Governance Guidelines for Implementation of AI Principles

Ver. 1.1

January 28, 2022

Expert Group on How AI Principles Should be Implemented

AI Governance Guidelines WG

Table of Contents

A. Introduction

1. The Aim of the AI Governance Guidelines

In March 2019, Japan published the document called “Social Principles of Human-centric AI” adopted by the Integrated Innovation Strategy Promotion Council, which contributed to the formulation of the OECD’s recommendations on Artificial Intelligence. The document set social principles for AI that is to be implemented in the society as a whole, and it states that, based on such principles, companies involved in AI business typically as a developer and operator should establish and comply with the goals to be implemented (AI development and utilization principles) according to the purpose and method of their AI business such as development and operation.

The social principles for AI are comprised of seven principles: (1) Human-centric, (2) Education/Literacy, (3) Privacy Protection, (4) Ensuring Security, (5) Fair Competition, (6) Fairness, Accountability, and Transparency, and (7) Innovation. The Governance Guidelines for Implementation of AI Principles (hereinafter referred to as the “AI Governance Guidelines”) present **action targets to be implemented** by an AI company, with the aim of **supporting the implementation of the AI principles** that is required for the **facilitation of deployment of AI**. In addition, **hypothetical examples of implementation corresponding to each of the action targets** and **practical examples for gap analysis between AI governance goals and current state** (hereinafter referred to as “examples for gap analysis”) are also shown. However, the examples of implementation and examples for gap analysis are for reference purposes only and are not intended to be exhaustive.

2. Legal Effect of the Guidelines

The Guidelines themselves are **not legally binding**. The Guidelines are comprised of action targets to be implemented, examples of implementation, and examples for gap analysis, etc., which summarize typical targets and practical examples that are shared in society to a certain extent. As the “Social Principles of Human-centric AI”, which are also not legally binding, are respected in society due to their universal nature, the Guidelines are also **expected to support company’s voluntary efforts as widely shared material that is referred to by AI companies who are involved in AI business, typically the development/operation of AI systems, in their business transactions, and through the development of a common understanding among stakeholders on the implementation of AI principles**. Please note that even if a company establishes structures, etc. in accordance with the Guidelines, it does not necessarily mean that the company complies with relevant laws, therefore please be mindful of complying with the relevant laws.

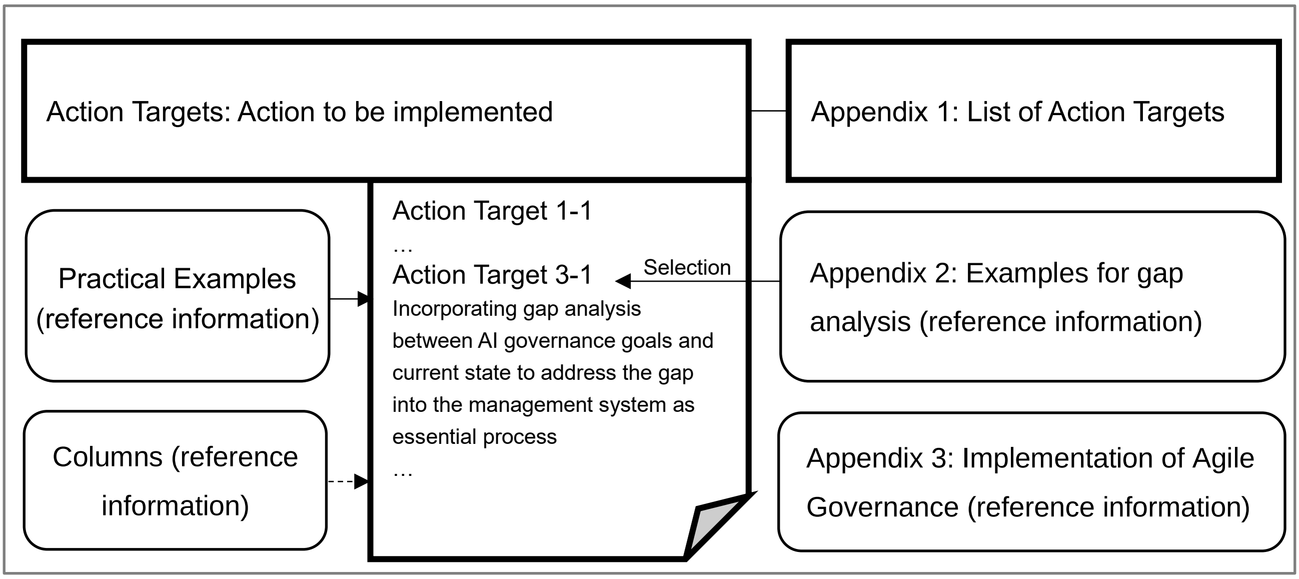
3. Relationship with other guidelines, etc.

While new elements have been added, the Guidelines focus on incorporating essence of various codes, guidelines, assessment lists and other documents published in and outside Japan, with the **aim of integrating the relevant guidelines** (which amounts to **guidelines of guidelines**). By cross-referencing a variety of relevant documents, the Guidelines are designed to serve as comprehensive guidelines as a whole.

4. How to Use the AI Governance Guidelines

The Guidelines are comprised of the main part (action targets, examples of implementation, and columns) and Appendices (list of action targets, examples for gap analysis (embodiment of Action Target 3-1), and implementation of agile governance). Among these, **the action targets are general and objective ones that should be implemented by every AI company involved in AI business, typically the development/operation of AI systems that could have a certain level of negative impacts on society** (relations between action targets will be discussed on pages 7 and 8). On the other hand, the examples of implementation and examples for gap analysis do not take into account any specific circumstances of individual AI company. The examples for gap analysis, in particular, could be insufficient or oversufficient depending on objectives and methods of AI system development, operation, etc. as well as subject under analysis. Therefore, the **decision on whether to adopt the examples of implementation and for gap analysis is obviously left up to the discretion of AI company, and if they decide to adopt those examples, they need to consider making modifications or selecting appropriate examples according to their circumstances**.

Illustration of the Structure of the Guidelines



Action Targets: Action to be implemented

Appendix 1: List of Action Targets

Practical Examples (reference information)

Columns (reference information)

Action Target 1-1
…
Action Target 3-1
Incorporating gap analysis between AI governance goals and current state to address the gap into the management system as essential process
…

Appendix 2: Examples for gap analysis (reference information)

Appendix 3: Implementation of Agile Governance (reference information)

5. Living Document

AI technologies are in the midst of rapidly evolving. Furthermore, it is anticipated that society will accumulate knowledge to maximize positive impacts brought by AI systems, while managing negative impacts at a level acceptable to stakeholders. Therefore, it is unlikely that the Guidelines will adequately function in the future without revisions being made. For the improvement of AI governance, **it is essential to continue reviewing how AI governance and the Guidelines should be and make necessary revisions** through multi-stakeholder participation, while referring to the design concept of agile governance.

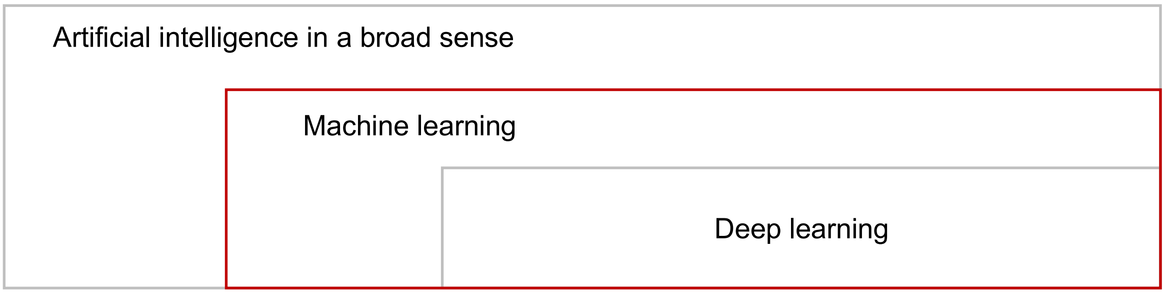
B. Definitions

Terms used in the Guidelines are defined as follows:

The scope of the Guidelines includes **AI systems in which a machine learning approach is used and which are at least partially created by inductively using data**, as shown below. However, **for software and other systems which could replace human decision making and in which the process of such decision making is less visible to users, the Guidelines are expected to be referred to as necessary, even if a machine learning approach is not used**[1].

AI system: A system that is developed with a machine learning approach, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning and that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy. It includes not only software but also a machine which contains software as an element[2].

The area in the red box is the scope of the Guidelines



The target of the Guidelines is **AI company** (AI system developers (i.e., companies that develop AI systems), AI system operators (i.e., companies that operate AI systems), and data providers).

AI system developer: An entity that develops an AI system for its own use or to provide it to others as a business (including an entity that conducts re-training, for example, to maintain the performance of an AI system).

AI system operator: An entity that operates an AI system for its own use or for the use of others as a business (for example, it includes an entity which do not engage in AI system development, but simply procure and operate an AI system), and is responsible to a certain extent for the operation of the AI system and/or maintenance of its performance. Although such an entity is not

[1] There is currently no clear definition of artificial intelligence (see "Social Principles of Human-centric AI" adopted by the Integrated Innovation Strategy Promotion Council (March 29, 2019)), and it is not appropriate to strictly define the scope of artificial intelligence in a broad sense.

[2] Guided by the OECD's definition of AI system. In the proposed AI regulation of the European Commission, AI system simply means software. The OECD's definition of an AI system is not limited to software.

necessarily the legal right holder of the AI system, it is generally believed that in many cases they are the same entity.

AI system user: An entity that simply uses an AI system developed by an AI system developer or an AI system provided by an AI system operator, and that is not responsible for the operation of the AI system and/or maintenance of its performance. Note that, while **AI system users include those who use AI systems for business purposes and consumers who use AI systems for non-business purposes**, in the Guidelines we believe **it is preferable to understand users in a more flexible manner based on their literacy levels** rather than explicitly dividing the users into the two groups[3].

Data provider: An entity that, as a business, provides others with data collected from a number of unspecified sources, data collected from specified people, data prepared by the data provider itself; a combination of them; or data created by processing the above-mentioned data, for the purpose of AI system training, etc.

* A single AI company may be classified into multiple roles at the same time. For example, if the development and operation of an AI system is performed by the same company, such a company is an AI system developer and an AI system operator.

