy protection machine learning, secure machine learning, etc. for those of high importance.
  - Since there is a possibility that AI users may input personal information or other privacy-related data when using the system, a mechanism to assist in determining the presence of privacy information and ensuring its confidentiality during system input should be introduced.
- Consider introducing a data management and restriction function for controlling access to data.
  - Introduce authorization to data access.
  - Set up a data management organization.
    - Install a CDO (Chief Data Officer).
    - Appoint a Privacy Officer.
    - Dedicate resources to privacy efforts.
      - Personnel allocation, human resource development, etc.
  - Develop and disseminate data operational rules.
- Conduct a privacy assessment.
  - Privacy Impact Assessment (PIA)
    - Visualize and organize the information collected and processed by the AI system, the flow of information, and stakeholders.
    - Identify privacy risks for AI systems.
    - Determine the impact and likelihood of each risk and assess the magnitude of the risk.
    - Determine the direction of risk response (reduction, avoidance, acceptance, transfer) according to the magnitude of the risk, and formulate a response plan.
  - Quality Management Implementation Items (see "Table 8. Overview of quality management implementation items")
- Acquisition of ISO standards related to the handling of personal data
  - ISO/IEC 27001

[100] When utilizing RAG and other external services or data, the risk of unintended leakage of important information (such as personal or confidential information) increases, making privacy protection and security assurance through data anonymization and access restrictions particularly important.

- It is an international standard for information security management systems (ISMS) that focuses on the maintenance and management of information security.

✧ ISO/IEC 27701
- It describes the extended requirements for personal information management systems (PIMS) based on ISO/IEC 27001.
- It is a standard that focuses on privacy protection and can be used by AI Developers to ensure proper management of personal data.

✧ ISO/IEC 29100
- It is an international standard for privacy that provides basic principles and requirements for the protection of personal data.

✧ ISO/IEC 27018
- It is an international standard for the protection of personal data in cloud services.
- It can be used by AI Developers providing cloud services to ensure the proper handling of personal data in the cloud environment.

[References]

- Ministry of Economy, Trade and Industry “Information Security Management Standards (2016 Revised Edition)” (March 2016)
- Information-technology Promotion Agency, Japan “SEC journal Vol. 45, Preface” (July 2016)
- Information-technology Promotion Agency, Japan “How to Make Safe Websites” (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 “Guidelines for Machine Learning Quality Management, 4th Edition” (December 2023)

Table 7. Examples of machine learning-specific threats, attack interfaces, attack execution phases, attackers, and methods for attack[101]

Threats	Assets on the attack interface	Attack execution phase	Examples of attackers	Typical examples of attack techniques
Data poisoning attack	The source of the training data	When collecting and processing training datasets	External attackers	Alteration of the source of the training data
	Training datasets	When collecting and processing training datasets When developing systems	Data providers System developers External attackers	Alteration of training datasets
Model poisoning attack	Pre-learning models	When training and providing pre-learning models When developing systems	AI model providers System developers External attackers	Installation of a backdoor in pre-learning models
	Learning mechanisms	When developing systems	System developers External attackers	Malicious training programs
	Trained AI models	When developing systems When operating systems		Alteration of AI models
Misuse of pollution of models	The source of the operational input data Operational input data systems	When operating systems	System users System operators	Operational input that misuses the backdoor Observation of output information, etc. during operation (to steal information embedded in the model)
Model extraction attack	The source of the operational input data Operational input data systems	When operating systems	System users System operators	Entry of data into the system during operation Observation of output information, etc. during operation
Evasive attack Sponge attack	Trained AI models	When obtaining trained AI models	System operators	Entry of malicious data into the system during operation Observation of output information, etc. during operation

[101] Quoted from National Institute of Advanced Industrial Science and Technology “Guidelines for Machine Learning Quality Management, 4th Edition” (December 2023)

	The source of the operational input data Operational input data	When operating systems	Operational input data providers System operators	Alteration of data to the system during operation
	Systems		System users System operators	Entry of malicious data into the system during operation Observation of output information, etc. during operation
Information leakage attack on training data	Pre-learning models	After obtaining pre-learning models	Model users (System developers)	Observation of input/output and internal information during operation of the obtained AI model
	Trained AI models	After obtaining trained AI models	System operators	
	The source of the operational input data Operational input data	When operating systems	Operational input data providers System operators	Alteration of the operational input data
	Systems		System users System operators	Entry of malicious data into the system during operation Observation of output information, etc. during operation

Table 8. Overview of quality management implementation items

Pre-analysis (Mainly AI Providers)	Data that needs protection		Handling of deliverables
	- Compliance with the governing law - Identification of sensitive personal data		- Determination of reuse deliverables - Confirmation of the consent agreement
Examination of methods (Mainly AI Developers)	Pre-stage	In-stage	Post-stage
	- Quality of training data - Protection processing - Data distribution (outliers)	- Generalization - PPML (differential privacy)	- Safeguard settings
	Trade-off analysis		
	- Accuracy of judgment vs. fairness - Data protection measures vs. usefulness		

[Contents of the main part (repeat)] *Excerpts of the main paragraph only

6) Transparency

When developing, providing, or using an AI system or service, based on the social context when the AI system or service is used, it is important that each AI business actor provides stakeholders with information to the reasonable extent necessary and technically possible while ensuring the verifiability of the AI system or service.

- Regarding "(3) Reasonable and truthful support"
  [Relevant contents (repeat of contents of the main part)]
  - The provision of information as described in the Main Part "(2) Providing relevant stakeholders with information" does not assume the disclosure of algorithms or source codes, but should be carried out to the extent that it is deemed socially reasonable in light of the characteristics and applications of the technology to be adopted, respecting privacy and trade secrets.
  - If any open technologies are used, conform to the rules specified for them.
  - When open sourcing developed AI systems, any potential social impacts should be considered.

  [Specific methods]
  - For the protection of privacy and trade secrets, for example, it is useful to create explanatory documents for non-engineers.
  - The provision of information is not completed once, and it is useful to accumulate dialog with stakeholders as much as possible in light of the characteristics and applications of the technology to be adopted. In addition, it is important to proactively design and maintain communication design in the same way as design and development
  - When using publicly available technologies or libraries, check the license and comply with it.
    - In particular, pay attention to prohibited commercial licenses, etc.
  - When making the developed AI system open-source, it is useful to identify and respond to the social impact that may arise from disclosure and risks through interviews with relevant stakeholders.

- Regarding "(4) Improving explainability and interpretability for relevant stakeholders"
  [Relevant contents (repeat of contents of the main part)]
  - To gain relevant stakeholders' understanding and sense of safety and display the proof of AI's behaviors, make sure how to explain to those who need explanation for those AI business actors who give explanation to analyze and understand, and take necessary steps.
    - AI Provider: Inform the AI Developer about things that are required to be explained.
    - AI Business User: Inform the AI Developer and AI Provider about things that are required to be explained.

  [Specific methods]
  - For example, it is useful to prepare explanatory documents for non-engineers about the principles of AI operation and decision-making processes.
    - In addition, please refer to Appendix 2. Behavioral Goal 4-1 [Ensuring that the operation of AI management system is explainable] as points to keep in mind when explaining.

[Contents of the main part (repeat)] *Excerpts of the main paragraph only

7) Accountability

When developing, providing, or using an AI system or service, it is important that each AI business actor executes its accountability to stakeholders within reasonable extent for ensuring traceability, conforming to common guiding principles, and the like based on each AI business actor's roles and the degree of risks posed by the AI system or service.

Excerpts from [Relevant contents (repeat of contents of the main part)] and Specific Methods

- Regarding "(1) Improving traceability"
  [Relevant contents (repeat of contents of the main part)]
  - Ensure that data sources and decision-making, etc. carried out during the development, provision, and use of AI can be traced and retracted to the extent technically feasible and reasonable.

  [Specific methods]
  - Data lineage (building provenance mechanisms)
    - To know where the data came from, how it was collected, managed, and moved within AI business actors over time.
    - Such data includes the identifier of the service or AI model that created the content, but it is not required to include user information.
  - Indication that the content is AI-generated (content authentication)
  - Version control of AI models
  - Obtaining training process logs
  - Backtracking and tracking of update history

- Regarding "(3) Designation of responsible persons"
  [Relevant contents (repeat of contents of the main part)]
  - AI business actors should appoint a person responsible for fulfilling accountability.

  [Specific methods]
  - When appointing a responsible person, it is useful to have a clear definition of roles and responsibilities.
  - In formulating policies for risk management and ensuring safety associated with the use of AI systems, collaborate with AI Providers as necessary.
  - To publicize the above policies, etc., it is useful to use the websites, etc. of AI business actors so that stakeholders can easily access them.

- Regarding "(4) Sharing responsibilities among actors"
  [Relevant contents (repeat of contents of the main part)]
  - As for responsibilities shared among actors, AI business actors including non-business users should clarify who should take the responsibilities through contracts or social commitments (voluntary commitments).

  [Specific methods]
  - For clarification of responsibility through contracts, if necessary, it is useful to refer to "Appendix 6. Major precautions for referring to "Contract Guidelines on Utilization of AI and Data."
  - Social commitments may include, for example, the formulation of ethical standards in cooperation with industry groups, etc.

- Regarding "(5) Specific actions for stakeholders"
  [Relevant contents (repeat of contents of the main part)]
  - Formulate and publicly report policies on AI governance and privacy policies, etc. of AI business actors to manage risks and ensure safety associated with the use of AI systems

and services as necessary (including social responsibilities such as sharing visions and disseminating and providing information to society and the general public).

- As necessary, opportunities should be set for accepting comments from stakeholders on incorrect AI output and the like, and objective monitoring of the output should be conducted.
- In the event of a situation that impairs the interests of stakeholders, formulate a policy on how to respond and steadily implement it, and periodically report the progress to stakeholders as necessary.

[Specific methods]

- When setting opportunities for accepting comments from stakeholders, there should be opportunities to receive feedback such as websites and contact points.
  - For details, please refer to Appendix 2. Behavioral Goal 5-2 [Considering opinions of outside stakeholders].
- To the extent possible, periodically disclose the monitoring results of AI systems.
- It is useful to prepare for a situation that impairs the interests of stakeholders by formulating a crisis management response plan, etc.
  - For details, please refer to Appendix 2. Behavioral Goal 3-4-2 [Preliminary consideration of response to incidents/disputes].

[References]

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

[Contents of the main part (repeat)] *Excerpts of the main paragraph only

8) Education/literacy

Each AI business actor is expected to provide the persons engaged in AI in the AI business actor with the necessary education to gain the knowledge, literacy, and ethical views to correctly understand and use AI in a socially correct manner. Each AI business actor is also expected to provide stakeholders with education, in consideration of the characteristics of AI, including its complexity and the misinformation that it may provide, and possibilities of intentional misuse of AI.

[Relevant contents (repeat of contents of the main part)]

- Take necessary measures to ensure that the persons engaged in AI in AI business actors acquire AI literacy of the level sufficient for the engagement.
- It is assumed that the division of tasks between AI and humans will change due to the expansion of generative AI use, so education and reskilling, etc. should be actively discussed to promote new ways of working.
- Provide educational opportunities taking into account differences in knowledge and skills among generations so that various people can acquire a deeper understanding of benefits of AI and increase their risk resilience.
- To improve the overall safety of AI systems and services, provide stakeholders with support to ensure education and literacy advancement as necessary.

[Specific methods]

- Education for AI Developers
  - Cultivate a mindset and culture that is willing to change, including the latest attack methods.
  - Promote cooperation across the entire value chain and understand the trade-offs arising from cooperation.
  - Appealing to the growing need for social responsibilities, etc.
- Education for AI Providers, AI Business Users, and non-business users, etc.
  - Education for AI Business Users and non-business users on how to properly use AI systems and potential risks
  - Information dissemination with the aim of increasing literacy about the appropriate use methods and benefits of AI systems developed by AI Developers, and how to deal with potential risks and risks, etc.

[References]

- Cabinet Office, “Tentative Arrangement of Issues related to AI” (May 2023)
- National Institute of Advanced Industrial Science and Technology “Guidelines for Machine Learning Quality Management, 4th Edition” (December 2023)
- NIST, “AI Risk Management Framework Playbook” (January 2023)

C. Matters to be observed in developing advanced AI systems

AI Developers developing advanced AI systems, including state-of-the-art foundational models and generative AI systems, should comply with the Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems[102] below.

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

I) **Take appropriate measures throughout the development of advanced AI systems, including prior to and throughout their deployment and placement on the market, to identify, evaluate, and mitigate risks across the AI lifecycle.**

- This includes employing diverse internal and independent external testing measures, through a combination of methods for evaluations, such as red-teaming, and implementing appropriate mitigation to address identified risks and vulnerabilities. Testing and mitigation measures, should, for example, seek to ensure the trustworthiness, safety and security of systems throughout their entire lifecycle so that they do not pose unreasonable risks. In support of such testing, developers should seek to enable traceability, in relation to datasets, processes, and decisions made during system development. These measures should be documented and supported by regularly updated technical documentation.
- This testing should take place in secure environments and be performed at several checkpoints throughout the AI lifecycle in particular before deployment and placement on the market to identify risks and vulnerabilities, and to inform action to address the identified AI risks to security, safety and societal and other risks, whether accidental or intentional. In designing and implementing testing measures, organizations commit to devote attention to the following risks as appropriate:
  - Chemical, biological, radiological, and nuclear risks, such as the ways in which advanced AI systems can lower barriers to entry, including for non-state actors, for weapons development, design acquisition, or use.
  - Offensive cyber capabilities, such as the ways in which systems can enable vulnerability discovery, exploitation, or operational use, bearing in mind that such capabilities could also have useful defensive applications and might be appropriate to include in a system.
  - Risks to health and/or safety, including the effects of system interaction and tool use, including for example the capacity to control physical systems and interfere with critical infrastructure.
  - Risks from models of making copies of themselves or "self-replicating" or training other models.
  - Societal risks, as well as risks to individuals and communities such as the ways in which advanced AI systems or models can give rise to harmful bias and discrimination or lead to violation of applicable legal frameworks, including on privacy and data protection.
  - Threats to democratic values and human rights, including the facilitation of disinformation or harming privacy.
  - Risk that a particular event could lead to a chain reaction with considerable negative effects that could affect up to an entire city, an entire domain activity or an entire community.

[102] For the entire text, refer to 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. It should be noted that this document is a living document built on the existing OECD AI principles in response to recent developments in advanced AI systems. Advanced AI systems are defined as the most advanced AI systems, including state-of-the-art foundational models and generative AI systems.

- Organizations commit to work in collaboration with relevant actors across sectors, to assess and adopt mitigation measures to address these risks, in particular systemic risks.
- Organizations making these commitments should also endeavor to advance research and investment on the security, safety, bias and disinformation, fairness, explainability and interpretability, and transparency of advanced AI systems and on increasing robustness and trustworthiness of advanced AI systems against misuse.

II) Identify and mitigate vulnerabilities, and, where appropriate, incidents and patterns of misuse, after deployment including placement on the market.

- Organizations should use, as and when appropriate commensurate to the level of risk, AI systems as intended and monitor for vulnerabilities, incidents, emerging risks and misuse after deployment, and take appropriate action to address these. Organizations are encouraged to consider, for example, facilitating third-party and user discovery and reporting of issues and vulnerabilities after deployment such as through bounty systems, contests, or prizes to incentivize the responsible disclosure of weaknesses. Organizations are further encouraged to maintain appropriate documentation of reported incidents and to mitigate the identified risks and vulnerabilities, in collaboration with other stakeholders. Mechanisms to report vulnerabilities, where appropriate, should be accessible to a diverse set of stakeholders.

III) Publicly report advanced AI systems' capabilities, limitations and domains of appropriate and inappropriate use, to support ensuring sufficient transparency, thereby contributing to increase accountability.

- This should include publishing transparency reports containing meaningful information for all new significant releases of advanced AI systems.
- These reports, instruction for use and relevant technical documentation, as appropriate as, should be kept up-to-date and should include, for example:
  - Details of the evaluations conducted for potential safety, security, and societal risks, as well as risks to human rights;
  - Capacities of a model/system and significant limitations in performance that have implications for the domains of appropriate use;
  - Discussion and assessment of the model's or system's effects and risks to safety and society such as harmful bias, discrimination, threats to protection of privacy or personal data, and effects on fairness; and
  - The results of red-teaming conducted to evaluate the model's/system's fitness for moving beyond the development stage.
- Organizations should make the information in the transparency reports sufficiently clear and understandable to enable deployers and users as appropriate and relevant to interpret the model/system's output and to enable users to use it appropriately; and that transparency reporting should be supported and informed by robust documentation processes such as technical documentation and instructions for use.

IV) Work towards responsible information sharing and reporting of incidents among organizations developing advanced AI systems including with industry, governments, civil society, and academia

- This includes responsibly sharing information, as appropriate, including, but not limited to evaluation reports, information on security and safety risks, dangerous intended or unintended capabilities, and attempts by AI actors to circumvent safeguards across the AI lifecycle.
- Organizations should establish or join mechanisms to develop, advance, and adopt, where appropriate, shared standards, tools, mechanisms, and best practices for ensuring the safety, security, and trustworthiness of advanced AI systems.
- This should also include ensuring appropriate and relevant documentation and transparency across the AI lifecycle in particular for advanced AI systems that cause

significant risks to safety and society.

- Organizations should collaborate with other organizations across the AI lifecycle to share and report relevant information to the public with a view to advancing safety, security and trustworthiness of advanced AI systems. Organizations should also collaborate and share the aforementioned information with relevant public authorities, as appropriate.
- Such reporting should safeguard intellectual property rights.

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

- Organizations should put in place appropriate organizational mechanisms to develop, disclose and implement risk management and governance policies, including for example accountability and governance processes to identify, assess, prevent, and address risks, where feasible throughout the AI lifecycle.
- This includes disclosing where appropriate privacy policies, including for personal data, user prompts and advanced AI system outputs. Organizations are expected to establish and disclose their AI governance policies and organizational mechanisms to implement these policies in accordance with a risk based approach. This should include accountability and governance processes to evaluate and mitigate risks, where feasible throughout the AI lifecycle.
- The risk management policies should be developed in accordance with a risk based approach and apply a risk management framework across the AI lifecycle as appropriate and relevant, to address the range of risks associated with AI systems, and policies should also be regularly updated.
- Organizations should establish policies, procedures, and training to ensure that staff are familiar with their duties and the organization's risk management practices.

VI) Invest in and implement robust security controls, including physical security, cybersecurity and insider threat safeguards across the AI lifecycle.

- These may include securing model weights and, algorithms, servers, and datasets, such as through operational security measures for information security and appropriate cyber/physical access controls.
- This also includes performing an assessment of cybersecurity risks and implementing cybersecurity policies and adequate technical and institutional solutions to ensure that the cybersecurity of advanced AI systems is appropriate to the relevant circumstances and the risks involved. Organizations should also have in place measures to require storing and working with the model weights of advanced AI systems in an appropriately secure environment with limited access to reduce both the risk of unsanctioned release and the risk of unauthorized access. This includes a commitment to have in place a vulnerability management process and to regularly review security measures to ensure they are maintained to a high standard and remain suitable to address risks.
- This further includes establishing a robust insider threat detection program consistent with protections provided for their most valuable intellectual property and trade secrets, for example, by limiting access to proprietary and unreleased model weights.

VII) Develop and deploy reliable content authentication and provenance mechanisms, where technically feasible, 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 with an organization's advanced AI system. The provenance data should include an identifier of the service or model that created the content, but need not include user information. Organizations should also endeavor to develop tools or APIs to allow users to determine if particular content was created with their advanced AI system, such as via watermarks. Organizations should

collaborate and invest in research, as appropriate, to advance the state of the field.
- Organizations are further encouraged to implement other mechanisms such as labeling or disclaimers to enable users, where possible and appropriate, to know when they are interacting with an AI system.

VIII) Prioritize research to mitigate societal, safety and security risks and prioritize investment in effective mitigation measures.
- This includes conducting, collaborating on and investing in research that supports the advancement of AI safety, security, and trust, and addressing key risks, as well as investing in developing appropriate mitigation tools.
- Organizations commit to conducting, collaborating on and investing in research that supports the advancement of AI safety, security, trustworthiness and addressing key risks, such as prioritizing research on upholding democratic values, respecting human rights, protecting children and vulnerable groups, safeguarding intellectual property rights and privacy, and avoiding harmful bias, mis- and disinformation, and information manipulation. Organizations also commit to invest in developing appropriate mitigation tools, and work to proactively manage the risks of advanced AI systems, including environmental and climate impacts, so that their benefits can be realized.
- Organizations are encouraged to share research and best practices on risk mitigation.

IX) Prioritize the development of advanced AI systems to address the world's greatest challenges, notably but not limited to the climate crisis, global health and education
- These efforts are undertaken in support of progress on the United Nations Sustainable Development Goals, and to encourage AI development for global benefit.
- Organizations should prioritize responsible stewardship of trustworthy and human-centric AI and also support digital literacy initiatives that promote the education and training of the public, including students and workers, to enable them to benefit from the use of advanced AI systems, and to help individuals and communities better understand the nature, capabilities, limitations, and impact of these technologies. Organizations should work with civil society and community groups to identify priority challenges and develop innovative solutions to address the world's greatest challenges.

X) Advance the development of and, where appropriate, adoption of international technical standards
- Organizations are encouraged to contribute to the development and, where appropriate, use of international technical standards and best practices, including for watermarking, and working with Standards Development Organizations (SDOs), also when developing organizations' testing methodologies, content authentication and provenance mechanisms, cybersecurity policies, public reporting, and other measures. In particular, organizations also are encouraged to work to develop interoperable international technical standards and frameworks to help users distinguish content generated by AI from non-AI generated content.

XI) Implement appropriate data input measures and protections for personal data and intellectual property
- Organizations are encouraged to take appropriate measures to manage data quality, including training data and data collection, to mitigate against harmful biases.
- Appropriate measures could include transparency, privacy-preserving training techniques, and/or testing and fine-tuning to ensure that systems do not divulge confidential or sensitive data.
- Organizations are encouraged to implement appropriate safeguards, to respect rights related to privacy and intellectual property, including copyright-protected content.
- Organizations should also comply with applicable legal frameworks.

The “Reporting Framework” for advanced AI system developers, agreed upon by the G7 in cooperation with the OECD, began operation in February 2025. This framework requires AI developers to voluntarily check and report their compliance with the “Hiroshima Process International Code of Conduct” for developing advanced AI systems[103]. Developers adhering to this code are expected to participate in the framework.

The framework’s questions are categorized into seven main areas:
(1) Identification and assessment of risks
(2) Risk management and information security
(3) Transparency reporting on advanced AI systems
(4) Organizational governance, incident management, and transparency
(5) Mechanisms for content authentication and provenance verification
(6) Research and investment to enhance AI safety and reduce societal risks
(7) Promotion of human and global benefits

[103] Hiroshima Process International Code of Conduct “Reporting framework”
https://transparency.oecd.ai/

Appendix 4. For AI Providers

In this chapter, "points" and "specific methods" are explained for the contents described in the Main Part "Part 4: Matters Related to AI Providers." After that, in "C. Common guiding principles" in the Main Part "Part 2: The Society to Aim for with AI and Matters to be Tackled by AI Business Actors," specific methods that should be especially considered by AI providers will be explained.

The "specific methods" described here is just an example. Some of them are written on both traditional AI and generative AI, or some are only applicable to either one of them. When considering specific responses, it is important to take into consideration the extent and probability of the risks posed by the AI system to be provided, the technical characteristics, and the resource constraints, etc. of AI business actors.

Also, AI providers who handles advanced AI system should conform to I) to XI) to a proper extent and should conform to XII), by reference to the description of "D. Guiding principles shared among business operators involved in advanced AI systems" in the Main Part "Part 2: The Society to Aim for with AI and Matters to be Tackled by AI Business Actors."

A. Descriptions of Part 4 "Matters Related to AI Providers"

[Contents of the main part (repeat)]

- When implementing an AI system

P-2) i. Actions against risks that consider the lives, bodies, properties, and minds of human and the environment

- ✧ Take measures that prevent AI from causing any harm on the lives, bodies, properties, and minds of stakeholders including AI business users, and the environment. The measures involve ensuring proper performances under usage conditions expected at the time of provision, enabling the AI system to maintain those performances in various situations, and minimizing (by guardrail technology or the like) risks caused by, for example, an uncontrollable robot linking to AI or improper output ("2) Safety").

[Points]

It is expected to ensure that the AI does not cause harm to the lives, bodies or properties of humans through actuator or the like by taking countermeasures as necessary based on the nature and mode of the expected damage and information from AI developers, etc.

It is expected to preliminarily adjust measures to be taken in case that the AI causes harm to the lives, bodies or properties of humans through actuator or the like. In addition, it is expected to provide AI Business Users or non-business users with necessary information about such measures.

In addition to observing the existing laws and regulations and guidelines, it is important to use a new technology to respond to issues caused by a new technology.

[Specific methods]

- Considerations for disadvantages to humans
  - ➢ Consideration for disadvantages to individual persons (e.g. when profiling using AI in a field where individual rights and interests can be severely affected; below are examples of disadvantage for which consideration is needed.)
    - ✧ An erroneous decision is made due to the result of profiling being different from facts.
    - ✧ An individual person is undervalued or overvalued due to only his/her certain characteristics being used for the profiling.

- ✧ If a part of the result of profiling of an individual person is common to characteristics of a particular group and a negative decision is made on such group, a negative decision may be made on such individual person, as well.
- ✧ As the result of profiling, human rights or interests are impaired, such as unreasonable discrimination of a particular individual person or group being promoted.
- ✧ A negative decision may occur in a course of predicting uncertain future in accordance with the result of profiling.
- ✧ An anonymous individual person may be identified by checking the result of profiling based on the information about an anonymous individual person with the result of profiling based on the information about a particular individual person.

- Prevention of incidents
  - ➢ Build a mechanism to secure the security in the entire AI system (achievement of fail safe).
  - ➢ If the existence of a risk that may not be known to AI Developer is recognized, notify AI Developer promptly to consult and consider countermeasures.
  - ➢ Consider a human involvement in advance and during action, such as confirming safe condition, and measures to prevent recurrence afterward.
  - ➢ Confirm the reliability of AI business users by means of a declaration of proper use (of AI) by AI business users.

[References]

- Cabinet Office, “Tentative Arrangement of Issues related to AI” (May 2023)
- The White House, “Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence” (October 2023)

[Contents of the main part (repeat)]
- When implementing an AI system

P-2) ii. Provision contributing to proper use (of AI)

  - ✧ Establish correct considerations to note for using AI systems and services (“2) Safety”).
  - ✧ Use AI within the expected scope of use set by AI developers (“2) Safety”).
  - ✧ Guarantee the accuracy of AI systems/services and recency as necessary (appropriateness of data) of training data at the time of its provision (“2) Safety”).
  - ✧ Examine how AI usage environments of the users of the AI system or service differ from those that AI developers expect (“2) Safety”).
- After an AI system or service starts to be provided
  - ✧ Periodically verify whether the AI system or service is used for proper purposes (“2) Safety”).

[Points]

When providing AI systems and services, AI providers are expected to cooperate with relevant stakeholders to take preventive measures and follow-up measures (information sharing, shutdown and recovery, clarification of the cause, and measures to prevent recurrence, etc.) according to the nature and mode, etc. of damage that may be caused or has been caused by incidents that may occur or have occurred when using AI, security breaches, privacy breaches, etc.

[Specific methods]
- Cooperation with stakeholders and preventive measures and follow-up measures
  - ➢ Provide information for use of AI to proper extent and in proper manner , etc.[104]
  - ➢ Prepare creating a list and procedures for items for which measures should be taken if AI causes harms on lives, bodies and properties of humans.
  - ➢ Take measures that should be taken in case of security breach.
  - ➢ Take measures that should be taken in case of breach of privacy of individual persons.
  - ➢ Share information with stakeholders if a new risk is recognized.
  - ➢ Awareness building activities for the society including potential users
  - ➢ Periodically check the proper use (of AI).