The structure of the Guidelines assumes that there are two layers of implementation bodies within a company. Please note that in a small company where the line between those two layers is not clear, it may be understood that a certain person or group bears the responsibility for both.

Top management: A person or group that is responsible for ensuring appropriate disclosure of information to various stakeholders including shareholders and non-shareholders; ensuring collaboration with stakeholders; and giving an overall direction to the operation level to establish a management system so that sound ethical standards for business are respected.

Operations: A person or group that designs and operates the management system according to the direction given by the top management so that sound ethical standard for business are respected, and that conducts or supports analysis of conditions and risks, goal setting, and evaluation of the management system.

[3] For user literacy, the OECD's "OECD FRAMEWORK FOR THE CLASSIFICATION OF AI SYSTEMS – PUBLIC CONSULTATION ON PRELIMINARY FINDINGS" (May 2020) provides a certain level of guideline. The document classifies users as follows: an amateur who has no training; a user with some specific training on how to use a certain system; and an expert with specific training and knowledge of AI. https://aipo-api.buddyweb.fr/app/uploads/2021/05/Report-for-consultation_OECD.AI_Classification_final.pdf.

C. AI Governance Guidelines

While AI systems can have a positive impact on companies, for example, helping them overcome HR shortages, improve productivity and develop high-value-added businesses, because the development and operation of AI systems is accompanied by risks unique to AI, such as unintentional impairment of fairness or safety issues, companies that develop and operate AI systems should take these issues as your own matter in understanding the overall conditions associated with AI systems and value provided by AI system. In implementing the Action Targets, the Guidelines expect that companies will **understand the gist of each of the Action Targets and make use of these Guidelines as a companion to their efforts instead of rigid adherence to them**. **And, once the gist of the Action Targets is understood, companies will most likely see that they are standard matters that the companies should implement, and at the same time are flexible in their implementation**.

These Guidelines begin with **conditions and risks analysis.** When deciding on a policy for a company (or for a business unit in some cases), the company should examine the **positive and negative impacts that the AI system can produce**, **the social acceptance related to the AI system's development and operation**, and, if the negative impacts are found not to be minor in light of scope of the company's business, etc., **their AI proficiency (how much the company is ready for development and operation of AI systems)**.

Based on these lines of thought, companies likely eventually formulate, for example, policies not to develop or operate AI systems in light of the significance of their potential negative impacts, the company's lack of experience in the field of AI, and current social acceptance; policies to limit their AI system development and operation to areas where potential negative impacts are minor; or policies to develop and operate AI systems while managing their potential negative impacts. And, when developing or operating AI systems, companies should consider the nature and significance of potential negative impacts and **examine whether to set AI governance goals for the company (or for a business unit in some cases)** that will serve as a compass for managing these impacts to levels that are acceptable to stakeholders. **If a company decides not to set AI governance goals** based on the assessment that their potential negative impacts are minor, **they should be prepared to explain their rationale to their stakeholders**.

Then, companies need to **design an AI management system** to achieve their AI governance goals. Specifically, the following targets are standard actions: **analyzing a gap between AI governance goals and current state and addressing the gap**, **improving literacy of human resources responsible for the AI management system**, **reinforcing the AI management through cooperation between companies/departments**, for example, sharing information properly, and **reducing incident-related burdens of AI system users** by preventing incidents and providing early response to incidents.

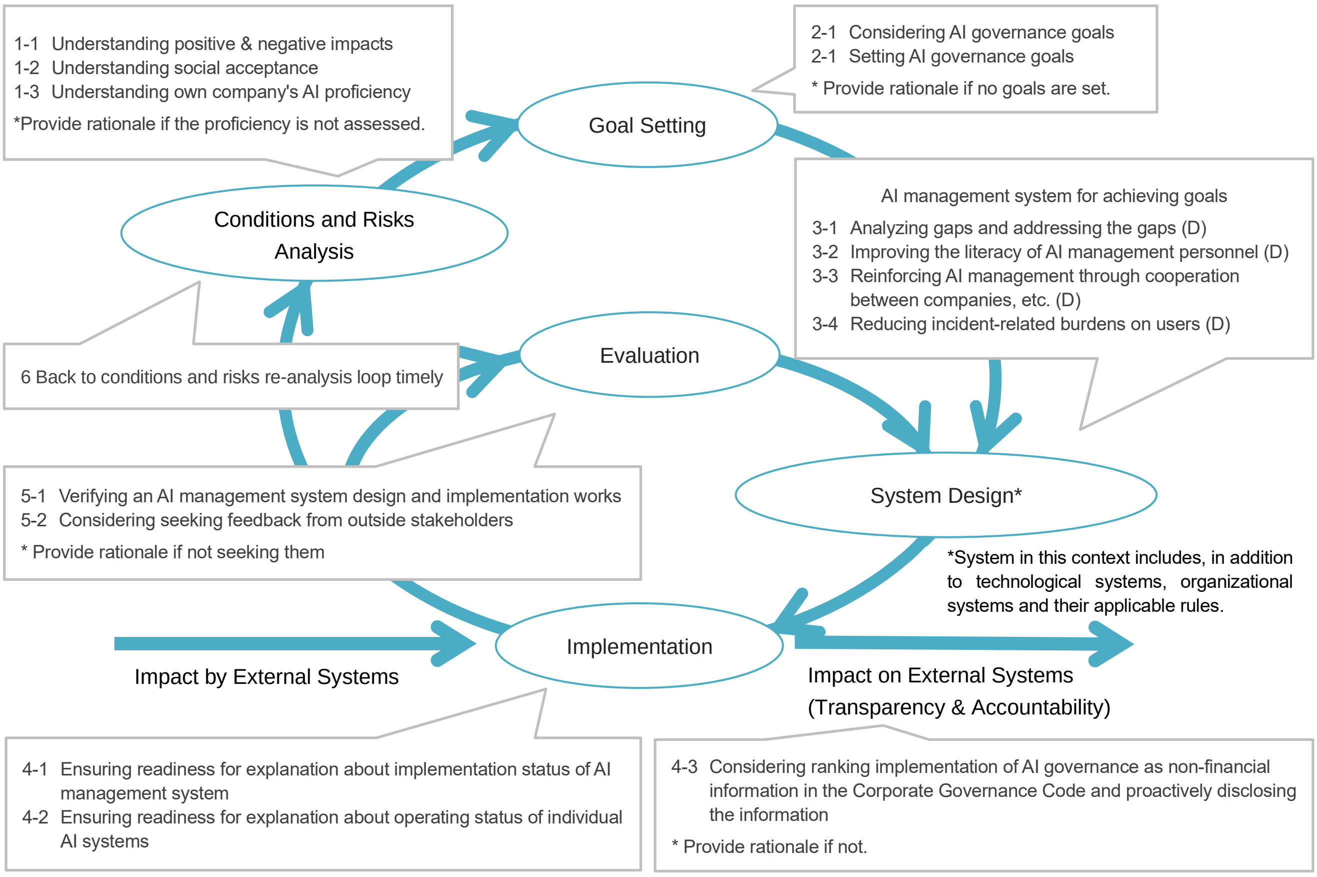
In order to agilely adapt to the development of AI technology, AI governance goals and AI management systems need to be **continuously evaluated**. The first two targets are **to ensure readiness for explanation about implementation status of AI management systems and operating status of individual AI systems externally**. In addition, **companies should consider ranking the information related to AI governance as non-financial information in the Corporate**

Governance Code and proactively disclosing such information for better communication with their stakeholders, and if they decide not to disclose such information, they should be prepared to explain their rationale.

Next, companies should have **individuals independent of the design and operation of the AI management system evaluate the validity of its design and operation**. While it goes without saying that **in-house evaluations on the system's validity should be carried out** using the above-mentioned information on the system's implementation status, **companies should also consider seeking opinions of** not only their shareholders, but also other **stakeholders** including business partners, consumers, and experts who are familiar with trends on the appropriate AI system operation and may actively set up opportunities to obtain such opinions as necessary.

Furthermore, in order to verify the validity of the AI governance goals themselves, **conditions and risks** consisting of positive and negative impacts that AI systems can have, social acceptance associated with AI system development and operation, and the company's AI proficiency **should be re-analyzed in a timely manner**.

Agile governance for AI system developers and operators (Data providers are mentioned in (D))


Goal Setting
Conditions and Risks Analysis
Evaluation
System Design*
Implementation
Impact by External Systems
Impact on External Systems (Transparency & Accountability)

1-1 Understanding positive & negative impacts
1-2 Understanding social acceptance
1-3 Understanding own company's AI proficiency
*Provide rationale if the proficiency is not assessed.

2-1 Considering AI governance goals
2-1 Setting AI governance goals
* Provide rationale if no goals are set.

AI management system for achieving goals
3-1 Analyzing gaps and addressing the gaps (D)
3-2 Improving the literacy of AI management personnel (D)
3-3 Reinforcing AI management through cooperation between companies, etc. (D)
3-4 Reducing incident-related burdens on users (D)

*System in this context includes, in addition to technological systems, organizational systems and their applicable rules.

4-1 Ensuring readiness for explanation about implementation status of AI management system
4-2 Ensuring readiness for explanation about operating status of individual AI systems

4-3 Considering ranking implementation of AI governance as non-financial information in the Corporate Governance Code and proactively disclosing the information
* Provide rationale if not.

5-1 Verifying an AI management system design and implementation works
5-2 Considering seeking feedback from outside stakeholders
* Provide rationale if not seeking them

6 Back to conditions and risks re-analysis loop timely

1. Conditions and Risks Analysis

(1) Understanding positive and negative impacts that AI systems may have

Action Target 1-1: Companies that develop and operate AI systems should, under the leadership of top management, understand not only positive impacts but also negative impacts, including unintended risks, that AI systems may have. This information should be reported to the top management and shared among those in top managerial positions, and their understanding should be updated in a timely manner.

[Practical Example 1]

AI systems have positive impacts, for example, helping companies create new businesses, add value to existing businesses, and improve productivity, but we must also remember that they may have negative impacts. In order to maximize the benefit of these positive impacts, we need to understand the negative impacts and unintended risks and examine how to balance them with the positive impacts. For this reason, **companies that develop and operate AI systems should examine, under the leadership of top management (in other words, by creating momentum within the organization and supporting the activities of those in operations positions), not only the positive but also the negative impacts that AI systems may have. The results from the examination should be shared among those in top managerial positions, and their understanding must be updated in a timely manner**.