[References]
- Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry, “Guidebook on Corporate Governance for Privacy in Digital Transformation (DX) ver. 1.3” (April 2023)

[104] In systems and services incorporating generative AI, it is important to consider designing user interfaces that facilitate human judgment in the use of AI-generated content.

[Contents of the main part (repeat)]
- When implementing an AI system

P-3) i. Consideration for bias in configurations or data of AI systems and services

- ✧ Guarantee fairness of data at the time of its provision and examine bias contained in referenced information and collaborating external services (“3) Fairness”).
- ✧ Regularly evaluate inputs/outputs of AI models and rationales of decisions made by AI models to monitor for any bias generated. As necessary, encourage AI developers to re-evaluate the bias generated by each technical element forming AI models and promote the improvement of AI models based on the re-evaluation results (“3) Fairness”).
- ✧ Examine the possibility where bias may be introduced that arbitrarily restricts business processes and decisions made by AI business users, or non-business users on AI systems, services, or user interfaces receiving AI output results (“3) Fairness”).

[Points]

AI providers are expected to note the possibility of bias being contained in decisions made by AI systems and services and are expected to take into consideration to prevent individual persons and groups from being unreasonably discriminated according to decisions by AI systems and services.

Note: It is important to pay close attention to the fact that there are multiple criteria for fairness, such as group fairness and individual fairness.

[Specific methods]

- Attention to AI outputs being defined by a variety of bias.
  - ➢ Bias by the data representativeness[105]
    - ✧ Non-securement of the data representativeness may cause biases.
    - ✧ Use of data containing a social bias may cause biases.
    - ✧ Manner of preprocessing may unintentionally cause biases in input data when it is used.
  - ➢ Handling personal data contained in data
    - ✧ When intending to collect a massive amount of data containing personal data to satisfy the data representativeness, handle the data by paying attention to the privacy by masking or deleting personal data.
  - ➢ Bias by algorithm
    - ✧ There may be a bias due to sensitive attributes (individual attributes such as gender and race of target persons that should be excluded from the perspective of fairness) depending on algorithm.
  - ➢ Clarification of sensitive attributes
    - ✧ When clarifying such attributes, consider the reasons listed in Article 14, paragraph (1) of the Constitution of Japan and the attributes referred to in international rules related to human rights.
  - ➢ Clarification of the details of fairness that should be secured regarding sensitive attributes
  - ➢ Addition of constraints to satisfy fairness criteria to machine learning algorithm
  - ➢ Use of tools to check biases (i.e. software)[106]
- Confirmation of fairness criteria (see “Column 15: Group fairness and individual fairness”)
  - ➢ Criteria for group fairness (Below are examples of criteria.)
    - ✧ Remove sensitive attribute and make a prediction only in accordance with non-sensitive attribute (unawareness)
    - ✧ Ensure the same predicted results across groups with different values for sensitive attributes (demographic parity).

[105] Data obtained through measurement, etc. is considered not to be biased but to suitably represent the entire population.
[106] Since a general bias check that is independent of the use case may not reveal the risk, conduct bias check using appropriate datasets tailored to the actually anticipated use cases.

- ✧ Adjust the ratio of the error of the predicted result to the actual result so that it does not depend on the value of the sensitive attribute (equalized odds).

➢ Criteria for individual fairness (Below are examples of criteria.)
- ✧ Individuals with equal attribute values other than sensitive attributes are given the same predicted result.
- ✧ Individuals with similar attribute values are given a similar predicted result (fairness through awareness).

Column 15: Group Fairness and Individual Fairness

Generally, when the fairness is required, the handling of attributes that may cause "unfairness" such as the race and gender (i.e. high-need attributes: same meaning as the sensitive attribute) is called to account. In this case, the group fairness is not to cause discrimination across different groups (e.g. unfavorable treatment of women) with respect to a certain high-need attribute, while the individual fairness is not to cause discrimination across "similar persons" not necessarily limited to a classification based on such particular attribute. Currently, the fairness evaluation (metrics) and measures for machine learning factors that can be used for all purposes are primarily based on the group fairness needing "high-need attributes" and, unless otherwise specified, are based on a viewpoint of group fairness. If the "legitimacy" for "individual party" is required as stated above, a viewpoint of individual fairness is required and general metrics are difficult to be defined. Therefore, measures must be considered to satisfy requirements for each AI system. For the individual fairness, a research of "degree of similarity" using the distance learning is proposed and expected for the future.

- Achievement of fairness through process of qualitative approaches and quantitative methods (see "Column 16: Process to secure the quality of fairness")
  - ➢ Risk analysis approach that qualitatively treats the data for social demands and quality of AI systems and services at the time of use and takes the occurrence of lack of fairness as a risk.
    - ✧ This first requires a qualitative assurance of fairness and then uses a quantitative metrics for fairness, such as the "equality in results," from the implementation stage as necessary and, as the contents of a system or AI factors are specified, will include quantitative approaches, as well.
    - ✧ This is intended to assure the reasonability of setting and selecting the metrics for fairness by analysis or design approaches and is considered to be similar to the risk analysis-based approaches to achieve the "risk avoidance," functional safety and so on (see "Figure 28. Illustration of process structure to secure the fairness quality").

[References]

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

Column 16: Process to assure the fairness quality

"Figure 28. Illustration of process structure to secure the fairness quality" qualitatively handles the "equality of treatment" at highly abstract stages, such as social demands and quality at the time of use, materializes it by a risk analysis approach that takes the occurrence of unfairness (with the same meaning as the impairment of fairness) as a risk and illustrates an example of a series of process achieved from any of the development stages as necessary through the quantitative fairness metrics, such as the "equality of results." This figure is not binding on any individual thinking ways for development or any stage setting but is a model to summarize a flow from the qualitative approach to the quantitative method.

(1) The most abstract demand for fairness: "Equality" and "equal treatment" are required at the level of "justice" or "human rights."

(2) Social demands: At a level of legal system or social rules, such as implied ethical conducts, the "equality of treatment" is required or the "equality of results" in a form of numerical target, etc. is required.

(3) (4) System demand/demand on AI factors: Target setting requires either of the numerical equality of results or the equality of treatment at a level corresponding to the design of entire system and the design of machine learning factors.

(5) Internal quality review: Also, in a process of building a part of internal quality system, the target setting requires either of the numerical equality of results and the equality of treatment.

(6) Internal quality achievement: At a level of the internal quality corresponding to the quality check means, there may be a method to analyze a statistic distribution of results, a method to monitor statistic and analytic indices other than the distribution of results or a method to explain the equality of treatment from a logical structure of implementation.

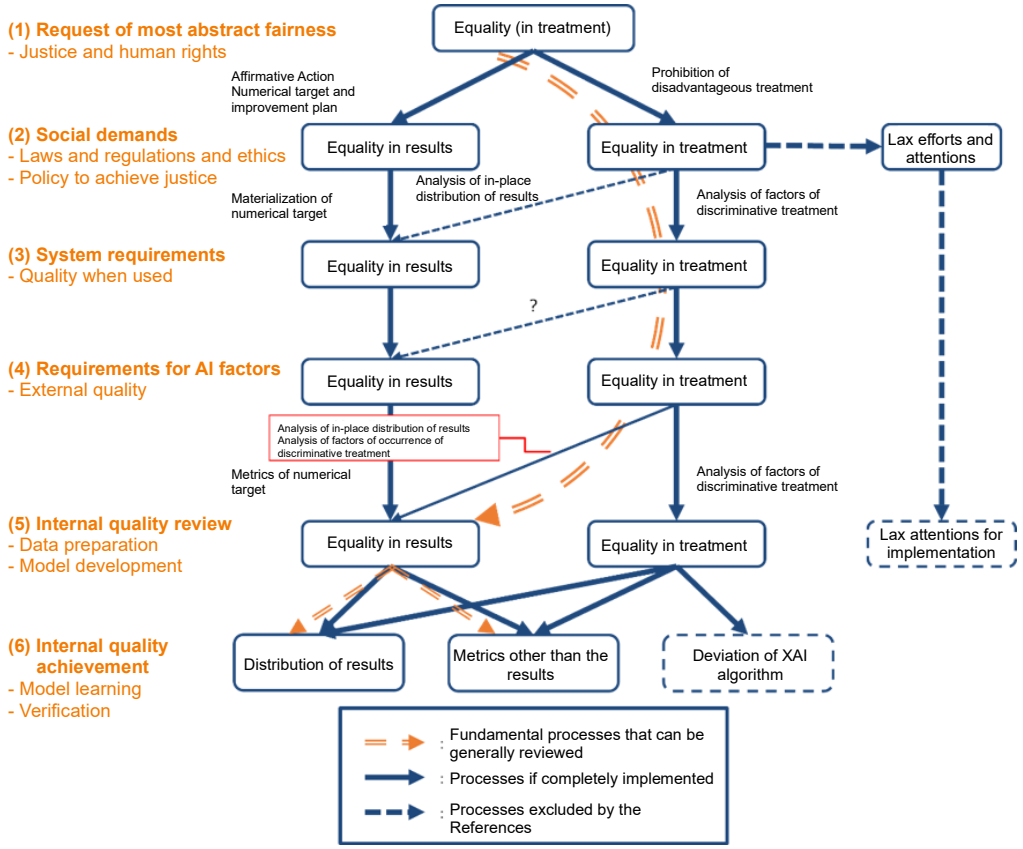


Figure 28. Illustration of a process structure to assure the fairness quality

[Contents of the main part (repeat)]
- When implementing an AI system

P-4) i. Deployment of mechanisms and measures for protecting privacy

✧ Throughout the implementation of an AI system, take privacy protection measures by, for example, introducing a mechanism that appropriately manages and restricts access to personal data based on the characteristics of the adopted technologies (privacy by design) ("4) Privacy protection").

[Points]

Privacy by design should be applied widely in three aspects.

1) IT system
2) Responsible business practices
3) Physical design and network base

The target of privacy by design, that is, "Assuring the privacy and acquiring sustainable competitive advantages for organizations are important" can be achieved by practicing the "Privacy by Design: The 7 Foundational Principles" stated below.

(1) Not afterward but beforehand and not remedial but preventive
(2) Privacy as a default setting
(3) Privacy incorporated in the design
(4) Totally functional - Not zero-sum but positive-sum (i.e. not a zero-sum approach which creates a trade-off relationship but an approach that contains all legitimate interests and targets)
(5) Security from the beginning to the end - protection of an entire lifecycle
(6) Visibility and transparency - maintenance of publication
(7) Respect on privacy of users - maintenance of user-centric principle

[Specific methods]

- Implementation of privacy measures based on the privacy by design[107]
  - ➢ Quality Management Implementation (see "Table 8. Overview of Quality Management Implementation Items")
    - ✧ Check that the data that needs a protection conforms to governing laws and regulations.
    - ✧ Identify high-need personal data as stipulated in the laws and regulations.
    - ✧ Identify reusable deliverables.
    - ✧ Confirm the arrangement agreed upon with data providers and handle the data in accordance with such arrangement.

[107] When utilizing RAG and other external services or data, the risk of unintended leakage of important information (such as personal or confidential information) increases, making privacy protection and security assurance through data anonymization and access restrictions particularly important.

- Respect on relevant stakeholders and personal privacy
  - Delete information that infringes the personal privacy and update AI algorithm and so on (when obtaining any information that infringes the privacy of relevant stakeholders, including AI business users, or individual persons).
  - Request a deletion of information that infringes the personal privacy and update AI algorithm and so on (when spreading any information that infringes the privacy of relevant stakeholders, including AI business users, or individual persons).
  - Since there is a possibility that AI users may input personal information or other privacy-related data when using the system, a mechanism to assist in determining the presence of privacy information and ensuring its confidentiality during system input should be introduced.

[References]

- Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry, “Guidebook on Corporate Governance for Privacy in Digital Transformation (DX) ver. 1.3” (April 2023)
- National Institute of Advanced Industrial Science and Technology “Guidelines for Machine Learning Quality Management, 4th Edition” (December 2023)
- Ann Cavoukian, “Privacy by Design: The 7 Foundational Principles” (September2011)

[Contents of the main part (repeat)]
- When implementing an AI system

P-5) i. Deployment of mechanisms for security measures

✧ Throughout the provision of an AI system or AI service, take security measures appropriately based on the characteristics of the adopted technologies (security by design) (“5) Ensuring security”).

[Points]

Keeping AI security in mind, in order to ensure the confidentiality, integrity, and availability of an AI system, it is expected that reasonable measures will be taken in light of the technical level at that time.[108] In addition, it is expected that the measures to be taken in the event of a security breach should be sorted out in advance, taking into account the intended use and characteristics of the AI system, the magnitude of the impact of the breach, etc.

The security of the AI system to be developed will be ensured by considering security from the early stage of the development process, referring to the security-by-design, etc. defined by the National Center of Incident Readiness and Strategy for Cybersecurity (NISC) in “Measures to Incorporate Information Security from the Planning and Design Stages.” If security functions are added later, or the security tool is run just before shipment, there is a possibility that many returns will occur, resulting in a large development cost. If security measures are taken at the early stage of the development, there will be few returns, resulting in the implementation and provision of AI system with good maintainability.

[Specific methods]
- Implementation of security measures based on the security by design
  - Implementation of threat assessment
    - ✧ Clarify threats faced by AI system and assumed attacks. Clarify “from what the AI system is protected.”
  - Definition of the security requirements
    - ✧ The secure behavior of AI system itself is defined. The types of requirements include system functionality requirements, availability, maintainability, performance. Security requirements are the requirements related to security among the system requirements and define the objectives necessary to safely operate the system. The security requirements are described as part of the system requirements definition document or as a security requirement definition document.
  - Selection of security architecture
    - ✧ Is based on the architecture information required of AI system provided by AI Developer.
    - ✧ Customize and use the architecture recommended by the platform provider with AI system rather than considering the own organization’s unique architecture.
- Classification of attacks against AI
  - System malfunction
    - ✧ Examples of damage as a result of lowering the risk avoidance include without limitation a failure of object detection in automated driving, driver’s missing of abnormality in driving assistance, evasion of malware detection in information security measures, failure to detect invasion in crime prevention system, failure to detect abnormal behaviors, increase in false-positive or false-negative in pathology diagnosis system.
    - ✧ Examples of damage as a result of lowering the AI system performance include without limitation the lowering efficiency in car allocation in traffic and logistics, increasing traffic jam and logistics costs, lowering accuracy rate in product recommendation, demand prediction and shop situation assessment in the field of

[108] When utilizing RAG, consider conducting regular management and audits of the searched and referenced data to detect any unauthorized changes or manipulations early.

retailers and lowering appropriateness in school admission, employment and personnel deployment.
    - ✧ Examples of damage as a result of lowering the fairness include without limitation unfair and discriminative loan in credit examination system, unfair and discriminative school admission, employment and personnel deployment in personnel assessment system and unfair and discriminative criminal risk judgment in crime prevention system.
  - Leakage of information about AI models
    - ✧ Attacks that may leak parameters, functions and other non-public information of AI model may cause a leakage of trade secret, etc. relating to the AI model functions.
  - Leakage of sensitive information contained in training data
    - ✧ If any sensitive information is contained in training data, an attack that leaks any information about training data may cause a privacy violation, leakage of trade secret and breach of legal regulations or contract.
    - ✧ If any medical information or other personal data, per-customer sales information, image data of military facilities where photography is prohibited, or the like is contained in training data set, an attack that leaks any information about training data may cause damage to individual persons (see “Table 6. Examples of damage and threats to the system using machine learning”).
- Consideration for measures upon occurrence of security violation
  - Initial action
    - ✧ Recovery by rollback of the AI system, use of alternative system, etc.
    - ✧ Stop the AI system (kill switch).
    - ✧ Disconnect the AI system from the network.
    - ✧ Confirmation of contents of security violation
    - ✧ Reporting to relevant stakeholders
  - Use of insurance to facilitate indemnification, compensation, etc.
  - Establishment of third-party organ and investigation and analysis of causes and proposals by such organ

[References]

- National Institute of Advanced Industrial Science and Technology “Guidelines for Machine Learning Quality Management, 4th Edition” (December 2023)
- Information-technology Promotion Agency, Japan “Security by Design Guide Instruction Book” (August 2022)
- NCSC, “Guidelines for secure AI system development” (November 2023)
- ACSC, “Engaging with Artificial Intelligence (AI)” (January 2024)

[Contents of the main part (repeat)]
- When implementing an AI system

P-6) i. Documentation of system architectures and the like

✧ In order to improve traceability and transparency, prepare documents describing the system architecture and data processing of the provided AI system or service that influences the decision-making (“6) Transparency”).

[Points]

AI Provider will record and store input/output and other logs of AI system and prepare documents for them in interpretable contents in order to ensure the explainability of AI’s input/output, etc. whereby it becomes easy to improve the process itself and communications and dialogues with relevant stakeholders will be enhanced. Publicize the risk management documents, if necessary. Documentation will enhance the transparency and enable a human review process, which will secure the explainability.

[Specific methods]

- Assurance of explainability
  - Recording and storage of logs of AI systems and services
    - ✧ Purpose of recording and storage of logs (e.g. whether it is intended to clarify causes and prevent recurrence of incidents in fields where the lives, bodies and properties of humans may be harmed)
    - ✧ Frequency of acquisition of logs, accuracy of logs and period of storage of logs
    - ✧ Protection of logs (ensuring confidentiality, integrity, availability, etc.)
    - ✧ Capacity of storage facility for logs
    - ✧ Recording of log time (securement of accuracy by time synchronization)
    - ✧ Scope of logs to be disclosed
    - ✧ Method to store logs (whether stored in a server, stored in recording media or otherwise)
    - ✧ Place to store logs (whether locally, on cloud or otherwise)
    - ✧ Procedures to check logs (method to access logs, etc.)
  - Adoption of AI system implementing interpretable algorithm
    - ✧ Adopt an interpretable AI model of high readability in advance in AI system to be used.
  - Adoption of technological methods to explain the result of decisions by algorithm to a certain extent
    - ✧ A global explanation method to make explanations by replacing interpretable AI model, such as “Make AI prediction and recognition processes readable” (for example, explain the result of inference by AI model with instances contained in data set or data processed therefrom, including SHAP.)
    - ✧ Local explanation method to provide reasons of prediction for particular inputs, such as “provision of important characteristics,” “provision of important training data” and “expression by natural language” (for example, explain inference logic and reason of decision in accordance with a rule of “if” and weight important factors, among input data, that materially influence the inference)
  - Management of historical trail of data
    - ✧ Manage when, where and for what purpose the data used for AI learning, etc. was collected (data provenance).
  - Analysis of input/output trend of AI model
    - ✧ Analyze the output trend of AI on the basis of a combination of multiple inputs and outputs in and from AI (for example, observe a change in output when changing an input pattern gradually).
  - Proper update of technical documents

[References]

- Consortium of Quality Assurance for Artificial-Intelligence-based Products and Services “AI Product Quality Assurance Guidelines” (June 2023)
- OECD, “Advancing accountability in AI” (February 2023)

[Contents of the main part (repeat)]
- After an AI system or service starts to be provided

P-4) ii. Countermeasures against privacy violation

✧ Properly collect necessary information concerning privacy protections on AI systems/services and discuss protection strategy when its violation is recognized to avoid repeated occurrence. (“4) Privacy protection”).

[Points]

Minds of persons and social acceptability for the privacy violation fluctuate due to contexts or passage of time and, therefore, it is expected to always collect relevant information (e.g. market trend, technologies, systems, etc.). Also, it is expected to build a relationship with learned individuals knowledgeable about the privacy violation (e.g. academic persons, consultants, attorneys, consumers associations) for consultation as necessary. Further, it is important to consider and improve the initial response in case of a privacy violation in actual business, follow-up response including subsequent remedy from damage, clarification of the cause and measures to prevent recurrence.

[Specific methods]
- Responses by a privacy protection organization
  - Aggregation of various information relating to new business and contents of service of each in-company department (intended to find every risk caused by a privacy violation in consumers and the society)
  - Initial response mainly by the privacy protection manager, follow-up response including subsequent remedy from damage, clarification of the cause and measures to prevent recurrence (in case that a privacy violation occurs)
  - Building a relationship with in-company departments
    - ✧ It is expected to keep in touch with departments handling AI systems and services not only by widely accepting privacy-relating consultations received from business departments, etc. but also by positively encouraging them to share problem consciousness. It is important to shape a framework and environment where departments that develop new businesses or new technologies can freely consult without bearing worries.
  - Building a framework of a privacy protection organization (Examples of framework patterns are stated below.)
    - ✧ It has no privacy protection organization but appoints a responsible person for each department that handles AI system and service.
    - ✧ It has a (concurrent) privacy protection organization and is affiliated with a department that handles AI system and service.
    - ✧ It has a (regular) privacy protection organization and is affiliated with a department that handles AI system and service.

[References]
- Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry, “Guidebook on Corporate Governance for Privacy in Digital Transformation (DX) ver. 1.3” (April 2023)

[Contents of the main part (repeat)]
- After an AI system or service starts to be provided

P-5) ii. Handling of vulnerabilities

✧ There are many new attack methods targeting AI systems and services, so identify trends in the latest risks and matters requiring attention in each provision step. And, discuss to deal with vulnerabilities ("5) Ensuring security").

[Points]

With respect to AI systems and services to be provided, it is important for AI Provider to provide AI business users or non-business users with services for security measures and to share the past incident information with them.

AI Provider is expected to pay close attention to risks of vulnerabilities existing on the security with respect to the management, improvement and adjustment of AI model. Also, AI Provider is expected to keep AI business users or non-business users informed of the existence of such risks in advance.

"Threats analysis" summarizes threats faced by AI system and service and assumed attacks and clarifies "from what the AI system and service is protected."

[Specific methods]

- Attention to risks for vulnerabilities of AI model (Below are examples of risks.)
  - Risk of intentional malfunction of AI model by adding minor fluctuation indiscernible by humans to data that can be accurately discerned by AI model as a result of insufficient learning or similar reasons and by entering such data (e.g. adversarial example attack).
  - Risk of erroneous learning by commingling inaccurately labelled data in learning with teacher.
  - Risk of easy reproduction by AI model
  - Risk of reverse engineering the data used for learning from AI model
- Measures of Machine Learning-specific Attacks (see "Table 7. Examples of Machine Learning-specific Threats, Attack Interfaces, Attack Execution Phases, Attackers, and Methods for Attack")
  - Data poisoning attack
    - ✧ Confirm the authenticity of dataset and reliability of dataset collection/processing process.
    - ✧ Use a data poisoning detection technology in dataset.
    - ✧ Use a technology to improve the robustness of dataset against data poisoning (by increasing the number of data to mitigate the effect of poisoning).
    - ✧ Do training by robust learning method against data poisoning (e.g. random smoothing and ensemble learning).
    - ✧ Remove and reduce poisoning from trained model.
    - ✧ Traditional security measures against vulnerabilities of software for development and environment of development
  - Manipulation of validation/test data
    - ✧ Confirm the reliability of dataset collection/processing process.
    - ✧ Traditional security measures against vulnerabilities of software for development and environment of development
  - Model poisoning attack
    - ✧ Confirm the reliability of AI model learning/provision process
    - ✧ Use a model poisoning detection technology.
    - ✧ Remove and reduce poisoning of pre-trained models and AI models.
    - ✧ Use the learning mechanisms to remove and reduce poisoning.
    - ✧ Traditional security measures against vulnerabilities of software for development and environment of development
    - ✧ Traditional security measures against vulnerabilities of system environment during operation and operational structures
  - Evasive attack
    - ✧ Method to improve and evaluate the robustness of AI model against hostile data

- ✧ Restriction on input in AI model (e.g. restriction on access right and restriction on the number of times and frequency of access)
- ✧ Use a hostile data detection technology.
- ✧ Concurrently use multiple different AI models or systems.
- ✧ Technological measures to prevent and reduce the model extraction attack

➢ Model extraction attack
- ✧ Use a model extraction attack detection technology.
- ✧ Process AI model's output information, etc.
- ✧ Ensemble learning
- ✧ Use a model extraction risk evaluation technology.

➢ Information leakage attack on training data
- ✧ Privacy protection leaning
- ✧ Generation of privacy protection data

[References]

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

[Contents of the main part (repeat)]

- After an AI system or service starts to be provided

P-6) ii. Providing relevant stakeholders with information

- ✧ Provide information on the AI system or service to be provided (for example, the items listed below) in a timely and appropriate manner so that it can be easily understood and accessed ("6) Transparency").
  - Fact that AI is used, appropriate/inappropriate use methods, etc. ("6) Transparency").
  - Information on safety, including technical characteristics of the AI systems and services provided, foreseeable risks that may arise as a result of using the AI systems and services, and remedies against them ("2) Safety").
  - Possibility of changes in output or programs due to learning by the AI systems and services ("1) Human-centric").
  - Information on the operational status of the AI systems and services, causes of failures, status of actions against them, incidents, etc. ("2) Safety").
  - Details of an update of the AI system, if any, and information on reasons for the update ("2) Safety").
  - Policies on collecting data learned by AI models, how AI models learn the data, and the system for implementing the learning ("3) Fairness," "4) Privacy protection," "5) Ensuring security").

[Points]

AI Provider is expected to secure the explainability of the result of AI output by taking into account the social context when using AI, including when using AI in fields where a material impact may be given to individual rights and interests, for the purposes of obtaining satisfaction and assurance of AI business users and showing evidences of AI behaviors to this end. At that time, it is expected to take necessary measures for analyzing and understanding what explanations are required.

After assessing and addressing the risks, it is important to verify whether the AI system conforms to the regulations, AI governance and ethical standards and to share the result with relevant stakeholders. This promotes the understanding of the risks and rational grounds behind decision-making and behaviors. It is important to surely share the information of AI's undesirable behaviors and incidents with the relevant stakeholders not only by establishing the process and tools for monitoring and review but also by communicating and reviewing them regularly.

[Specific methods]

- Information sharing about AI system and services
  - The fact that the AI system and services to be provided use AI and the intended use and method of use
  - The scope of use of AI
  - Benefits and risks according to the nature, mode of use, etc. of AI
  - Method of regularly checking the scope and method of using AI system and services to be provided (especially a method of observing and confirming if AI system is autonomically updated), importance and frequency of check, risks caused by non-check, etc.
  - Update of AI system and inspection, repair, etc. of AI system implemented to improve AI functions and mitigate risks in a course of use
  - Details of the evaluations conducted for safety, security, and societal risks, as well as risks to human rights,
  - Appropriate field of use and the limit of ability and performance of AI model or AI system and services that has effect on the use
  - Discussion and assessment of the AI model's or AI system's and service's effects and risks to safety and society such as harmful bias, discrimination, threats to protection of privacy or personal data, and effects on fairness
  - The results of red-teaming conducted to evaluate the AI model's or AI system's and service's fitness for moving beyond the development stage.

- Attention to the provision of information
  - ✧ AI business users will share necessary information in a timely manner.
  - ✧ Information to be provided about AI system and services shall be provided before use thereof.
  - ✧ If the above information cannot be provided before use of AI system and services, develop a system to respond to feedback from AI business users or non-business users according to the risks assumed in accordance with the nature, mode of use, etc. of AI.

[References]
- Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry, "Guidebook on Corporate Governance for Privacy in Digital Transformation (DX) ver. 1.3" (April 2023)
- NIST, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)" (January 2023)
- OECD, "Advancing accountability in AI" (February 2023)
- AISI, "Guide on Red Teaming Method for AI Safety" (September 2024)

[Contents of the main part (repeat)]
After an AI system or service starts to be provided

P-7) i. Explanation to AI business users of conformity to common guiding principles

- ✧ Encourage AI business users to use AI properly and provide them with the following information ("7) Accountability"):
  - Call attention to the use of data for which accuracy, and recency as necessary (appropriateness of data), are guaranteed ("2) Safety").
  - Call attention to the learning of inappropriate AI models during in-context learning ("2) Safety").
  - Precautions for when inputting personal data ("4) Privacy protection").
- ✧ Call attention to inappropriate input of personal data into the AI systems and services to be provided ("4) Privacy protection").