While it is reasonable to assume that positive impacts of AI systems are well known among companies that are going to develop and operate AI systems, **we have reviewed the positive impacts that AI technology can have** by referring to comprehensive and exhaustive materials such as "The AI White Paper" compiled by the Information-Technology Promotion Agency (IPA).

In addition, **we investigated whether any incidents were reported in the past regarding AI systems whose functions or fields of application are the same or similar to those of the system that we are going to develop and operate, and even if there were no reports of such incidents, we examined whether there were reports anticipating specific possibilities of incidents**. Incident information is available from a variety of documents and on the Internet. We started by collecting information shared in Japan, as we plan to develop and operate only in Japan. To this end, the "**AI Utilization Handbook - How to Use AI Wisely -" published by the Consumer Affairs Agency can be a good starting point**[4]. For example, as a point of concern for consumers, the handbook states that "AI may mis-recognize voices and give incorrect instructions, or collect information from our day-to-day conversations," which can be said as a potential incident seen from the consumer's perspective. Furthermore, **there are also a wide variety of books on AI** that describe

[4] Consumer Affairs Agency "AI Utilization Handbook - How to Use AI Wisely -" (published in July 2020), available only in Japanese. https://www.caa.go.jp/policies/policy/consumer_policy/meeting_materials/review_meeting_004/ai_handbook.html.

incidents and those that may occur in the future. The Japan Deep Learning Association (JDLA)'s G-Test covers ethical matters, and information associated with incidents can be acquired as a part of the G-Test certification. The "Interim Report Attached to the Draft Recommendation on Profiling" provides easy-to-understand explanations of a number of cases[5]. Furthermore, **while recognizing that social acceptance of AI systems can differ from country/region to country/region, we also referred to the incident database listed in "Column: Sharing Incidents" later**. In our analysis, we see that many incidents relate to the management of personal information, fairness, and security. Moreover, we plan to perform in-depth impact analyses of individual AI systems through the gap analysis, as described in Action Target 3-1.

[Practical Example 2]

Because the AI systems that we develop and operate span across a diverse range of fields, **in addition to Practical Example 1, we have roughly classified incidents that have had a negative impact on society and potential negative impacts that some anticipate, in light of general frameworks in order to develop an overall picture about them**. While we currently use our own framework, we have recently been paying close attention to the progress of discussions on the OECD classification framework[6]. We understand that the official version will be published between September and October 2021. The CONTEXT chapter, which generally corresponds to conditions and risks analysis, presents a general framework from the perspective of the relationship between the OECD AI principles and industrial sectors, as well as from the perspectives of business function, potentially impacted stakeholders, and scope of impacts. We are currently looking into how we can incorporate the classifications into our framework, keeping in mind that they are no more than ancillary tools for gaining rough understanding of negative impacts. Moreover, we plan to perform specific impact analyses of individual AI systems through the gap analysis, as described in Action Target 3-1.

[Practical Example 3]

We are developing and operating a wide range of AI systems, so we are aware that any incidents with these systems, if they occur, may have a significant impact on society. For this reason, we carry out analyses at our in-house, cross-functional study group, etc., based on our understanding that **analyses of positive and negative impacts of AI can be made more useful by combining information gleaned from our own experience with information from the experience of other companies in our industry, and in some cases, from the experience in other industries**. And, by continuing these analyses at a certain interval, **we are able to look into revising our AI governance goals in a timely manner before an incident occurs**.

[5] Personal Data +α Study Group "Interim Report Attached to the Draft Recommendation on Profiling" (Dec 19, 2018), available only in Japanese.

[6] OECD, OECD FRAMEWORK FOR THE CLASSIFICATION OF AI SYSTEMS – PUBLIC CONSULTATION ON PRELIMINARY FINDINGS (May 2020), https://aipo-api.buddyweb.fr/app/uploads/2021/05/Report-for-consultation_OECD.AI_Classification_final.pdf.

Column: Sharing incidents

It is often pointed out that an effective way to address the issue of negative impacts that may result from AI system development and operation is in most cases to learn from past incidents. Because AI systems are built inductively based on data sets and many of their negative impacts are unintentional, it is useful to understand past incidents in order to reduce the negative impacts. While cases of these incidents are typically sourced from public information such as news reports, research papers, etc., it can be quite difficult to access the information that we need.

To address this issue of accessibility, the Partnership on AI released "The AI Incident Database (AIID)" in November 2020[7]. More than 1000 incidents are posted on AIID with their URL links, and a search app is also provided. In addition to the Partnership on AI, an AI Incident Tracker is available on GitHub[8].

However, keeping this type of database going appears to be a challenge. Most of the AIID incident cases reportedly consist of initial lists provided by academics. Some point out that, as more and more AI systems are provided and the volume of information increases, accumulating information with a focus on important information will also become a challenge. In addition, some point out that actively collecting incident cases to make them "shared property" is not easy because unpublished "near misses" at companies constitute important experience for these companies, and information on these "near misses" can in some cases be deemed to be intellectual property of these companies.

[7] Sean McGregor, When AI Systems Fail: Introducing the AI Incident Database (November 18, 2020), https://www.partnershiponai.org/aiincidentdatabase/.

[8] jphall663, awesome-machine-learning-interpretability, https://github.com/jphall663/awesome-machine-learning-interpretability/blob/master/README.md#ai-incident-tracker.

(2) Understanding social acceptance of AI system development and operation

Action Target 1-2: Companies that develop and operate AI systems should, under the leadership of top management, understand the current state of social acceptance based on opinions of not only direct stakeholders, but potential stakeholders before full-scale provision of the AI systems. In addition, even after the full-scale operation, companies should obtain opinions of stakeholders again and update their perspectives in a timely manner.

[Practical Example 1]

Under the leadership of top management, companies that develop and operate AI systems should understand the current state of social acceptance based on the opinions of potential stakeholders. **Because AI is a relatively new technology**, we should be acutely aware that **there will tend to be differences in how users and companies that develop and operate AI systems understand these systems.**

We referred to **consumer questionnaire surveys** published by governments, public institutions, think tanks, etc., **as our first lead**. For example, **the Consumer Affairs Agency**, in the AI Working Group under the Committee on Challenges to Consumer Digitalization, **carried out questionnaire surveys and published their results** on (1) the status of consumers' understanding of AI, (2) consumers' expectations for AI, their perceived challenges with respect to AI, and their intentions to use AI, (3) AI-provided services that consumers use (and the types of risks they are concerned about), and (4) how much they recognize and understand the risks associated with the AI services that they use[9]. **Since we are considering international expansion, we also referred to questionnaire surveys of overseas consumers**[10]. In addition, **we also referred to the views of citizens' groups regarding AI systems**.

Because information on social acceptance obtained here will eventually be used in the overall design of AI governance, we need to remove incidental details and distill it to core information so that management can use them in their decision making. At our company, based on information and analyses obtained in Action Target 1-1, **we have developed a risk-based overview of social acceptance by, for example, categorizing various AI systems, in accordance with the magnitude of their negative impacts**, for example, into application that is highly unlikely to be understood and accepted by society regardless of any explanation that we can provide, application that is likely to be understood and accepted by society if we actively provide sufficient explanation,

[9] The Consumer Affairs Agency carried out consumer questionnaire surveys on AI in their AI Working Group under the Committee on Challenges to Consumer Digitalization. For example, the results from the Working Group's Second Consumer Questionnaire Survey are available at the following website: https://www.caa.go.jp/policies/policy/consumer_policy/meeting_materials/assets/consumer_policy_cms101_200616_1.pdf. It is available only in Japanese.

[10] For example, BEUC, etc. https://www.beuc.eu/publications/survey-consumers-see-potential-artificial-intelligence-raise-serious-concerns/html.

application that is likely to be understood and accepted by society if we provide explanations as necessary, and application that is unlikely to have a negative impact on consumers.

[Practical Example 2]

In addition to Practical Example 1, **we actively send personnel to seminars and conferences on AI ethics and quality** that are organized by universities and industry organizations. In recent years, these seminars are often held in a webinar format, and this has allowed us to acquire information more efficiently than in the past. And having access to webinars overseas enables us to understand international trends in AI ethics and quality.

[Practical Example 3]

We had practiced those in Practical Example 2 until recently. But since we are engaged in the full-scale and wide range of development and operation of AI, we understand that our stakeholders have relatively high expectations for our proper use of AI. For this reason, under the leadership of top management, we switched from a policy of acquiring information on our stakeholders' views indirectly and passively, to one where we do so directly and actively.

Under this new policy, **we invite experts who are familiar with social acceptance of AI and hold regular expert meetings**. **We draw on the expert meetings** not only to have them assess our AI management system and our operations, but also **to deepen our understanding of the conditions in which we operate, such as general social acceptance of AI**. In addition, compared to the general information that we obtain in Practical Examples 1 and 2, we understand that a characteristic of the **information we gain in the expert meetings is that it has been explored in-depth for our company** and is in many cases not widely known. Then, we carry out detailed risk-based analyses of social acceptance by combining the information obtained at the expert meetings with general information obtained in Practical Examples 1 and 2. The results of the analyses are organized by those in operations positions, and then they report them to top management (those in charge of business execution).

[Practical Example 4]

The actions we undertake at our company are generally the same as those described in Practical Example 3 but differ in that **we report the views obtained in expert meetings directly to our board of directors**. We recognize that our **top management have heightened their sensitivity to AI ethics and quality because they now have direct access to information that has been explored in-depth for our company**. For example, we see tangible effect of having AI ethics and quality elevated to mandatory training subjects for all employees. These mechanisms are effective in motivating top management to take ownership of AI ethics and quality issues.

Column: Improving AI proficiency

Before moving on to Action Target 1-3, we would like to provide general information on AI proficiency. The Landing AI Playbook, compiled by Andrew Ng, co-founder of Google Brain and professor at Stanford University, presents a model for improving AI proficiency that a company starts off by aiming for small but meaningful successes and then leverages these successes to extend AI systems in-house and to third parties. At the RIETI's BBL seminar “Forefront of Deep Learning and Challenges for Utilization,” the speaker, Takeshi Izaki, pointed out that successful cases of AI utilization in Japan are in line with the Andrew Ng's Playbook, and that the Landing AI Playbook is an effective tool not only for Silicon Valley software companies, but also for manufacturing industry.