[Points]

In accordance with the intent of other "Common guiding principles," AI Providers are expected to perform its reasonable accountability by providing AI business users or non-business users with information and explanations on AI system characteristics and conducting dialogues with various stakeholders according to the size of their own knowledge and ability in light of the nature, objectives, etc. of AI used by them in order to obtain the reliability on AI from people and the society.

When using AI in fields where the lives, bodies and properties of humans may be harmed, AI Providers are expected to take measures as necessary and explain the contents thereof to AI business users or non-business users to a reasonable extent in accordance with the information from AI Developer, etc. in light of the nature, mode, etc. of assumed damage.

[Specific methods]
- Call attention to AI business users or non-business users for use of AI system and services and take relevant measures.
  - Inspect and repair AI, update AI system and promote responses to AI business users or non-business users. (The objective is that AI does not cause harm to the lives, bodies or properties of humans through actuator or the like. For a period from discovery of problems to provision of updates, provide information about, and call attention to, the problems timely and appropriately.)
  - Provide information about measures to be taken if the AI causes harm to the lives, bodies or properties of humans (if necessary).

- Confirm input, output and other logs by recording and storing them and call attention to inappropriate input (intended to inhibit malicious use by AI business users and non-business users).
- Measures to be taken in case of violation of privacy of AI Business User, non-business users and other relevant stakeholders or individuals persons.
- Proper scope of use and manner of use of AI system and services (Provide information after having confirmed the purpose of use, intended use, nature, ability, etc. of AI based on the information and explanations provided by AI Developer, etc.)

[References]

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

[Contents of the main part (repeat)]
- After an AI system or service starts to be provided

P-7) ii. Documentation of service agreements or the like

- ✧ Compile service agreements for AI business users or non-business users ("7) Accountability").
- ✧ Present privacy policies ("7) Accountability").

[Points]

It is helpful to use the Service Level Agreement (hereinafter "SLA"), which is a common recognition of guarantee standards for the contents, scope, quality, etc., of services, in order to dissolve uncertainty in providing AI services and achieve maintenance and management of proper service level. SLA is expected to clarify the scope, details and preconditions of services and the level of requirement for the service level and to shape a common recognition of both AI business users/non-business users and AI Providers.

To build a reliable relationship with AI business users or non-business users and secure the social reliability on business activities, it is expected to formulate and publicly report the "concept and policy of personal data protection (so-called privacy policy, privacy statement, etc.)."

[Specific methods]

- Preparation of service agreement
  - ➢ Setting the level of requirements for the subject AI services
    - ✧ In setting the level, determine the priority by considering the effect on business in case of occurrence of an incident and define mainly those of high importance.
    - ✧ In determining the service level, define objective items (i.e. quantitative values, measurement by formula, etc.) to prevent a difference in recognition between AI business users/non-business users and AI Provider.
- Formulation and public report of the concept and policy of personal data protection (privacy policy and privacy statement, etc.)
  - ➢ Clarify contracted processing.
    - ✧ Make contracted processing more transparent by clarifying whether the processing is contracted or not and what tasks are contracted.
  - ➢ Public report
    - ✧ After having formulated a policy, publicly report it by posting it on the website and explain it externally in advance in an easy-to-understand manner.
- Appropriate updating of documents

[References]

- The 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. Descriptions of "Common guiding principles" in Part 2

Although not mentioned in the Main Part, "Part 4 Matters Related to AI Providers," specific methods for the Main Part, "Part 2 Common Guiding Principles," which are especially important for AI providers, are explained here.

[Contents of the main part (repeat)] *Excerpts of the main paragraph only

1) Human-centric

When developing, providing, or using an AI system or service, each AI business actor should act in a way that does not violate the human rights guaranteed by the Constitution of Japan or granted internationally, as the foundation for accomplishing all matters to be conducted, including the matters described later. In addition, it is important that each AI business actor acts so that the AI expands human abilities and enables diverse people to seek diverse well-being.

- Regarding "(1) Human dignity and autonomy of individuals"
  [Relevant Description]
  - Based on the social context of AI use, respect human dignity and the autonomy of individuals.

  [Specific methods]
  - Promote researches to mitigate societal, safety and security risks.
    - Research to mitigate societal, safety and security risks and investment in effective mitigation measures (Below are examples of research contents.)
      - Maintenance of democratic value
      - Respect on human rights
      - Protection of children and socially vulnerable people
      - Protection of intellectual property rights and privacy
      - Avoidance of harmful bias
      - Avoidance of disinformation/misinformation
      - Avoidance of information manipulation, etc.
    - Research of risk mitigation and sharing of best practices (to be conducted to a possible extent)

- Regarding "(2) Paying attention to decision-making and emotional manipulation by AI"
  [Relevant Description]
  - Do not develop, provide or use AI systems and services for the purpose of unjustly manipulating or on the precondition of manipulating human emotions, such as decision-making and cognition.

  [Specific methods]
  - Measures against manipulations on decision-makings and emotions
    - Alert AI business users or non-business users
    - Promote the sharing and recognition of the dependence on AI in the field of education and the existence of risks for manipulations on decision-makings and emotions
    - Consider incentives to promote the detection and report of vulnerabilities after introduction (e.g. incentive scheme, contest, goods, etc.)

- Regarding "(3) Countermeasures against disinformation"
  [Relevant Description]
  - Generative AI has enabled everyone to forge fake information that seems to be true and fair, so recognize the increasing risk of destabilizing and confusing the society through disinformation, misinformation, and biased information generated by AI, and take necessary countermeasures.

[Specific methods]
- Consideration for measures to avoid risks of disinformation, misinformation and biased information
  - Develop and introduce a technology to enable AI business users or non-business users to identify the information generated by AI.
  - Use technologies to prevent the output of false or misleading information (such as hallucinations) by generative AI, including RAG[109]
  - Awareness-raising and information literacy education for a wide range of ages

- Regarding "(4) Ensuring diversity/inclusion"
  [Relevant Description]
  - In addition to ensuring fairness, to prevent information poverty and digital poverty and allow more people to enjoy the benefits of AI, pay attention to make it easy for socially vulnerable people to use AI.

  [Specific methods]
  - Mind to leave nobody behind AI use
    - Improvement of UI (user interface)/UX (user experience)
    - Development of safe and assured environment of use
    - Development of public digital platform

- Regarding "(6) Ensuring sustainability"
  [Relevant Description]
  - Examine the impact of the whole lifecycle on the global environment during the development, provision, and use of AI systems and services.
  - Provide information to ensure that users and non-business users can appropriately differentiate and utilize models, such as employing lightweight models with minimal environmental impact according to their purpose.

  [Specific methods]
  - Consideration of globally common issues
    - Support the progress of the sustainable development goals of the United Nations and encourage the development and use of AI for global interest. (Below are examples of global issues.)
      - Climate control
      - Health and welfare of humans (WHO)
      - Education of high quality
      - Eradication of poverty and elimination of starvation from the world
      - Maintenance of hygiene
      - Clean energy at reasonable prices
      - Eradication of inequality
      - Responsible consumption and production, and so on

[References]
- The Ministry of Internal Affairs and Communications "White Paper on Telecommunications 2021 version" (July 2021)
- United Nations "Sustainable Development Goals" (September 2015)

[109] The accuracy of RAG depends on the information it utilizes. Therefore, when employing RAG, it is necessary to use highly reliable and accurate databases or data sources, pay attention to the update status of the databases used by users, and regularly manage and monitor the quality of the information.

[Contents of the main part (repeat)] *Excerpts of the main paragraph only

2) Safety

Each AI business actor should avoid damage to the lives, bodies, minds, and properties of stakeholders during the development, provision, and use of AI systems and services. In addition, it is important that the environment is not damaged.

- Regarding "(1) Taking into consideration the lives, bodies, properties and minds of humans and the environment"
  [Relevant Description]
  - Determine the responses for cases where the safety of AI systems or services is endangered so that the steps can be quickly taken in such cases.

  [Specific methods]
  - Summarize incident measures and consider measures upon occurrence thereof.
    - ✧ Summarize incident measures.
      - Preliminarily arrange emergency contact system in case of occurrence of harm.
      - Summarize methods of cause investigation and method of recovery works.
      - Consider for measures to prevent recurrence and summarize response policy.
      - Establish the method to share incident-related information.
    - ✧ Initial action
      - Recovery by rollback of the AI system, use of alternative system, etc.
      - Stop the AI system (kill switch).
      - Disconnect the AI system from the network.
      - Confirmation of the details of the harm
      - Reporting to relevant stakeholders
    - ✧ Use of insurance to facilitate indemnification, compensation, etc.
    - ✧ Establishment of third-party organ and investigation and analysis of causes and proposals by such organ

[References]

- Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry, "Guidebook on Corporate Governance for Privacy in Digital Transformation (DX) ver. 1.3" (April 2023)

[Contents of the main part (repeat)] *Excerpts of the main paragraph only

3) Fairness

During the development, provision, or use of an AI system or service, it is important that each AI business actor makes efforts to eliminate unfair and harmful bias and discrimination against any specific individuals or groups based on race, gender, national origin, age, political opinion, religion, and so forth. It is also important that before developing, providing, or using an AI system or service, each AI business actor recognizes that there are some unavoidable biases even if such attention is paid, and determines whether the unavoidable biases are allowable from the viewpoints of respect for human rights and diverse cultures.

- Regarding "(2) Intervention by decisions made by humans"
  [Relevant Description]
  - To prevent AI from outputting unfair results, consider implementing interventions by having humans make decisions at the appropriate time, rather than letting AI make the decisions alone.

  [Specific methods]
  - Determination of whether the intervention by decisions made by humans is necessary (Below are examples of criteria for determination.)
    - Nature of the rights and interests of AI business users or non-business users who are affected by output of AI and intentions of AI business users or non-business users
    - Degree of reliability of output of AI (superior or inferior to the reliability of decisions made by humans)
    - Temporal grace necessary for decisions made by humans
    - Ability expected of AI business users or non-business users who make decisions
    - Need to protect matters to be decided (e.g. response to individual applications by humans or response to volume applications by AI systems and services)
    - Uncertainty of statistical future prediction
    - Need and degree of satisfactory reasons for decision-makings (decisions)
    - Degree of assumed discrimination based on race, creed, or gender due to the fact that the training data contains social bias against minorities, etc.
  - Ensure the effectiveness of decisions made by humans
    - On the assumption that an explanation is obtained from AI that has explainability, preliminarily clarify items for which decisions should be made by humans (if it is considered appropriate for humans to make final decisions on output of AI).
    - Provide information and explanations so that AI business users or non-business users may acquire necessary ability and knowledge in order to appropriately assess output of AI (if it is considered appropriate for humans to make final decisions on output of AI).
    - Preliminarily summarize responses to ensure the effectiveness of decisions made by humans.

[References]

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

[Contents of the main part (repeat)] *Excerpts of the main paragraph only

6) Transparency

When developing, providing, or using an AI system or service, based on the social context when the AI system or service is used, it is important that each AI business actor provides stakeholders with information to the reasonable extent necessary and technically possible while ensuring the verifiability of the AI system or service.

- Regarding "(4) Improving explainability and interpretability for relevant stakeholders"
  [Relevant Description]
  - Share necessary explanations for those to be explained with actors who explain to analyze requirements of such explanation to gain relevant stakeholders' understanding and sense of safety to provide proof of AI operations.

  [Specific methods]
  - Ensuring explainability
    - ✧ Identify a part of insufficient explanation in light of needs, opinions, etc. of AI business users or non-business users and assess the contents of explanation in cooperation with AI Developer.
    - ✧ Analyze the context in cooperation with stakeholder(s) including AI Developer and investigate and document potential risks, including affected matters and situations that may be affected.
    - ✧ Ensure the monitoring and review interface to follow the risks and the frequency, function and effectiveness thereof.
    - ✧ Establish and communicate the remedial mechanism, including the process by which stakeholders can lodge complaints.

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

[Contents of the main part (repeat)] *Excerpts of the main paragraph only

8) Education/literacy

Each AI business actor is expected to provide the persons engaged in AI in the AI business actor with the necessary education to gain the knowledge, literacy, and ethical views to correctly understand and use AI in a socially correct manner. Each AI business actor is also expected to provide stakeholders with education, in consideration of the characteristics of AI, including its complexity and the misinformation that it may provide, and possibilities of intentional misuse of AI.

[Relevant Description]
- Take necessary measures to ensure that the persons engaged in AI in AI business actors acquire AI literacy of the level sufficient for the engagement.
- It is assumed that the division of tasks between AI and humans will change due to the expansion of generative AI use, so education and reskilling, etc. should be actively discussed to promote new ways of working.[110]

[Specific methods]
- Ensuring AI literacy
  - Formulate AI policies clarifying roles and responsibilities and keep the persons involved in AI in AI business actors informed thereof.

[110] The Ministry of Health, Labour and Welfare's Employment Policy Study Group discusses career development and skill education based on technological changes under the theme of "Improving Labor Productivity Utilizing New Technologies." https://www.mhlw.go.jp/stf/shingi/other-syokuan_128950.html

  - Define characteristics of reliable AI and keep the persons involved in AI in AI business actors informed thereof.
  - Collect information about laws and regulations applicable to AI systems and keep the persons involved in AI in AI business actors informed thereof.
  - Collect potentially adverse effects that may be generated from AI systems and keep the persons involved in AI in AI business actors informed thereof.
  - Keep the persons involved in AI in AI business actors informed of the fact that the data digital technologies including AI are being used for various tasks.
- Education and reskilling
  - Provide trainings that comprehensively respond to technological and socio-technological aspects of AI risk management.
  - Education for improvement of resilience to environmental changes, etc.
    - ✧ Mental rotation that flexibly switch between “vertical thinking” and “lateral thinking”
    - ✧ Enhance the modelling skill required for organizational evaluation.
    - ✧ Use an agile thinking to extend the learning horizontal line of a field of skill one is bad at.
    - ✧ Improve the evaluation skill to analyze uncertainty that is difficult to predict with past experiences and know-hows.
    - ✧ Convert the “important points” of organizational evaluation (to an agile “instantaneous force plus vitality”).
    - ✧ Method of evaluation of resynchronization (Configuration, Architecture, Synthesis and Dissemination) of more complexed and advanced management in situations where the AI governance is formed in a hybrid of centralization and dispersion.