The Landing AI Playbook contains many descriptions on technical aspects, but there is almost no mention of how risk management should be developed to match AI proficiency. In order to further promote social implementation of AI technology, it will be important to organize and share best practices regarding the practical implementation of AI principles using the Action Targets given in the AI Governance Guidelines.

(3) Understanding your company's AI proficiency

Action Target 1-3: Companies that develop and operate AI systems should, under the leadership of top management, evaluate and re-evaluate in a timely manner their AI proficiency based on the extent of the company's experience in developing and operating AI systems, the number of employees, including engineers, involved in the development and operation of AI systems and their degree of experience, and the degree of AI literacy of these employees with respect to AI technology and ethics, except in situations where a company assesses negative impacts of their AI system are minor based on analyses of Action Targets 1-1 and 1-2 in light of the company's business domain and scale, etc. If the negative impacts are assessed to be minor and no evaluation of AI proficiency is carried out, companies should be prepared to explain their rationale to their stakeholders.

[Practical Example 1]

If a company succeeds in introducing an AI system into their business, or using the AI system to streamline their production processes and service provision operations, the AI system can have positive impacts on their business, for example, providing solutions to HR shortage issues, improving productivity, and helping the company to develop a new high-value-added business. On the other hand, haphazard provision of AI to businesses can entail risks peculiar to AI, such as unintentionally impairing fairness and causing safety problems. As such, AI companies are required to understand the risks, which can be described as the negative aspects of AI introduction, before they move forward with introduction. Therefore, **it becomes important to look at an index of AI proficiency (how much the company is ready for developing and operating AI systems), which makes it visible how much the company can respond to the negative impacts of AI**.

At our company, **under the leadership of top management, we evaluate and re-evaluate our AI proficiency in a timely manner** to prevent ourselves from focusing perhaps too much on the positive impacts of developing and operating AI systems at the cost of failing to pay enough attention to the negative impacts and risks, and consequently ending up causing significant damage to other businesses with our AI systems.

To evaluate AI proficiency, we use the AI-Readiness Guidelines from the Japan Economic Federation's "AI Utilization Strategy - For an AI-Ready Society" (February 19, 2019).[11] **With the AI-Readiness Guidelines we evaluate whether the magnitude of impacts that our company's AI systems may have on society and the scope of relevant stakeholders are proportionate to our company's level of AI proficiency**. And **we use the AI proficiency as a basis for examining our overall AI governance, including discussing our AI governance goals**.

[11] Japan Business Federation, "AI Utilization Strategy - For an AI-Ready Society" (February 19, 2019). https://www.keidanren.or.jp/policy/2019/013_honbun.pdf. A table in a list format is available at https://www.keidanren.or.jp/policy/2019/013_sanko.pdf. They are written in Japanese.

[Practical Example 2]

We do not develop AI systems ourselves. We outsource our AI systems development, operate the delivered AI systems, and provide them to AI system users. As such, since we are merely an AI system operator, we did not initially pay attention to our AI proficiency when we began providing AI systems to our users. However, we began to see a rise in the number of complaints from AI system users, and it soon became clear that the AI systems were not working as we had “expected.” We subsequently learned that not only did we have problems in terms of proper communication with our AI system developer, but problems with the literacy that is basis for our communication.

We understood that the “AI-Readiness Guidelines” from the Japan Economic Federation were primarily for developers of AI systems, **but we found that the guidelines also contained indices of AI proficiency for AI system operators who outsource their AI development**. So, we extracted some of the elements in the guidelines for our own purposes and incorporated them into our guidelines, and now evaluate and re-evaluation our AI proficiency in a timely manner under the leadership of top management. For example, **we use them to evaluate whether the magnitude of impacts that our company’s AI systems may have on society and the scope of relevant stakeholders are proportionate to our company’s level of AI proficiency**. And **we use AI proficiency as a basis for examining our overall AI governance, including discussing our AI governance goals**.

2. Goal setting

(1) Considering setting AI governance goals

Action Target 2-1: Companies that develop and operate AI systems should, under the leadership of top management, consider whether or not to set their own AI governance goals (for example an AI policy*) based on the “Human-Centric Social Principles of AI”, keeping in mind importance of the process leading up to goal setting, and taking into consideration the positive and negative impacts that AI systems can have, social acceptance regarding the development and operation of AI systems, and their own AI proficiency. And if a company decides not to set AI governance goals based on the assessment that their potential negative impacts are minor, they should be prepared to explain their rationale to their stakeholders. If it is determined that the “Human-Centric Social Principles of AI” function fully satisfactorily, this may be set as the goals instead of the company's own AI governance goals. Note that even if no goal is set, it is desirable to understand the importance of the “Human-Centric Social Principles of AI” and to implement measures associated with Action Targets 3 to 5 as needed.

* AI governance goals include not only an AI policy that consists only of actions for complying with the Social Principles of AI, but also include a data utilization policy that covers other elements while including actions for complying with the Social Principles of AI. AI governance goals may include a policy of enhancing a positive impact such as improvement of inclusion with AI. Obviously, it is up to each company whether or not to call this AI policy. Representative examples of AI governance goals can be found in various literature[12].

[Practical Example 1]

Since we have only recently begun developing and operating AI systems and our AI proficiency is not very high, **we plan to use only AI systems for applications that have minor negative impacts on society** for the time being. For this reason, **we have not set any AI governance goals**, but we will consider setting goals when we decide to expand our business to applications where potential negative impacts cannot be deemed to be minor. We record the details of our discussion, as a matter of course, to ensure that we are ready to explain our rationale for not setting AI governance goals to our stakeholders.

[Practical Example 2]

While we have only recently begun developing and operating AI systems, we considered setting AI governance goals because the potential negative impacts of some of the applications of our AI

[12] One example of such literature is Satoshi Funayama, “Responsibility and Ethics of AI (No. 2), Ethics Initiatives by Company (1)”, p. 75 NBL No. 1170 (2020.5.15), which lists examples from Japan, the United States, Germany, etc. It is available only in Japanese.

systems cannot be deemed to be minor. **Because the "Social Principles of Human-Centric AI" appropriately assemble all elements that we should focus on, including privacy protection, we decided to define the Principles as our goals for the time being**, and we work hard to respect the seven principles in the "Social Principles of Human-Centric AI." For example, in order to encourage operations personnel to respect the seven principles, we are moving forward with training programs including e-learning to share awareness in individual workplaces. We believe that we will need to define our own AI governance goals when we expand the scope of our AI system development and operations, so **we currently hold study group sessions on cases at other companies as our lead up to setting these goals**.

[Practical Example 3]

We are a company with a diverse business portfolio, and each business unit has a different approach to AI technology. In addition, it is not easy to agree on a single set of AI governance goals because we operate under an in-house company system, where each company or business unit is independently operated. For this reason, as things currently stand, we respect the "Social Principles of Human-Centric AI," and at the same time incorporate AI ethics and quality training into our company-wide AI training to raise understanding of AI ethics and quality. We have also set up an AI consultation desk to collect information on cases from business units. While our actions may not appear agile from the outside, we believe that **the value is in the process of reaching agreement on AI governance goals**. We also believe it would be possible to **examine the need for and elements of AI governance goals at each business unit that develop and operate AI systems** at a preliminary stage leading up to setting AI governance goals for the entire company.

[Practical Example 4]

We have extensive experience in developing and operating AI systems as well as in supporting companies that operate AI systems, and we also develop and operate AI systems for applications whose potential negative impacts are not considered to be minor. While we have not encountered any serious incidents to date from AI systems that we operate or provide to third parties, we understand that there are **numerous applications of AI systems that we provide, social acceptance of which has yet to be established**. For this reason, **we have defined and published AI governance goals to enhance our communication with our stakeholders including consumers**. **These efforts**, thanks to the fact that our stakeholders understand our policies, have been commended for **enabling basic approaches toward AI technology to be shared among stakeholders including customers and persons engaged in developing AI system, and for facilitating communication among them**.

3. System design (building an AI management system)

(1) Incorporating gap analysis between AI governance goals and current state to address the gap into the management system as essential process

Action Target 3-1: Companies that develop and operate AI systems should, under the leadership of top management, identify a gap between AI governance goals and current state in the AI systems that they are developing and operating, and if any negative impacts are found upon evaluating the impacts resulting from the gap, determine whether or not the negative impacts would be acceptable, taking into account their severity, scope, and frequency of occurrence. They should incorporate processes that prompt a reexamination of how the AI systems should be developed and operated in an appropriate stage such as during AI system design, development, before they are used, and after their usage begins, to address cases where the negative impacts are found not to be acceptable. Those in operations positions should make these processes concrete. In addition, those who are not directly involved in the development and operation of AI systems should be included in the gap analysis between AI governance goals and current state. It should be noted that it would not be appropriate to stop the development or provision of AI simply because a gap was found. As such, gap analysis is merely a step for evaluating negative impacts and simply serve as a starting point for improvement.

* In implementing this Action Target, please refer to not only the following practical examples but also Attachment 2 (Practical Examples for Gap Analysis between AI Governance Goals and Current State) as needed.

[Practical Example 1]

We are a small company, our chief technical officer and development personnel work closely with each other, and the number of projects we undertake is not so large. So, our **chief technology officer has a good understanding of all our projects**. **The technology officer defines perspectives for gap analysis from the “Social Principles of Human-Centric AI,” and instructs development personnel to identify gaps for each perspective, evaluate impacts caused by the gaps, and report the results to the technology officer as soon as practically possible in all AI system development projects**. **The technology officer** then **re-evaluates the impacts caused by the gaps based on the reports received from the development personnel at a meeting of development personnel, with participation from non-development personnel, and, if any negative impacts are found, examine whether or not they are reasonably acceptable. If it is found that they are not reasonably acceptable, the officer with the meeting members reexamine a better way of providing the AI system**.

With respect to implementing the above process, **we are working to standardize internal operations** in accordance with Action Target 3-1-1, **using standard gap analyses in our industry**

and Attachment 2 of the AI Governance Guidelines as references.