[References]
- The Ministry of Economy, Trade and Industry, Information-technology Promotion Agency, Japan “Digital Skill Standards ver. 1.1” (August 2023)
- NIST, “AI Risk Management Framework Playbook” (January 2023)

[Contents of the main part (repeat)] *Excerpts of the main paragraph only

10) Innovation

Each AI business actor is expected to make efforts to actively contribute to the promotion of innovation for the whole society.

[Relevant Description]
- Ensure the interconnectivity and interoperability between your AI systems/services and other AI systems/services.
- When there are standard specifications, comply with them.

[Specific methods]
- Standardization of data format, protocol, etc.
  - Data format (syntax and semantics) in input/output of AI
  - Method of connection for linkage between AI systems and services (or protocol of each layer, especially via the network)
  - In implementing multiple AI models or using new dataset, confirm that the same language is used. In case of differences, consider adjustments such as tokenization methods and vocabulary expansion.

[References]
- National Institute of Advanced Industrial Science and Technology “Guidelines for Machine Learning Quality Management, 4th Edition” (December 2023)

Column 17: The Japan Digital Health Alliance (JaDHA)
"Generative AI Guideline for Healthcare Providers"

The Japan Digital Health Alliance[111] (thereafter "JaDHA") released "Generative AI Guideline for Healthcare Providers (self reference guideline for healthcare providers using Generative AI for their services)[112]" to make safe digital environment for AI users to safely select services on January 18, 2024. On February 7, 2025, Version 2.0 was formulated as a revised edition based on technological and institutional trends.[113]

This guideline presumes that AI guidelines for business and provides a set of checkpoints for AI providers to safely provide their services using Generative AI based on discussions about privacy issues which are especially important in healthcare compared with other industries as well as strong impact caused by false or misleading information.

The guideline includes realistic and concrete cautions about data treatment in specific cases where readers, as AI providers on AI guidelines for business, should pay attention concerning "3)Fairness" and "4)Privacy protection" after correctly recognizing AI actors and related value chains for specific occasions.

In addition, it structures accountable explanation and presentations to AI users based on "7) Accountability" and presents applicable checklist for AI providers as well as references.

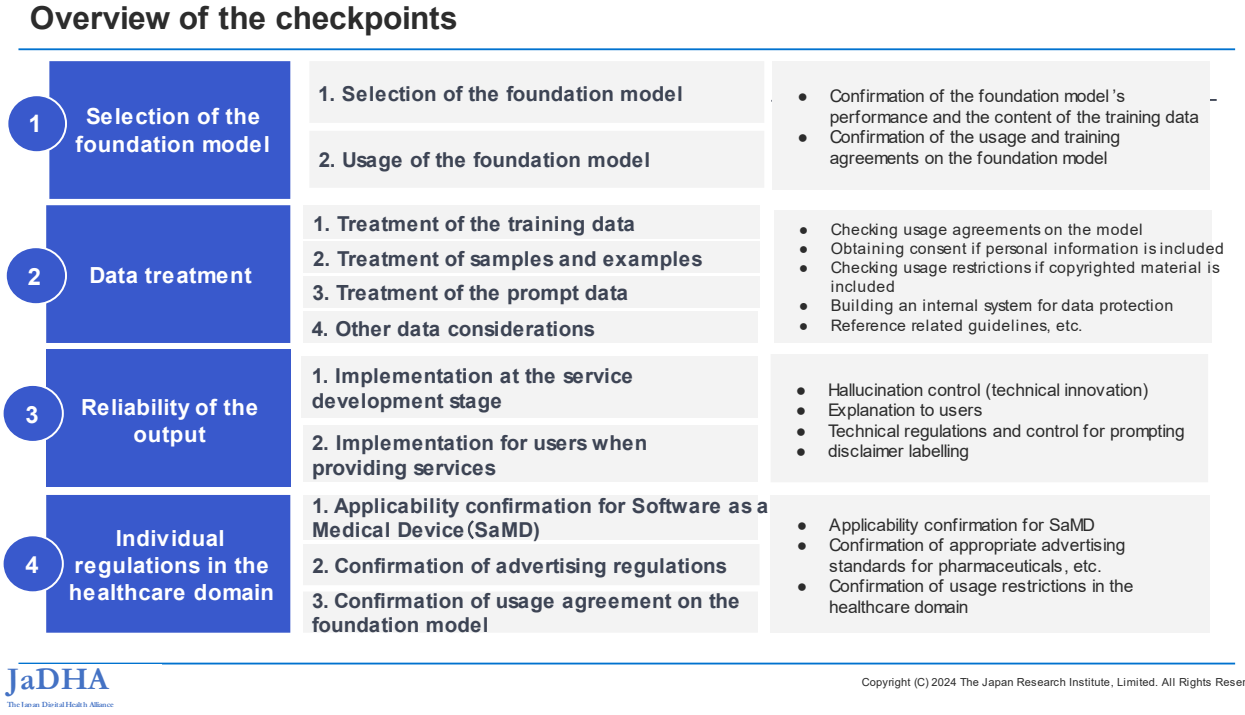


Figure 29. "Generative AI Guideline for Healthcare Providers" by JaDHA

JaDHA expects to promote service promotion and innovation in healthcare using Generative AI. It also aims to provide tools to apply generative AI for healthcare services especially for start-ups and SME.

The guideline by JaDHA is planned to be updated occasionally following technology advancements as well as related legal changes concerning Generative AI to provide valid and timely information for healthcare providers.

[111] Established in March 2022 as an industry organization to consider the development and issues of the digital health industry in Japan. Currently, companies with diverse attributes are participating, from pharmaceutical and medical device manufacturers to health tech startups.
[112] https://jadha.jp/news/news20240118.html
[113] https://jadha.jp/news/news20250207.html

Appendix 5. For AI Business Users

In this chapter, "points" and "specific methods" are explained for the contents described in the Main Part "Part 5: Matters Related to AI Business Users." After that, in "C. Common guiding principles" in the Main Part "Part 2: The Society to Aim for with AI and Matters to be Tackled by AI Business Actors," specific methods that should be especially considered by AI business users will be explained.

The "specific methods" described here is just an example. Some of them are written on both traditional AI and generative AI, or some are only applicable to either one of them. When considering specific responses, it is important to take into consideration the extent and probability of the risks posed by the AI system and services to be used, the technical characteristics, and the resource constraints, etc. of AI business actors.

Also, AI business users who handles advanced AI system should conform to I) to XI) to a proper extent and should conform to XII), by reference to the description in "D. Guiding principles shared among business operators involved in advanced AI systems" of the Main Part "Part 2: The Society to Aim for with AI and Matters to be Tackled by AI Business Actors."

A. Descriptions Part 5 "Matters Related to AI Business Users"

[Contents of the main part (repeat)]

- When using AI systems and services

U-2) i. Proper use (of AI) that considers safety

- ✧ Conform to instructions for use specified by AI providers, and use AI systems and services within the expected scope of use set by AI providers during the design process ("2) Safety").
- ✧ Input data for which accuracy, and recency as necessary (appropriateness of data), are guaranteed ("2) Safety").
- ✧ Understand the degrees of precision and risks of AI output and use AI output after confirming various risk factors ("2) Safety").

[Points]

AI business users should use AI based on the information provided by AI Provider (including the one from AI Developer) and explanations by considering the social context when they use AI.[114] [115]

In using AI operating through actuator or the like, if a transition to human operation is scheduled due to satisfaction of certain conditions, AI business users or non-business users are expected to preliminarily recognize who is responsible for situations before, during and after transition. Also, they are expected to acquire necessary abilities and knowledge by receiving explanations from AI Provider for conditions and method of transition.

When using AI, it is important for AI business users to cooperate with relevant stakeholder(s) to take preventive measures and follow-up measures (information sharing, shutdown and recovery, clarification of the cause, and measures to prevent recurrence, etc.) according to the nature and mode, etc. of damage that may be caused or has been caused by incidents that may occur or have occurred when using AI based on the information provided by AI Provider (including the one from AI Developer) , security breaches, privacy breaches, etc.

[Specific methods]

- Information acquisition about AI system and services

[114] When utilizing RAG, pay attention to ensuring the recency and reliability of the data being searched and referenced. It is important to note that when using RAG (Retrieval-Augmented Generation), the convergence of responses generated by AI may accelerate, which may not be suitable for tasks requiring content diversity and originality.

[115] If program code generated using generative AI contains security vulnerabilities, it may lead to information tampering or leakage. Additionally, if incorrect or inefficient code is generated, there is a concern that it may result in performance degradation or accidents. It is also important to be aware of the possibility that the generated code may infringe on others' intellectual property rights.

  - Proper intended use and method of AI system and services to be used
  - Benefits and risks according to the nature, mode of use, etc. of AI
  - Method of regular confirmation of the scope and method of using AI (especially the method of observation and confirmation if AI is autonomically renewed), importance and frequency of confirmation, risks from non-confirmation and so on.
  - Update of AI system and inspection and repair of AI to be conducted to improve AI functions and mitigate risks through a course of use.
- Use in proper scope and manner
  - Recognize benefits and risks according to the nature, mode of use, etc. of AI and understand proper intended use (before use).
  - Acquire knowledge and skills necessary for proper use (of AI) (before use).
  - Regularly confirm that AI is used in a proper scope and manner (during use).
  - Update AI system and inspect and repair AI or request AI Provider to do so (for the purpose of improving AI functions and mitigating risks through a course of use).
    - ✧ Keep it in mind that the update may affect other collaborating AIs.
  - Feedback the incident information to AI Provider (or AI Developer through AI Provider) (including a case where any incident has occurred or is predicted to occur).
- Preventive measures and follow-up measures in cooperation with relevant stakeholders
  - Provide information for use in proper scope and manner.
  - Take measures that should be taken if AI causes harm to the lives, bodies and properties of humans.
  - Take measures that should be taken in case of security breach.
  - Take measures that should be taken in case of breach of privacy of individual persons.
  - Awareness building activities for the society including potential users
  - Promptly share the information on incidents, etc. with AI Provider and AI Developer and consider measures.

[References]

- Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry, “Guidebook on Corporate Governance for Privacy in Digital Transformation (DX) ver. 1.3” (April 2023)

[Contents of the main part (repeat)]
- When using AI systems and services

U-3) i. Consideration for bias in input data or prompt

✧ Input data for which fairness is guaranteed to avoid significant lack of fairness, pay attention to bias in prompts, and be responsible for determining whether to use AI output results for business (“3) Fairness”).

[Points]

In the event of any doubt for the result of AI output, AI business users are expected to inquire AI Provider (or AI Developer through AI Provider) as necessary.

AI business users are expected to pay close attention to the representativeness of data used for AI learning, etc., social biases contained the data and the like in light of the possibility of AI output being determined by data used for learning and in accordance with the social context in which AI is used.

In order to maintain the fairness of judgments made by AI, in light of the social context in which AI is used and the rational expectations of people, AI business users are expected to involve human judgment as to whether or not to use such judgments, or how to use them, etc.

[Specific methods]

- Pay close attention to the fact that AI output is determined by various biases (which is a point to determine as to whether or not inquiry to AI Provider is necessary).
  - Bias by the data representativeness
    - ✧ Non-securement of the data representativeness may cause biases.
    - ✧ Use of data containing a social bias may cause biases.
    - ✧ Manner of preprocessing may unintentionally cause biases in input data when it is used.
  - Handling personal data contained in data
    - ✧ When intending to collect a massive amount of data containing personal data to satisfy the data representativeness, handle the data by paying attention to the privacy by masking or deleting personal data.
  - Bias by algorithm
    - ✧ There may be a bias due to sensitive attributes (individual attributes such as gender and race of target persons that should be excluded from the perspective of fairness) depending on algorithm.
  - Clarification of sensitive attributes
  - Clarification of the details of fairness that should be secured regarding sensitive attributes
  - Addition of constraints to satisfy fairness criteria to machine learning algorithm
- Confirmation of fairness criteria (see “Column 15: Group fairness and individual fairness”)
  - Confirmation of criteria for group fairness (Below are examples of criteria.)
    - ✧ Remove sensitive attribute and make a prediction only in accordance with non-sensitive attribute (unawareness).
    - ✧ Ensure the same predicted results across groups with different values for sensitive attributes (demographic parity).
    - ✧ Adjust the ratio of the error of the predicted result to the actual result so that it does not depend on the value of the sensitive attribute (equalized odds).
  - Confirmation of criteria for individual fairness (Below are examples of criteria.)
    - ✧ Individuals with equal attribute values other than sensitive attributes are given the same predicted result.
    - ✧ Individuals with similar attribute values are given a similar predicted result (fairness through awareness).

[References]

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

[Contents of the main part (repeat)]
- When using AI systems and services

U-4) i. Countermeasures against inappropriate input of personal data and privacy violation

- ✧ Refrain from improperly inputting personal data to AI systems and services ("4) Privacy protection").
- ✧ Collect information on privacy violation in AI systems and services properly and take the necessary steps to prevent violations ("4) Privacy protection").

[Points]

When using AI system and services, the personal data should be handled properly in accordance with the rules under the personal information protection act by reference to the "Remarks for the use of generative AI services" of the personal data protection committee, as well.

Establishment of a privacy protection organization will substantially promote activities, such as close communication between in-company departments using AI system and service, including a new business department, collection of relevant information from outside learned individuals and consideration of multidirectional measures. Technical innovation and consumers' increased awareness of privacy are day-to-day expanding the extent that should be considered from perspectives of privacy protection. Therefore, it is important to build a privacy protection organization that can assure multidirectional assessment of and agile response to social demands, including technical innovation and consumers' awareness for privacy issues. In addition, to handle consumers' and other personal data globally, sufficient care should be paid and a global system should be built with respect to application of overseas laws and regulations to address the privacy protection.