[Practical Example 2]

Our company, which has many business units, **has appointed an AI governance officer and set up an AI Ethics Review Committee headed by this officer**. The committee is **composed of individuals other than those engaged in development and operation projects of each AI systems, and their mission is to analyze gaps on a project basis between the current state and our company's AI policy** based on the "Social Principles of Human-Centric AI." Specifically, **the committee is to create lists for gap analysis based on our AI policy, use the lists to identify gaps** in AI system development and operation, and **evaluate the impacts caused by the gaps**. **If any negative impacts are found, the committee is to determine whether or not they are reasonably acceptable. If it is found that they are not reasonably acceptable, project personnel are advised to reexamine a better way of developing and providing the AI**. **We create the lists for gap analysis in accordance with Action Target 3-1-1, using standard gap analyses in our industry and Attachment 2 of the AI Governance Guidelines as references**. **We also select actual projects and have AI management personnel work alongside project personnel to refine the lists and make them standard process in the management system operation**. The AI Ethics Review Committee asks the project personnel to report the results of their reexamination, and if there is any concern about the appropriateness of the report, the AI governance officer will communicate his/her concern to the officer overseeing the project and make some necessary adjustment.

The negative impacts of AI systems vary greatly depending on their application, scope, and mode of use. And because it can be surmised that persons engaged in these projects are likely to be the most knowledgeable on the nature and degree of their impacts, **in cases where it is clear that the negative impacts are minor, other modes of gap analysis operation may be allowed, for example, having AI management personnel present at project meetings carry out simple gap analysis instead of conducting a uniform and strict gap analysis**. That being said, we have not yet accumulated sufficient in-house know-how at this point for gap and risk analysis, and for this reason, we conduct a uniform gap analysis by the AI Ethics Review Committee as a mandatory gate for all projects, and we will see how it goes.

[Practical Example 3]

There may be instances where the process of gap analysis may have to be carried out by plural companies. For example, where an AI system operator that provides services to other users outsources AI system development to an AI system developer instead of developing the system on their own, it may be reasonable in some instances for the AI system developer and operator to apportion the gap analysis processes between them. In such cases, it is **obviously important for the developer and operator to share the expected flow of processes from AI system development to operation, as well as the method and benchmark for gap analysis**. If an AI system operator makes little of the risks associated with providing AI system-based services, this will place AI system developers in a difficult position. Thus, arrangement such as the above is important.

As a company that may receive such consignment orders to develop AI systems, we enter an agreement with operators which prescribes that, unless there are circumstances attributable to our company, the operator providing the service is responsible for any accidents in the operation of the AI system. However, our company is still at risk of becoming involved in a dispute if such an accident occurs. For this reason, we cannot afford to be indifferent to how the AI systems we deliver are being operated. In fact, there was one instance where we became aware of operational risks towards the end of a project. We advised the operator that we had to redesign the project and we had no alternative but to bear a part of the redesign cost. Learning from this experience, **we referred to standard gap analyses in our industry and Attachment 2 of the AI Governance Guidelines to establish our gap analysis process, thoroughly understood gist of each analysis item, and set up a policy of sharing our gap analysis process with AI system operators who focus solely on providing services to others without developing systems themselves**. Utilizing our gap analysis process which covers all items of concern and, above all, carrying out the gap analysis at an early stage have allowed us to conduct our negotiations with our customers more smoothly.

* As in the following practical examples, there may be instances where companies will need to expand the scope of their discussions, in addition to conducting their usual gap analysis process.

[Practical Example 4]

We are a small company, and our main business is to develop AI systems. Our chief technology officer receives progress reports on all projects, and these reports contain sections on fairness and other matters associated with AI ethics. **Some of the ethical issues of AI can be addressed through technical approaches, such as preparing sufficient data sets to obtain reasonable output, but this may not be enough in socially sensitive areas in some cases**.

As such, **when working on AI system projects in these sensitive areas, we discuss the matter with our chief legal officer and others who are not involve in the projects themselves**. To identify sensitive areas, we refer to, among other sources, approaches of leading companies that have already developed and operated AI systems in a wide range of areas. Specialist journals are useful for collecting this type of information[13]. These journals often publish summary articles, and following the leads contained in these summary articles is an efficient and effective way of finding in-depth information on the Internet or through other source.

We are aware that **some companies invite external experts and specialists to exchange views on individual projects**. We would also like to set up a similar forum for exchanging views as our business expands.

[13] For example, Satoshi Funayama, “AI Ethics Initiatives by Companies (1)” NBL No. 1170 (May 15, 2020) provides examples of how sensitive areas should be addressed. It is available only in Japanese.

[Practical Example 5]

We are a large company with a mixture of departments that develop AI systems and those that operate them. We have already established an AI policy and evaluate deviations from the policy in all projects. If a project is in a field that we have previous experience in, it is good enough to have AI management personnel take action in the early stages of the project, **but for developing or using AI systems in sensitive fields where we have no previous experience, we make sure that individual consultations are made instead of going through our regular process.** And **when such consultation is requested, the matter is to be discussed at a cross-sectional meeting consisting of persons in charge at the developing and operating departments, and legal department, etc**. The same applies if AI management personnel identify such a project during regular gap analysis process.

We **regularly invite external experts and specialists so that we have early access to information on recent AI incidents and sensitive areas**. Therefore, for the time being, we are sufficiently capable of addressing these matters through cross-sectional discussions based on information and general advice obtained from these experts and specialists. Meanwhile, **as the applications of our AI systems are diversifying, we believe, going forward, that we may eventually need to seek the opinions of external experts on individual projects as well**.

① Ensuring consistency with industry-standard gap analysis processes

> Action Target 3-1-1: Companies that develop and operate AI systems should, under the leadership of top management, check whether a standard gap analysis process in their industry is available and incorporate it into their own process if such a process is available.

[Practical Example 1]

Because a diversity of perspectives, as well as the sharing of understanding with other companies is indispensable for the **practical implementation of AI principles, companies must refer to efforts of other companies and organizations, and not simply consider the matter themselves**. Based on this line of thinking, we instructed our AI management personnel to investigate outside efforts to build a gap analysis process.

Since our main business is to develop AI systems for industrial applications, we conducted a survey with a focus on industrial applications. As we proceeded with the survey, we found that, for example, the Ministry of Economy, Trade and Industry, the Ministry of Health, Labor and Welfare, and the Fire and Disaster Management Agency have published the "**Guidelines on Assessment of AI Reliability in the Field of Plant Safety**" and "**Format for Recording Details of Implementation**" for implementing the Guidelines, as well as "**Practical Examples of Reliability Assessment Records**," a compilation of examples on how to fill these out. **In addition, we learned that "Guidelines for Quality Assurance of Machine Learning-based Artificial Intelligence**" published by the Consortium of Quality Assurance for Artificial-Intelligence-based Products and Services contains examples of voice user Interface, industrial process, autonomous driving, and OCR. Furthermore, we learned that the National Institute of Advanced Industrial Science and Technology, a national research and development corporation, has published their "**Machine Learning Quality Management Guidelines**" **and has a plan to create a reference guide** in the form of a compilation of specific examples of actual applications grouped by industrial application. **The gap analysis process we currently operate incorporates some of these specific efforts**.

[Practical Example 2]

We develop and operate AI systems based on data obtained from AI system users. **In the practical implementation of AI principles, particularly as it relates to the practical implementation of the privacy principle**, we understand that **we need to pay attention to** not only AI model building and their output, but also **how we manage input data to the AI model**. While we have rich experience in managing personal information, we nonetheless believe that we should actively learn from outside. Accordingly, we instructed our AI management personnel to investigate outside efforts to build our gap analysis process.

As for things to be considered regarding AI model building and their output, one example we found was the "**Checklist of Voluntary Actions**" **for profiling presented by the Personal Data +α Study Group**. Furthermore, regarding input data management to AI models, we revisited the **guidelines published by the Personal Information Protection Commission**. The "**Guidebook on Corporate Governance for Privacy in Digital Transformation (DX) ver. 1.1**" was also informative as it includes descriptions regarding AI from the perspectives of both input and output. **The gap analysis process we currently operate incorporates some of these specific efforts.**

② Providing users with sufficient information on potential gaps and measures to address them

Action Target 3-1-2: In case that a certain degree of gaps may potentially occur with AI systems of AI system operators that provide services to AI system users, they should, under the leadership of top management, provide sufficient information about the gaps and measures to address the gaps, as well as make a contact point easily accessible.

[Practical Example 1]

We operate AI systems and provide services to AI system users, primarily to large numbers of unspecified consumers. Because **AI literacy at service recipients is expected to range over a wide spectrum**, **we compile and provide risk related information** ― information that we implement proper risk management, take measures to minimize negative impacts, and implement rigorous information security management ― **on our AI systems operations in a way that inexperienced consumers are able to understand, and also make a contact point easily accessible**. In addition to this information, because AI literacy at service recipients is expected to range over a wide spectrum as mentioned above, **except in situations where it is obvious to users that AI system outputs are used** in the information and other services we provide, **we indicate in an easy-to-understand way that AI is used and clarify the advantages and disadvantages of using AI**. We also let users know that **we offer alternative services for AI system users who prefer not to use the AI mode**. Since we also manage personal information in some instances, we not only comply with the Personal Information Protection Commission guidelines, but **have also established ongoing communication with consumers, using the "Guidebook on Corporate Governance for Privacy in Digital Transformation (DX) ver 1.1" as a reference**.

[Practical Example 2]

We operate AI systems and provide services to outside parties in a similar manner described in Practical Example 1, but our services differ from the Example in that they are **provided to companies for business purposes**. **Since AI literacy is relatively high at our service recipients, we provide succinct explanations with technical terms** about the possibility that certain gaps can be found in the AI system that we provide and measures we take to address the gaps. We also make a contact point easily accessible.

We may provide AI system-based services to general consumers in the future. If we do, we think we will **provide sufficient information in accordance with the AI literacy of the recipients of our services**.

[Practical Example 3]

Our operations are similar to those described in Practical Example 1, but **we make it clear in our agreement with our AI developer that we can have the developer provide us with the information that we will need to respond to inquiries from AI system users**. The developer acts promptly on these matters, partly because "feedback" from AI system users is valuable information for the developer, too.

[Practical Example 4]

While our operations are similar to those described in Practical Example 1, **we also believe that giving AI system users the option to choose AI system-based services at their own discretion provides additional value, so we are designing ways of providing information that will differentiate us from competitors**. In addition, we are working on ways **of receiving feedback not only on the AI systems but also on how we should provide information**.

③ Providing AI system developers with sufficient information for gap analysis (for data providers)

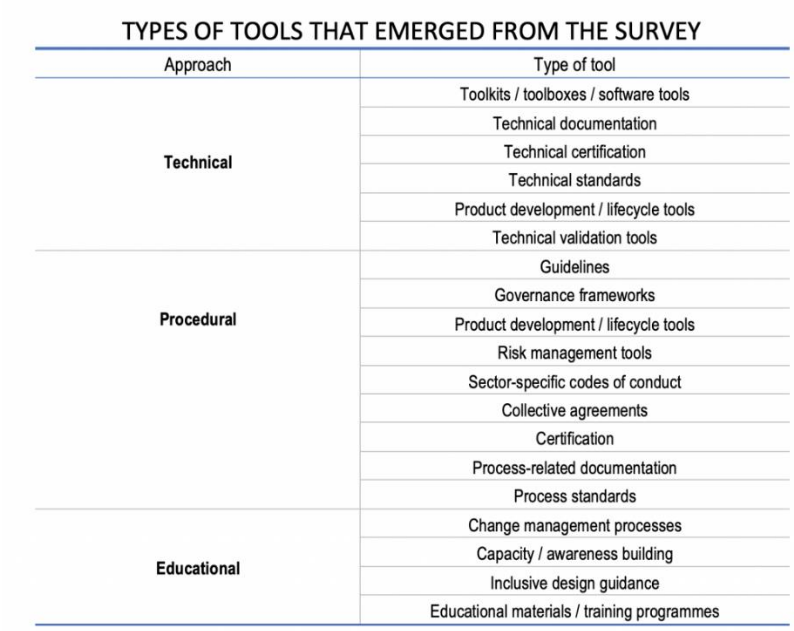
Action Target 3-1-3: Companies that provide data should provide information on the data sets including data collection sources, collection policies, collection criteria, annotation criteria, and limitation on use to ensure that companies that develop AI systems are able to appropriately conduct gap analysis, and AI system developers should acquire data sets from data providers that provide sufficient information.

[Practical Example 1]

We are a data provider that provides data to companies that develop AI systems, and we **provide information on data sets**, including data collection sources, collection policies, collection criteria, annotation criteria, limitation on use and other information **to ensure that companies that develop AI systems are able to appropriately conduct gap analysis**. Moreover, **even if a data set that we provide is not sufficiently organized, we provide sufficient basic information, including the data collection source, that will be needed for performing gap analysis**.

Column: Initiatives for ensuring fairness

It is difficult to define standards of fairness and to practically apply fairness principles. To address this issue, the OECD is working to support practicing AI principles. The OECD Network of Experts Working Group on Implementing Trustworthy AI has put together a collection of initiatives for putting principles such as fairness into practice, and has classified these into three categories: technical, procedural, and educational. As for technical tools for ensuring fairness, the working group reports efforts carried out at AT&T, Microsoft, LinkedIn, Google, and IBM, but none at Japanese companies. The sample tool featured in the OECD Framework of Tools for Trustworthy AI is LinkedIn's Fairness Toolkit (LiFT).



TYPES OF TOOLS THAT EMERGED FROM THE SURVEY

Approach	Type of tool
Technical	Toolkits / toolboxes / software tools
	Technical documentation
	Technical certification
	Technical standards
	Product development / lifecycle tools
	Technical validation tools
Procedural	Guidelines
	Governance frameworks
	Product development / lifecycle tools
	Risk management tools
	Sector-specific codes of conduct
	Collective agreements
	Certification
	Process-related documentation
	Process standards
Educational	Change management processes
	Capacity / awareness building
	Inclusive design guidance
	Educational materials / training programmes

Classification-of-tools chart, excerpted from AI Wonk at OECD.AI.[14]

Standards of fairness and common understanding on practice regarding fairness may be defined de facto in international discussions like one at the OECD. Because standards of fairness may differ due to cultural differences and differences in business customs, we will need to disseminate efforts of Japanese companies to ensure that the diversities found in Japan are also well recognized in the discussion. In fact, financial loan screening by AI in three different regions — Japan, US, and UK — was found to be different depending on what constituted being fair in each region[15].

[14] Carolyn Nguyen, Adam Murray, and Barry O'Brien, "What are the tools for implementing trustworthy AI? A comparative framework and database," The AI Wonk, OECD.AI (May 25, 2021), https://oecd.ai/wonk/tools-for-trustworthy-ai.

[15] Fujitsu, "Developing a 'Fairness by Design' AI development method that considers fairness — which differs depending on culture and business customs — in the design phase" (March 31, 2021), https://pr.fujitsu.com/jp/news/2021/ 03 / 31-1.html. Although the website is available only in Japanese, the research interview related to the website is available in English at https://www.fujitsu.com/global/about/research/article/202105-ai-ethics.html.

(2) Improving literacy of human resources responsible for AI management systems

Action Target 3-2: Companies that develop and operate AI systems should, under the leadership of top management, strategically improve AI literacy in order to properly operate their AI management system, considering outside learning materials as an option. For example, this may include education to boost general literacy on AI ethics for those in the top management, management teams, and those in operations positions responsible for legal and ethical aspects of AI system development and operation, as well as training for those responsible for AI system development and operation not only on AI ethics but also on AI technology. Companies that provide data should take steps to improve the general literacy of AI ethics of employees engaged in data provision by referring to practical examples for AI system developers and operators.

[Practical Example 1]

Because we are a small company and **we do not have many trainees for AI literacy, we do not develop training programs in-house for boosting AI literacy and have instead decided to make use of third-party learning materials.** A variety of learning programs are available domestically and overseas, including online courses and textbooks provided by Coursera, an American commercial organization that provides educational technology, and by the Japan Deep Learning Association (JDLA), as well as the Stay-at-Home DX Step Course introduced by the Ministry of Economy, Trade and Industry. One of the examples of online learning materials for people without a science background is "AI for Everyone" (Coursera, about 4 hours, with transcript) lectured by Professor Andrew Ng of Stanford University. It is pointed out that "learners can reach near-expert levels by studying for about 40 hours on a learning site on the Internet, such as Coursera[16]," and based on this, we decided that **third-party learning materials would be sufficient**.

We utilize a program based on the syllabus of the JDLA certification test to measure trainees' achievement levels. The JDLA's G-test covers a wide range of topics from the basics of AI technology to AI ethics. In addition, at the "Online Seminar on the Experience of G-Test Passers" organized by the JDLA held on May 30, 2020, an individual who had two years of experience in sales at an AI venture company stated based on her experience that she needed 25 to 30 hours of study to pass the test, which was one of the bases for our confirmation that this test would not be an excessive burden on trainees.

Reviewing our training program, we believe that it is as effective as we expected. For example, an individual, whose knowledge of AI systems was limited to fragmentary information that she had heard in news reports, **is now able to consider the negative impacts of AI, taking these issues as your own matter**, because she learned the basics of AI technology and its ethical aspects.

[16] Yutaka Matsuo, "Leading Researcher's Thoughts on 'How to Face AI'", p. 56, Weekly Toyo Keizai (May 16, 2020), available only in Japanese.

[Practical Example 2]

We are a large company, and AI system development and operation is one of the pillars of our business. We are aware that third-party learning materials on AI technology and ethics are available, but since **we provide a large number of AI systems and they can have a significant impact on society, we use in-house learning materials with extensive coverage of case examples based on specific applications of our own AI systems** instead of general-purpose third-party training materials.

In the AI training program that we initially created, the AI ethics component came at the end of the AI technology course. However, **recommendations given by a committee made up of invited external experts raised our management's interest in AI ethics, and we have since created an independent e-learning course solely on AI ethics which all employees are now required to take**. This e-learning course includes a lecture and check test and is designed so that even people who are not familiar with AI ethics are able to complete it in about an hour. **Regardless of the short amount of time it takes to complete,** we believe **the learning effects of this course are considerable, because it is related to the applications of our own AI systems**.

Column: Examples of utilizing the JDLA certification test

Many companies use the JDLA's G-test[17]. Some companies have incorporated the G-test into their human resources development curriculum on company-wide DX promotion projects, with hundreds of people taking group exams. For example, ENEOS divides digital literacy for all their employees into four subjects: AI Analytics, Business Intelligence, Cyber Security, and Design Thinking, and incorporates the G-test into AI Analytics[18]. Another company says that their activities evolved from voluntary activities at the IT department to G-test preparation study group sessions with participation from about 100 people across departments. The G-test is also used in training programs provided by public institutions[19].

G-test users speak highly of the test, saying they obtained "knowledge that future project managers and solution architects must have," and some say that segments who seek to take the test now extend beyond young engineers to executives, managers, and team leaders. It appears the G test gives impetus for people to learn the basics of AI technology and AI ethics.

[17] Comments from group test companies at the website of JDLA, available only in Japanese at https://www.jdla.org/certificate/general/#general_No04.

[18] In slide 22 below, the E-test and G-test are elements of AI Analytics. https://www.hd.eneos.co.jp/csr/meeting/pdf/esg_ex_20201202.pdf. It is available only in Japanese

[19] The Saitama Industrial Promotion Public Corporation incorporates the G-test in their "AI / IoT Human Resources Development Training (engineer training course)" for SMEs. https://www.saitama-j.or.jp/iot/jinzai/. It is available only in Japanese.

(3) Reinforcing AI management through cooperation between companies/departments such as by proper information sharing

Action Target 3-3: Under the leadership of top management and with due consideration for trade secrets, companies that develop and operate AI systems and those that provide data should, except where everything from the preparation of data sets for training and other purposes to AI system development and operation is performed entirely within their own department, clarify and actively share, in accordance with the Principle of Fair Competition, AI system operational issues that the company or department is unable to fully address on their own and the information necessary to address these issues. In doing so, in order to facilitate the exchange of information clarified above, the AI system developer, AI system operator, and data provider are encouraged to agree on scope of information disclosure in advance and consider measures to protect trade secrets, for example, by entering a non-disclosure agreement.

[Practical Example 1]

We deliver developed AI systems to customers and these customers operate the AI systems. The accuracy of these AI systems may become lower due to changes of their operating environment, which in some cases may cause damage including damage to equipment. For this reason, **we ask our customers to monitor the AI system's output and we also provide instructions on how to determine if there has been any deterioration in quality**.