[Specific methods]

- Responses by a privacy protection organization
  - ➢ Aggregation of various information relating to new business and contents of service of each in-company organization (intended to find every risk caused by a privacy violation in consumers or the society)
  - ➢ Initial response mainly by the privacy protection manager, follow-up response including subsequent remedy from damage, clarification of the cause and measures to prevent recurrence (in case that a privacy violation occurs)
  - ➢ Building a relationship with in-company organization
    - ✧ It is expected to keep in touch with organizations using AI systems and services not only by widely accepting privacy-relating consultations received from each organization but also by positively encouraging them to share problem consciousness. It is important to shape a framework and environment where organizations that develop and use new businesses or new technologies can feely consult without bearing worries.
  - ➢ Building a framework of a privacy protection organization (Examples of framework patterns are stated below.)
    - ✧ It has no privacy protection organization but appoints a responsible person for each organization that uses AI system and service.
    - ✧ It has a (concurrent) privacy protection organization and is affiliated with an organization that uses AI system and service.
    - ✧ It has a (regular) privacy protection organization and is affiliated with an organization that uses AI system and service.
- Preliminary adjustment and implementation of measures that should be taken in case of privacy violation
  - ➢ Preliminary adjustment of measures that should be taken in case of privacy violation
    - ✧ If any information is provided by AI Provider (including the one from AI Developer) about measures against violation of privacy of individual person, pay close attention to it and consider the measures.

- Delete information that may lead to a violation of privacy of individual person and update AI algorithm (when obtaining any information that may lead to a violation of privacy of individual person).
- Request AI Provider, etc. to delete information that may lead to a violation of privacy of individual person and request AI Developer, AI Provider, etc. to update AI algorithm (when obtaining any information that may lead to a violation of privacy of individual person).

- Input a prompt including personal data
  - For example, when using generative AI services, if any personal data to be input is scheduled to be used as AI training data by a provider of generative AI service, pay close attention not to input any prompt including personal data for which no consent is obtained.
  - Pay close attention to information to be input in AI
    - ✧ Do not give any sensitive information (including not only own information but also others' information), among others, to AI without due reason due to excessive emotional involvement in AI or similar reasons.
  - Respect on privacy
    - ✧ If AI collects and uses data by itself for its learning, respect the privacy of individual persons at the time of collection, etc.

[References]

- Personal Data Protection Committee "Remarks for the Use of Generative AI Services" (June 2023)
- Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry, "Guidebook on Corporate Governance for Privacy in Digital Transformation (DX) ver. 1.3" (April 2023)

[Contents of the main part (repeat)]
- When using AI systems and services

U-5) i. Implementation of security measures

- ✧ Conform to instructions for security specified by AI providers (“5) Ensuring security”).
- ✧ Pay attention not to improperly input secured information into AI systems/services (“5) Ensuring security”).

[Points]

It is desirable for AI business users to pay close attention when using AI system and services if they are provided with information by AI Provider (including the information from AI Developer) about the measures that should be taken in the event of a security violation. In the event of any doubt for security in using AI system and services, they are expected to report to AI Provider (or AI Developer through AI Provider) to that effect.

If non-business users are assumed to implement security measures, AI business users are expected to take necessary security measures in cooperation with non-business users based on the information provided by AI Provider (including the one from AI Developer) by paying close attention to AI system security.

[Specific methods]

- Recognition of risks relating to vulnerabilities
  - Risk of intentional malfunction of AI model by adding minor fluctuation indiscernible by humans to data that can be accurately discerned by AI model as a result of insufficient learning or similar reasons and by entering such data (e.g. adversarial example attack).
  - Risk of erroneous learning by commingling inaccurately labelled data in learning with teacher.
  - Risk of easy reproduction of AI model
  - Risk of reverse engineering the data used for learning from AI model
- Consideration for measures upon occurrence of security violation
  - Initial action
    - ✧ Recovery by rollback of the AI system, use of alternative system, etc.
    - ✧ Stop the AI system (kill switch).
    - ✧ Disconnect the AI system from the network.
    - ✧ Confirmation of contents of security violation
    - ✧ Reporting to relevant stakeholders
  - Use of insurance to facilitate indemnification, compensation, etc.
  - Establishment of third-party organ and investigation and analysis of causes and proposals by such organ
- Input a prompt including secured information
  - For example, when using generative AI services, if any secured information to be input is scheduled to be used as AI training data by a provider of generative AI service, pay close attention not to input any prompt including secured information for which no consent is obtained.
  - Pay close attention to information to be input in AI
    - ✧ Do not give any secured information to AI due to excessive emotional involvement in AI or similar reasons.

[References]

- Information-technology Promotion Agency, Japan “AI Handbook for Parties Involved in Security” (June 2022)
- Information-technology Promotion Agency, Japan “Security by Design Guide Instruction Book” (August 2022)
- ACSC, “Engaging with Artificial Intelligence（AI）” (January 2024)

[Contents of the main part (repeat)]
- When using AI systems and services

U-6) i. Providing relevant stakeholders with information

✧ Input data for which fairness is guaranteed to avoid significant lack of fairness, and pay attention to bias in prompts when obtaining the output result from the AI system or service. When using the output result for business decision-making, provide the relevant stakeholders with information to the reasonable extent. ("3) Fairness," "6) Transparency").

[Points]

AI business users are expected to secure the explainability of the result of AI output by taking into account the social context when using AI, including when using AI in fields where a material impact may be given to individual rights and interests, for the purposes of obtaining satisfaction and assurance of non-business users and showing evidences of AI behaviors to this end (i.e. plain information easy to understand in relation to the factors on which AI system bases its prediction, recommendation or decision and to its decision-making processes). At that time, AI business users are expected to improve the explainability of the result of AI output by analyzing and understanding what explanations are required of them to build and maintain the reliability of individual persons and taking necessary measures.

[Specific methods]
- Providing relevant stakeholders with information
  - Clarification of matters to be explained
    - ✧ Clarify the scope of non-disclosure through execution of a contract with AI Provider in which AI Provider limits the scope of non-disclosure (including the scope set by AI Developer).
  - Test of the manner of explanation before introduction of AI system and service and of the explanation itself
  - Acquisition of feedback for explanation
    - ✧ Acquire feedback for accuracy and definiteness of explanation from stakeholders, including non-business users, and individual persons or groups that may be affected.
  - Provision of information about AI models
    - ✧ Include the provision of information, including the type and source of input data, high-level data conversion process, standards and grounds of decision-makings, risks and measures to mitigate them.
  - Attention to the provision of information
    - ✧ Non-business users will share necessary information in a timely manner.
    - ✧ Information to be provided about AI system and services shall be provided before use thereof.
    - ✧ If the above information cannot be provided before use of AI system and services, develop a system to respond to feedback from non-business users according to the risks assumed in accordance with the nature, mode of use, etc. of AI.

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

[Contents of the main part (repeat)]

- When using AI systems and services

U-7) i. Explanation to relevant stakeholders

- ✧ Provide information, including instructions for proper use, in a plain and accessible manner to the reasonable extent according to the nature of the relevant stakeholders (“7) Accountability”).
- ✧ If planning to use data provided by relevant stakeholders, let the stakeholders know in advance how to provide the data and its formats based on the characteristics and use purposes of AI, contact points with the relevant stakeholders as data providers, privacy policies, and the like (“7) Accountability”).
- ✧ If intending to use the AI output result as a reference for an evaluation of a specific individual or group, notify the specific individual or group to be evaluated about the use of AI, follow procedures for guaranteeing the accuracy, fairness, transparency, etc., of the output result as recommended by the Guidelines, and make a reasonable judgment by humans taking into account automation bias. If the individual or group evaluated demands you to give an explanation, fulfill your accountability by accepting the demand (“1) Human-centric,” “6) Transparency,” “7) Accountability”).
- ✧ In accordance with the characteristics of the AI systems and services to be used, set up a help desk, at the reasonable level, that handles inquiries from relevant stakeholders to give explanations and receive requests in cooperation with the AI providers (“7) Accountability”).

U-7) ii. Effective use of provided documents and conformity to agreements

- ✧ Properly store and use the documents about the AI systems and services provided by the AI providers (“7) Accountability”).
- ✧ Conform to the service agreements specified by the AI providers (“7) Accountability”).

[Points]

AI business users are expected to prepare, publicly report and notify the AI usage policies to enable non-business users to recognize the use of AI appropriately.

[Specific methods]

- Disclosure of usage policies for AI containing the following matters
  - ➢ The fact that AI is being used (Identify the name and details of specific functions and technologies, if possible.)
  - ➢ Scope and manner of use of AI
  - ➢ Grounds for output of AI
  - ➢ Risks associated with the use of AI
  - ➢ Consultation desk
  - ➢ Precautions for disclosure and notification of usage policies
    - ✧ In the event of use of AI in a mode that AI output may directly affect non-business users or third parties, prepare and disclose the usage policies pertaining to AI and make explanations if inquired in order to enable non-business users or third parties to recognize the use of AI appropriately.
    - ✧ In the event of a possibility of material effect on the rights and interests of non-business users or third parties, notify to that effect positively. (AI Provider and AI business users are required to publicly report the usage policies pertaining to AI when the output of AI to be used may directly affect non-business users or third parties. In other words, the usage policies pertaining to AI are not necessarily required to be publicly reported if AI is used just as an analysis tool made available to human thinking or if it is substantially assured that an original idea created by AI is finally judged by humans, provided that a voluntary public report is expected).

- ✧ Notification or public report are expected to be made not only before commencement of use but also upon change in AI behaviors and end of use (especially if assumed risks change due to change in AI behaviors).
- ✧ If AI is used to detect fraudulent acts or if a risk of misuse of AI is concerned about, consider the necessity, details and manner of disclosure or notification to determine whether or not to disclose or notify.

[References]

- 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. Descriptions of “Common guiding principles” in Part 2

Although not mentioned in the Main Part, “Part 5 Matters Related to AI Business Users,” specific methods for the Main Part, “Part 2 Common Guiding Principles,” which are especially important for AI business users, are explained here.

[Contents of the main part (repeat)] *Excerpts of the main paragraph only

1) Human-centric

When developing, providing, or using an AI system or service, each AI business actor should act in a way that does not violate the human rights guaranteed by the Constitution of Japan or granted internationally, as the foundation for accomplishing all matters to be conducted, including the matters described later. In addition, it is important that each AI business actor acts so that the AI expands human abilities and enables diverse people to seek diverse well-being.

- Regarding “(3) Measures against false information, etc.”
  [Relevant Description]
  - Recognize the increasing risk of destabilizing and confusing society through disinformation, misinformation, and biased information generated by AI, as generative AI has enabled anyone to create information that appears to be true and fair, and take necessary countermeasures.

  [Specific methods]
  - Consideration of risk avoidance measures for disinformation, misinformation, and biased information
  - Development and introduction of technology to enable AI users to identify information generated by AI
  - Ensuring the recency of data searched and referenced by RAG (when users prepare the data)
  - Conducting awareness-raising and information literacy education for a wide range of ages

- Regarding “(4) Ensuring diversity/inclusion”
  [Relevant Description]
  - In addition to ensuring fairness, to prevent information poverty or digital poverty and allow more people to enjoy the benefits of AI, pay attention to make it easy for socially vulnerable people to use AI.

  [Specific methods]
  - Mind to leave nobody behind AI use
    - ✧ Improving AI literacy
    - ✧ Securement and development of digital/AI personnel
    - ✧ Development of safe and assured environment of using AI

- Regarding “(6) Ensuring sustainability”
  [Relevant Description]
  - Consider the impact on the global environment throughout the entire lifecycle in the development, provision, and use of AI systems and services.

  [Specific methods]
  - Utilize lightweight models with minimal environmental impact according to the purpose, and appropriately differentiate and use models.

[References]

Appendix 5. For AI Business Users
When developing, providing, or using an AI system or service, each AI business actor should act in a way that does not violate the human rights guaranteed by the Constitution of Japan or granted internationally, as the foundation for accomplishing all matt

- The Ministry of Internal Affairs and Communications “White Paper on Telecommunications 2021 version” (July 2021)

[Contents of the main part (repeat)] *Excerpts of the main paragraph only

2) Safety

Each AI business actor should avoid damage to the lives, bodies, minds, and properties of stakeholders during the development, provision, and use of AI systems and services. In addition, it is important that the environment is not damaged.

- Regarding "(1) Taking into consideration the lives, bodies, properties and minds of humans and the environment"
  [Relevant Description]
  - Determine the responses for cases where the safety of AI systems or services is endangered so that the steps can be quickly taken in such cases.

  [Specific methods]
  - Summarize incident measures and consider measures upon occurrence thereof.
    - ✧ Summarize incident measures.
      - Preliminarily arrange emergency contact system in case of occurrence of harm.
      - Summarize methods of cause investigation and method of recovery works.
      - Consider for measures to prevent recurrence and summarize response policy.
      - Establish the method to share incident-related information.
    - ✧ Initial action
      - Recovery by rollback of the AI system, use of alternative system, etc.
      - Stop the AI system (kill switch).
      - Disconnect the AI system from the network.
      - Confirmation of the details of the harm
      - Reporting to relevant stakeholders
    - ✧ Use of insurance to facilitate indemnification, compensation, etc.
    - ✧ Establishment of third-party organ and investigation and analysis of causes and proposals by such organ

[References]

- Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry, "Guidebook on Corporate Governance for Privacy in Digital Transformation (DX) ver. 1.3" (April 2023)

[Contents of the main part (repeat)] *Excerpts of the main paragraph only

3) Fairness

During the development, provision, or use of an AI system or service, it is important that each AI business actor makes efforts to eliminate unfair and harmful bias and discrimination against any specific individuals or groups based on race, gender, national origin, age, political opinion, religion, and so forth. It is also important that before developing, providing, or using an AI system or service, each AI business actor recognizes that there are some unavoidable biases even if such attention is paid, and determines whether the unavoidable biases are allowable from the viewpoints of respect for human rights and diverse cultures.

- Regarding "(2) Intervention by decisions made by humans"
  [Relevant Description]
  - To prevent AI from outputting unfair results, consider implementing interventions by having humans make decisions, rather than letting AI make the decisions alone.

  [Specific methods]
  - Determination of whether the intervention by decisions made by humans is necessary and ensuring the effectiveness of such decisions

- ✧ Determination of whether the intervention by decisions made by humans is necessary (Below are examples of criteria for determination.)
  - Nature of the rights and interests of AI business users or non-business users who are affected by output of AI and intentions of AI business users or non-business users
  - Degree of reliability of output of AI (superior or inferior to the reliability of decisions made by humans)
  - Temporal grace necessary for decisions made by humans
  - Ability expected of AI business users or non-business users who make decisions
  - Need to protect matters to be decided (e.g. response to individual applications by humans or response to volume applications by AI systems and services)
  - Uncertainty of statistical future prediction
  - Need and degree of satisfactory reasons for decision-makings (decisions)
  - Degree of assumed discrimination based on race, creed, or gender due to the fact that the training data contains social bias against minorities, etc.

➢ Ensure the effectiveness of decisions made by humans
- ✧ On the assumption that an explanation is obtained from AI that has explainability, preliminarily clarify items for which decisions should be made by humans (if it is considered appropriate for humans to make final decisions on output of AI).
- ✧ AI business users or non-business users will acquire necessary ability and knowledge to appropriately assess output of AI (if it is considered appropriate for humans to make final decisions on output of AI).
- ✧ Preliminarily summarize responses to ensure the effectiveness of decisions made by humans.

[References]
- National Institute of Advanced Industrial Science and Technology “Guidelines for Machine Learning Quality Management, 4th Edition” (December 2023)

[Contents of the main part (repeat)] *Excerpts of the main paragraph only

8) Education/literacy

Each AI business actor is expected to provide the persons engaged in AI in the AI business actor with the necessary education to gain the knowledge, literacy, and ethical views to correctly understand and use AI in a socially correct manner. Each AI business actor is also expected to provide stakeholders with education, in consideration of the characteristics of AI, including its complexity and the misinformation that it may provide, and possibilities of intentional misuse of AI.

[Relevant Description]
- Take necessary measures to ensure that the persons engaged in AI in AI business actors acquire AI literacy of the level sufficient for the engagement.
- It is assumed that the division of tasks between AI and humans will change due to the expansion of generative AI use, so education and reskilling, etc. should be actively discussed to promote new ways of working.

[Specific methods]
- Matters that should be contained as the literacy education and skills for use of AI
  - ➢ Knowledge for AI, mathematics and data science
  - ➢ Understand the characteristics of AI and data, that is, the inclusion of bias in data and the possibility of causing bias depending on the way of use.
  - ➢ Understand the issues pertaining to the accuracy, fairness and privacy protection of AI or data and the details of security and AI technology’s limit.
  - ➢ Understand that the data digital technologies including AI are being used for various tasks.

  - Use AI appropriately to improve the productivity by combining it with skills, including “asking questions,” “proposing and testing a hypothesis” and so on.
  - Recognizing the possibility that false information generated by AI is included in information circulating in the media and the necessity to discern and select such information
  - Skills for creating prompts to derive appropriate answers when using generative AI
- Education for improvement of resilience to environmental changes, etc.
  - Mental rotation that flexibly switch between “vertical thinking” and “lateral thinking”
  - Enhance the modelling skill required for organizational evaluation.
  - Use an agile thinking to extend the learning horizontal line of a field of skill one is bad at.
  - Improve the evaluation skill to analyze uncertainty that is difficult to predict with past experiences or know-hows.
  - Convert the “important points” of organizational evaluation (to an agile “instantaneous force plus vitality”).
  - Method of evaluation of resynchronization (Configuration, Architecture, Synthesis and Dissemination) of more complexed and advanced management in situations where the AI governance is formed in a hybrid of centralization and dispersion.

[References]

- The Ministry of Economy, Trade and Industry, Information-technology Promotion Agency, Japan “Digital Skill Standards ver. 1.1” (August 2023)
- Ministry of Internal Affairs and Communications, “Guidebook for AI Utilization and Introduction in Local Governments <Supplementary Appendix> Case Studies of Generative AI Introduction in Leading Organizations” (July 2024)
- Ministry of Internal Affairs and Communications, “Collection of Multistakeholder Initiatives on Countermeasures Against False and Misleading Information on the Internet” (May 2024)

Appendix 6. Major precautions for referring to “Contract Guidelines on Utilization of AI and Data”

As explained in the Main Part “Part 2,” multiple actors are involved in each situation of the development, provision and use of AI. Therefore, it is expected to set forth the rights and obligations of the parties in a contract as clearly as possible for each transaction relating to the development, provision and use of AI as the guiding principles for settlement of dispute, if any, in order to facilitate respective transactions and prevent needless dispute associated with such transactions.

“Contract Guidelines on Utilization of AI and Data” whose initial version was formulated and publicized in June 2018[116] (in December 2019, a partly updated version 1.1 was publicized; hereinafter referred to as the “Contract Guidelines”) summarizes basic concepts for contracts for development/use of software using AI, contracts for provision/use of data, matters that should be understood in advance as preconditions of these contracts, and so on against a background of issues of those days.

The Contract Guidelines were formulated in a trend where the development of AI would be more promoted and put to practical use and, under the objective of the Guidelines promoting the development and use of AI, the following issues were listed that should be resolved through the Guidelines.

- Regarding contracts for provision/use of AI and data, practical operations are not accumulated enough.
- There is a gap between concerned parties for recognition and understanding of the technological characteristics of AI and the value of data and AI development know-hows.

When the Contract Guidelines were formulated, the removal of obstacles to transactions between a party developing the software using AI and a party using the outcomes of development was considered as an important issue under the objective of facilitating such transactions to promote the development and practical use of AI.

Six years passed from the formulation and publication of the initial version of the Contract Guidelines and, during that span, situations relating to the development and use of AI remarkably progressed and new technologies and ways of use were created day-to-day, besides many AI-related technologies are entering a phase of prevalence in the society. It is important to pay close attention to the fact that, due to this background, the Contract Guidelines contains both the contents, reference to which remains useful, and the contents, for which situational changes after publication should be considered.

As an example, among other contents of the Contract Guidelines, the following contents primarily referred to in the AI Part, Part 2 (Explanation of AI Technology) and Part 3 (Fundamental Concepts) and Data Part, Part 3 (Legal Fundamental Knowledge for Assessment of Data Contract) are considered to be as useful as before in general when referred to.

(AI Part)
- Features of development of software using AI and characteristics of AI
- Summarization of intellectual property rights, etc.
- Fundamental viewpoints for belongingness of rights and setting of usage conditions
- Fundamental viewpoints for distribution of responsibilities

[116] The Ministry of Economy, Trade and Industry “Contract Guidelines on Utilization of AI and Data version 1.1” (December 2019), https://warp.da.ndl.go.jp/info:ndljp/pid/12685722/www.meti.go.jp/press/2019/12/20191209001/20191209001-1.pdf

(Data Part)
- Legal nature of data and means of protecting data

Reference to explanations of various contract models contained in the Contract Guidelines is considered to be still useful but, as the transactions relating to the development, provision and use of AI are more diversified in comparison with when the Contract Guidelines were formulated, the difference between various contract models and actual transactions needs to be assessed carefully.

It is important to pay close attention especially to the following points as the matters for which situational changes after publication of the Contract Guidelines should be considered.

(1) Diversified contract models

Based on the dichotomy theory consisting of a party developing AI (i.e. vendor) and a party using AI (i.e. user), the Contract Guidelines provide and explain two model contracts, that is, a development contract for (1) a transaction in which a user commissions a vendor to develop AI; and a usage contract for (2) transaction in which a vendor allows a user to use AI developed by the vender.

This categorization is considered to be still applicable presently, as it is, to some transactions for development or use of AI. However, since 2022, with the emergence of generative AI possessing versatility, AI services utilizing such technology have rapidly proliferated. Consequently, the number of contracts for utilizing generative AI services (general-purpose AI) in various scenarios has increased, as well as contracts where a vendor customizes general-purpose AI services from other vendors to provide them to users, whereby transactions not fitted into any category sorted out in the Contract Guidelines become more and more important.

Contracts related to the use and development of AI can be broadly classified into the following three types, with each type having different contractual considerations and negotiation points.

- Type 1: Usage of General-purpose AI Service

This involves cases where AI users utilize general-purpose AI services provided by AI providers.

- Type 2: Customization

This involves cases where AI users utilize AI services customized by AI providers for specific AI users (customized services). The AI provider offering customized services combines additional functions (non-AI models) they developed with general-purpose AI services provided by other vendors.

- Type 3: New Development

This involves cases where AI users collaborate with AI developers and AI providers to develop and utilize unique AI systems.

To address these market changes, the Ministry of Economy, Trade and Industry formulated a checklist in February 2025, which outlines considerations for contracts related to AI utilization by Japanese businesses (hereinafter referred to as the “Contract Checklist”).[117] For contract types that emerged after the publication of the Contract Guidelines, please refer to the Contract Checklist.

[117] The Ministry of Economy, Trade and Industry’s “Contract Checklist for AI Utilization and Development” (Study group on contractual considerations for the utilization of AI, February 2025)
（https://www.meti.go.jp/press/2024/02/20250218003/20250218003.html）

Furthermore, with the anticipated further advancement of AI technology, it is possible that the number of contracting parties will increase, making contractual relationships more complex and potentially leading to new contract types. In such cases, it is important to review conditions based on the actual circumstances of each transaction, while referring to the concepts presented in the Contract Guidelines and Contract Checklist.

(2) Risk distribution under a complex value chain

For either type of (1) transaction in which a user commissions a vendor to develop AI or (2) transaction in which a vendor allows a user to use AI developed by the vender, the Contract Guidelines spare much of description in rearranging interpretation about the legal relationship between vender and user, especially the belongingness and use relation for outcomes, and burden of risk of damage to concerned parties and infringement of intellectual property rights and other rights of third party, keeping in mind an adjustable simple interest relation model between vendor and user.

Recently, however, as mentioned above in (1), AI’s value chain tends to diversify or become more complex, and recently, various business operators have become involved in the development, provision, and use of AI. Besides, as AI prevails not only to business activities but also to daily life, the use by non-business users is increasing., resulting in occurrence of problems which cannot be caught sufficiently only by looking at a bilateral relationship between vendor and user.

(Examples of business operators)
- Business operator who develops AI (e.g., a business operator that develops the generative AI model itself. AI Developer in these Guidelines)
- Business operator who develops software incorporating a developed AI (e.g., a business operator that develops software to enable the use of developed generative AI in a chat format. AI Developer in these Guidelines)
- Business operator who provides outside parties with such software (e.g., a business operator that provides software to general consumers to use generative AI in a chat format. AI Provider in these Guidelines)
- Business operator who provides outside parties with services for which the provided software has been customized (e.g., a business operator that provides services specialized for specific applications to companies using existing generative AI models. AI Provider in these Guidelines)
- Business operator who uses such services (AI Business User in these Guidelines)

※Note that many business operators may hold multiple positions.

One of those issues is how to distribute the responsibilities on the value chain of AI. For example, in cases where a service customizing software incorporating AI is provided to non-business users, the issue may be who should assume a risk of damage arising out of AI which may occur to such non-business users. Details and degree of those risks are largely affected not only by the quality of AI but also by the way of providing software and the significance of provision and use of services. It may be difficult to set reasonable boundary to the scope of responsibilities without looking at the role of each player in the value chain beyond the bilateral relationship between vendor and user.

The context of this reasonability of boundary poses a problem that a party who cannot directly control a risk assumes the risk. Taking the above case as an example, if AI Developer is fully liable for any damage arising out of AI irrespective of the way of provision and use, AI Developer will have to assume a risk it cannot directly control. This situation tends to take place between

parties with different negotiating powers, while influential AI sometimes should think extensively about the scope of risks AI Developer should control.

The Contract Guidelines put focus on bilateral transactions between vendor and user and the way of distributing responsibilities based on this diversified or complexed value chain needs to be considered according to individual situations. In this respect, the practical examples in Appendix 2, 3. System Design (Building of AI management system) may be helpful.

(3) Allocation of responsibility and accountability

As mentioned above in (2), when considering risk distribution among multiple parties, it is necessary to analyze new types of risks as well. As AI becomes popular and applied more and more, it is expected that risks associated with the development, provision and use of AI will increasingly come to the surface.

Those risks include a risk of accident related to software incorporating AI and services customizing it, which may cause damage to a party who develops, provides and uses AI or third parties. However, currently, there are no clear standards regarding which entity bears what kind of responsibility in the event of an accident, and there may be cases where businesses abandon or hesitate to introduce AI due to the difficulty in outlining risk scenarios. For these new types of risks, it is important to distribute risks cooperatively among the relevant parties rather than imposing all risks on one party. In some cases, the use of insurance, such as liability insurance, may also be useful. The following are beneficial points to consider contractually in relation to accident risk.

- Organizing new types of risks
  There have already been cases domestically and internationally where accidents have occurred due to intellectual property rights infringement, violation of personal information protection laws, and breach of confidentiality agreements associated with the use of AI services. It is important to identify potential risks according to the content and nature of AI services, considering precedents as above and the trend of discussions, and to define allocation of risks (which entity bears what kind of responsibility). On the other hand, since accidents that were not anticipated at the time of the contract may occur, it is also important to review the contract content as necessary, taking changes in the surrounding environment into account.

- Reasonable Explanations
  In the event of an accident, important points of discussion include what caused the accident (it may be due to the actions of AI developers, AI providers, and AI users, or it may be something that arises inevitably due to the nature of AI) and whether each party exercised due care to avoid the accident. Parties who develop, provide or use AI may be required to give reasonable explanations as to how they were involved in respective processes. Party who is primarily responsible for the accident may be held responsible for such explanations irrespective of whatever contract was executed among all parties who develop, provide or use AI. What can be set forth in a contract is limited to the sharing of responsibilities only among parties to the contract. All parties in AI value chain may be put in a position where they are required to give explanations to a certain level if they are held by non-contract parties accountable on the accident.

- Presentation of Objective Evidence
  To provide reasonable explanations, objective grounds in addition to the details of explanations are needed and it is expected to sort out such grounds before and after execution of a contract for development, provision and use of AI. Although not mentioned in the Contract Guidelines, it is useful to consider responses after execution of contract by referring to practical examples for Appendix 2, 3. System Design (Building of AI management system). When presenting objective evidence, there may be an

approach to disclose information that is beneficial for monitoring (such as log data). However, it is also necessary to be aware that the disclosure of such internal materials may stand in a trade-off relationship with risks to security and competitiveness.