This arrangement does not work by simply asking customers who are not familiar with AI to monitor or perform other tasks on the system. **We need to take our time to explain and get customers to understand reasons why the AI systems need maintenance, causes that make maintenance necessary (deviation of the distribution of input data during operation from that of training data and, etc.), and trends of changes of output that result from the causes**. It may be sufficient in some cases to simply provide standard information, but even if AI system developers think it should be sufficient, they should encourage their customers to actively ask questions and make efforts to share understanding between the parties as much as possible. It is also important to enter a maintenance service contract, etc., as necessary, and to create a system in which the developer willingly answers questions even after system delivery. Also, **after the AI system is re-trained, the developer should give a detailed explanation on how its outputs have changed after re-training to the customers**.

In order to facilitate such information sharing, our standard practice requires an agreement in advance on the scope of information to be disclosed between the AI system developer and AI system operator as well as non-disclosure agreement.

[Practical Example 2]

The AI systems we develop were trained by data sets of particular cases and may produce unfavorable output if they are applied to subjects that are not covered by the data sets. For this reason, **in addition to providing explanations on the data used for training, etc., as well as overviews of the models used and explanations on accuracy and other performance parameters, we also explain, to AI system operators who intend to provide the AI system to AI system users, the situations and targets that the AI system should not apply**. In order to make the information sharing process more thorough, we communicate not only in paper or electronic form, but also set up a meeting to explain matters verbally and ask the operators to sign off on a form acknowledging their receipt of such explanation.

① Understanding the current state of information sharing between plural companies

Action Target 3-3-1: Under the leadership of top management and with due consideration for trade secrets, companies that develop and operate AI systems and those that provide data should, except where everything from the preparation of data sets for training and other purposes to AI system development and operation is performed entirely within their own company, understand the current state of information sharing between companies, and update their understanding in a timely manner.

[Practical Example 1]

When developing AI systems, AI system developers need to consider situations that the systems are intended to be used in, and AI systems need to be operated with proper knowledge about conditions under which they were developed. For this reason, in situations where provision of data, data annotation, AI system development, and AI system operation are carried out by plural companies, it becomes important to share information among the companies. **It is desirable to make such information standardized for the sharing purpose to promote social implementation of AI technology**. With this issue in mind, **under the leadership of top management and with due consideration for trade secrets, we take steps to understand the current state of information sharing among plural companies and update our understanding in a timely manner in deciding our approach to information sharing**.

As we proceeded with information gathering, we learned that **various efforts have been made to standardize information sharing among plural companies**. For example, we learned that the National Institute of Advanced Industrial Science, a national research and development corporation, has published their “Machine Learning Quality Management Guidelines” with the aim of setting up socially agreed standards on quality of machine-learning-based systems, which is one of the purposes of the Machine Learning Quality Management Guideline. We also learned that the Ministry of Economy, Trade and Industry, Ministry of Health, Labor and Welfare, and Fire and Disaster

Management Agency have created, based on these Guidelines, a format for recording details of the implementation of reliability assessments in the area of plant security. In addition, we also learned that model cards are proposed with the belief that AI models should have labelling about their performance just as ingredient labelling on food products help consumers make responsible choices[20].

At the moment, although there is no standard documenting procedure for sharing, performance and quality information of trained machine learning models, etc. between plural companies, **we intend to refer to various other efforts instead of considering our own standards from scratch in establishing our in-house system**.

[Practical Example 2]

We **belong to an organization engaged in AI ethics and quality, and actively exchange views with other affiliated companies on best practices for providing information on AI system's performance and other attributes**. AI system users should be provided with sufficient information about AI systems, but because not all users, whether they are consumers or other users, are knowledgeable about details of characteristics and limitations of AI, it is not appropriate to think that it is enough merely to give AI system users information that is difficult for anyone other than experts to understand or huge amounts of detailed information unilaterally. In order to consider appropriate ways of providing information, it is important for companies to not only have direct user experience of their own, but also indirect contact with large numbers of users through exchanges of views with other companies.

Information that AI system developers should give to AI system operators includes, for example, information about the data used to develop their AI systems. For example, this may include information on the source of data (which may sometimes be open data), the amount of data and their distribution, and overviews for each of the categories covered in the data. It is also important to describe the algorithms chosen (or not chosen) during development, provide an overview of the generated model, and in particular, to explain the conditions under which tests were performed and how much accuracy was attained.

While these perspectives are not new to companies with rich experience in developing and operating AI systems, we believe that "**how we communicate" is important;** that is, **types of information we should provide and how in-depth our explanations should be.** **Understanding the current state of information sharing among plural companies is important in considering the overall design of AI governance**, and herein lies the meaningfulness of participating in organizations that focus on AI ethics and quality.

[20] Google, "The value of a shared understanding of AI models," https://modelcards.withgoogle.com/about.

② Encouraging gathering information and exchanging views routinely for conditions and risks analysis

Action Target 3-3-2: Companies that develop and operate AI systems should, under the leadership of top management, regularly collect relevant information such as formulation of rules for the development and operation of AI systems, best practice, and incidents, and encourage the exchange of views within and outside the company.

[Practical Example 1]

Agile governance is sustained by routinely gathering information and exchanging views for such information. As it is said that governance for ensuring proper AI system development and operation requires involvement of various stakeholders, we need to listen to voices of various stakeholders in gathering information and exchanging views routinely. For example, **even if there is an in-house AI management team, it is necessary to hold in-house study group meetings with different departments and to be involved in group activities that other companies also participate in**.

To date, when a person in charge went out to gather information and exchange views outside the department/company, we asked the person to explain how much likely he/she would obtain important information. Although we are reaching a consensus about principles of AI ethics, **because we have little choice but to explore how to respect the principles in a situation where there are no correct answers and because other companies also take a similar approach, we now encourage AI management personnel to gather information and exchange views regarding the proper development and operation of AI, and instruct them to share the information they obtained at in-house study group meetings across the departments**.

Continuing such activities gradually enables us to get a picture of major trends, although we have no definitive solutions. And we utilize what we got in these activities in the analysis of conditions and risks we carry out in a timely manner.

[Practical Example 2]

We are a small company that develops AI systems. Because some in the company were of the view that we should focus on growth more than respecting AI ethics, **we decided to start off by organizing joint in-house study group meetings on AI ethics at our legal and technical departments**. We assigned a facilitator for these meetings because the definitions and usages of terminology may differ from department to department. As a result, discussions proceeded smoothly, and we learned that there were no significant differences in the understanding of AI ethics as engineers who had been of the view that we should focus on growth were by then familiar with papers dealing with matters such as fairness. **Ever since our engineers began to show an interest in respecting AI ethics through technological means, our development processes have been gradually becoming consistent with AI ethics**. **In the next step, we would like to progress further by exchanging views outside**

the company.

Column: An environment of co-creation created by diverse stakeholders

While Action Target 3-3 is compiled from the perspective of reinforcing AI management as an approach to address the negative impacts of AI systems, from another perspective, information sharing among plural companies can be regarded as a move to create the foundations of an environment of co-creation. Even in the development and operation of conventional software and systems, the cooperation of plural companies has been indispensable for ensuring safety and addressing other issues, but further cooperation is required for AI systems.

In the context of this situation, various efforts have been made for further collaboration. Regarding contracts and legal liability, the JDLA has held study group meetings on “AI Quality Assurance in Contracts.” The AI Law Study Group, which is a group of volunteers, is divided into subcommittees on data, privacy, intellectual property, etc., to exchange opinions on legal and ethical issues. In the area of standardizing shared information among multiple companies, the “Partnership on AI” is running a project called “ABOUT ML,” which provides samples from Google, Microsoft, and IBM.

In Agile Governance, which is required in Society 5.0, a diverse range of stakeholders are expected to be actively involved in designing governance. And we hold out expectations that by organically combining these activities, we will be able to create the environment of co-creation needed for AI system development and operation. AI Governance Guidelines are expected to make a positive contribution in making organic linkages.

(4) Reducing incident-related burdens on users by preventing incidents and through early response

Action Target 3-4: Companies that develop and operate AI systems and those that provide data should, under the leadership of top management, reduce incident-related burdens on users by preventing incidents and through early response.

[Practical Example 1]

In order to improve users' confidence in AI systems, we believe it is important for us to reduce the burden on users as much as possible by devising ways to prevent problems from occurring, and by responding promptly when problems do occur. **AI system development and operation often involves companies and individuals in a variety of different positions**, such as data providers, AI system developers, AI system operators, business users of AI systems, and consumers. And **because of the so-called black-box aspect of AI, it tends to be difficult to find out who is responsible**. In order to prevent incidents, **it is important to allocate responsibilities to those who are able to mitigate negative impacts**. **It is also important to improve our ability to respond early to incidents by preparing for possible incidents**.

[Practical Example 2]

With the implementation of Practical Example 1 forming the basis of our operations, we are looking into using insurance for some purposes. **In applications where certain uncertainties are unavoidable regarding AI system operation and where an incident rarely occurs but if it occurs it may come with certain economic losses despite great benefits the AI system brings to society as a whole,** we believe that **it is important to reduce the burden on users with insurance to provide remedies for damage as soon as possible**. We certainly recognize the importance of reducing the uncertainties of AI systems to continuously improve user confidence, and we continue research and development to achieve this objective.

① Allocating burdens of addressing uncertainties among companies appropriately

Action Target 3-4-1: Under the leadership of top management, companies that develop and operate AI systems and those that provide data should, except where everything from the preparation of data sets for training and other purposes to AI system development and operation is performed entirely within their own company, allocate burdens of addressing uncertainties of AI systems among companies so that negative impacts may be minimized as a whole.

[Practical Example 1]

The use of AI systems that are inductively built with training data and machine learning techniques is accompanied by the problem of uncertainty wherein inferred results derived by AI systems may not always be correct. Measures to address such uncertainties include approaches aimed at reducing uncertainties in the AI system development stage, including preparing appropriate data sets, selecting appropriate models, and conducting verification and testing before the AI system is put into operation. And equally important are approaches that aim to control uncertainties through measures during the operation of AI systems, including the monitoring of AI system operations. **In cases where an AI system is developed and operated by different parties, in principle, the parties should appropriately allocate burdens of addressing uncertainties associated with the provision of AI in contract and etc., taking into account options that are realistically available to each party**. This principle is confirmed in the "Contract Guidelines on Utilization of AI and Data[21]."

Based on the principle, as a company that develops AI systems, we believe that **ensuring that AI system operators use the system properly contributes to increasing society's confidence in AI technology**. In our information gathering, we have found that some AI system operators see AI systems as nothing more than conventional software and that AI system developers should therefore take full responsibility for the quality of AI systems. In contrast, we have also found that, **in some cases, once AI system operators understand their expectations for the AI system and are given detailed explanations so that they themselves are clear on the matter, they are able to determine the re-training timing on their own**. And we came to understand that the idea that it is important for "quality assurance engineers, teams, and organizations, along with development and sales" to "engag[e] in activities that enhance their understanding of AI products" as described in the "Guidelines for Quality Assurance of Machine Learning-based Artificial Intelligence" is gradually spreading. However, **since the idea that AI system developers should guarantee quality is still deep-rooted**, we intend to continue to conduct regular surveys on burdens of addressing uncertainties while holding out expectations for a spread of the positive impacts of activities such as the "Guidelines for Quality Assurance of Machine Learning-based Artificial Intelligence."

[Practical Example 2]

We are an AI system operator and provide services, where we use AI systems developed by another company. We entered an agreement with an AI system developer based on the model contract in the "Contract Guidelines on Utilization of AI and Data." According to the Contract Guidelines, a developer of an AI system (trained model) guarantees neither perfect completion of the work nor the performance, quality, etc. of deliverables, but owes a certain level of duty of care in performing the

[21] Ministry of Economy, Trade and Industry "Contract Guidelines for the Use of AI and Data: AI Section" (June 2018), page 106. For your convenience, see page 103 of the English translation available at https://www.meti.go.jp/press/2019/04/20190404001/20190404001-2.pdf. "[T]he User's issues will be resolved through the decision-making of the User because that decision-making is deeply related to the User's business and the existing internal rules, restrictions, and organizational units and … it is difficult for the Vendor to make a performance warranty with respect to the behavior of a Trained Model, etc. in response to unknown input (data) that is not under the control of the Vendor."

work. We assumed that we were merely operating an AI system developed by another company and had not seriously considered our accountability as the operator in a case where any inappropriate incidents associated with the AI system operation occurred or where end users asked us for explanations.

However, **no matter who is held legally liable ultimately,** since we came to recognize that **if a user asked us for explanations about AI systems that we operate, we would not be exempt from all of the responsibilities to respond to these requests, at least as the first responder** and that **we might place our reputation at risk if we failed to provide sufficient explanations** **only because we were the one who AI system users directly receive services from**, **we have shifted our policy to one where we do whatever is in an AI system operator's capacity to reduce risk and explain these efforts as necessary** with help from AI developers.

[Practical Example 3]

We have plans to outsource the development of AI systems that utilize data that we possess. But since we lack know-how regarding data management, we initially intended to outsource the entire project to a third party, including not only data preprocessing such as cleansing, but also data quality management. We mistakenly assumed that by simply collecting data that we currently possess and providing the AI system developer with the data, the AI system developer, who was an expert about data management, would take care of the data processing necessary to develop the AI system that met our requirements.

However, once we began gathering information before outsourcing the development, we learned of the "Practical Guidebook on Data Provision for Fostering Human Resources of Experts in AI and Data Science"[22] which is a compilation of information and is a useful reference even for the provision of data between companies in general. This Guidebook outlines ideas such as "**the quality of data to be outsourced before the provision can only be controlled by the outsourcer**," and reminders for data providers which say that in some cases and under certain conditions, "**the outsourcer, in principle, should be held liable for damages caused by the use or implementation of the created deliverables** … **based on the ideas that a risk provider should be held liable and a party who reaps a profit should be held liable** … because profits from the use of the deliverables are allocated solely to the outsourcer."

We now understand that what kind of data is required for AI system development is determined according to what kind of AI system we seek to have developed, and that there are limitations to what the AI system developer is able to deal with. We reconsider how to allocate burdens of addressing uncertainties among companies, taking into account that **even the data provision is one of the important stages in the lifecycle of AI system development and operation**.

[22] Ministry of Economy, Trade and Industry "Practical Guidebook on Data Provision for Fostering Human Resources of Experts in AI and Data Science" (March 1, 2021), available only in Japanese.

② Considering in advance actions to take in response to incidents/disputes

Action Target 3-4-2: Companies that develop and operate AI systems should, under the leadership of top management, consider defining response guidelines and plans so that upon occurrence of an AI incident or dispute, they can promptly give an explanation to AI system users, identify the extent of the impact and damage, clarify legal responsibilities, consider relief measures and measures to prevent the spread of damage and recurrence, or take other relevant actions. Further, they should consider conducting rehearsal exercise relevant to such guidelines and plans, as appropriate.

[Practical Example 1]

We are a smaller company that develops and operates AI systems. It is obviously important to reduce risks of AI incidents as much as possible, but because it is difficult to completely eliminate the possibility of incidents, it is our understanding that **it is important to formulate and implement plans designed to minimize the damage from incidents when they occur**.

Specifically, **as part of our readiness for incidents when they occur, we have set up a contact point from outside, assigned an officer in charge of response, and established a contact system within the company, as well as a contact system for external parties and experts**. While it is difficult to cover all potential incidents, we have formulated general response guidelines by classifying major potential incidents that we can anticipate based on features of our AI systems into some categories. In addition, we carry out a rehearsal exercise regularly to verify the feasibility of the response guidelines we have formulated.

[Practical Example 2]

We are a large company that operates AI systems. **In order to promptly respond to incidents, we have set up a contact point from outside, assigned an officer in charge of response, set up a communication and coordination system for the risk management, legal, public relations, and crisis management divisions, as well as a contact system for external parties and experts**.

In addition, we sketch out multiple potential types of incidents, consult with experts in advance to identify the kinds of legal liabilities we may incur, and then carry out risk assessment. Because various types of damage such as personal, privacy infringement, and property damage can occur, it is useful to sort out in advance how the legal liabilities of developers, operators, and users, etc., relate to each other for each type of damage. In addition, what we need to keep in mind is that there are a variety of factors (problems in algorithm, training data authenticity, biases in the training data, etc.) that can potentially cause abnormal outputs and that unexpected things are prone to occurring, which is unique to AI system. We strive to regularly update our technical and operational mechanisms that are designed to reduce system risks even in the event of unexpected situations.

We have a company-wide BCP (Business Continuity Plan) in place. Because we may encounter business continuity issues if any of the AI systems we run shuts down, we decided to include AI incidents as one of triggers of BCP, and we are currently formulating plans on initial response and business continuity to address situations where we are forced to shut down all or some of our AI systems. In addition, we understand that simply formulating plans is meaningless and that any inability to actually execute the plans during an emergency poses a major risk. So, we carry out a rehearsal exercise for executing our plans at least once a year.

4. Implementation

(1) Ensuring readiness for explanation about implementation status of AI management systems

Action Target 4-1: Companies that develop and operate AI systems should, under the leadership of top management, make sure that they are ready for explanation about the implementation status of AI management systems externally by recording the gap analysis process under Action Target 3-1 and by taking other relevant actions.

[Practical Example 1]

Of the dynamic processes, which include conditions and risks analysis, goal setting, system design, implementation, evaluation, and re-analysis of conditions and risks, it has been pointed out that the “implementation” process can lapse into a “passive” exercise and can actually be difficult to put into practice. Stated simply, the task of “implementation” is to keep records and publish the status of implementation as necessary as shown in Action Target 4-3. Because **the acquisition of data and information in the “implementation” process leads to decision-making for improvement**, we believe that “implementation” is the key to achieving improvements by means of re-analysis of conditions and risks, evaluation, and other means.

We place great importance not only on the practical application of AI governance, but on keeping records for further improvement, and we take it as a matter of course that we maintain the records of implementation status of Action Targets 3. For example, we keep records of gap analysis carried out in all individual AI system development projects, create an implementation outline when AI-related training is provided, maintain minutes of internal meetings and meetings with other companies regarding AI system development and operation, and make sure that relevant persons other than those engaged in these tasks are able to access the records and documents mentioned above.

Because we are a relatively large company, we do not have difficulties with addressing action targets associated with general corporate governance. However, **because functions are considerably subdivided and allocated within our company, we are concerned that gaps may eventually emerge between different departments in the degree of expertise and in understanding of AI, and that this may have an effect on inter-departmental coordination and cooperation, since AI is a relatively new technology**. For example, a concern we have with the contact point we have set up in accordance with Action Target 3-1-2 is the possibility of **being late in identifying signs of an impending serious incident due to the inability of an inquiry service rep to comprehend technical details**. We currently make efforts to improve employee literacy in accordance with Action Target 3-2. For the time being, however, **we have the inquiry service rep actively report the details of incoming inquiries, in addition to overviews, to AI management personnel.**

As for “recording the gap analysis process under Action Target 3-1 and by taking other relevant actions,” for the purpose of explanation to other departments and outside the company, we refer to a proposal by U.S National Institute of Standards and Technology (NIST)[23] and endeavor to make the record and other relevant actions accurate and meaningful as far as possible and be aware of limit of explainability.

[Practical Example 2]

We are a small company that develops AI systems. Our chief technical officer is aware of all projects, is very familiar with AI ― as demonstrated in his/her ability to write programs and analyze research papers ― and has a strong interest in AI ethics issues. Therefore, we do not believe that the gaps in expertise between departments will be a problem. Meanwhile, **because of the high level of expertise of the people involved in the project, it is easy for them to assume that action targets are achieved without them having to check the details.** For this reason, **we take the step of attaching a gap analysis checklist to project progress reports so that the chief technical officer can interview those involved as needed**.

In addition, because we are a highly specialized group, we analyze that there is a tendency for our perceptions to diverge from those of the public. For this reason, we endeavor to direct our awareness to social acceptance by **regularly sharing the current status based on daily information gathering and exchange of opinions** in accordance with Action Target 3-3-2 while checking our implementation status.

[23] P J. Phillips, Amanda C. Hahn, Peter C. Fontana, David A. Broniatowski, Mark A. Przybocki, “Four Principles of Explainable Artificial Intelligence (Draft), NIST Interagency/Internal Report (NISTIR) - 8312-draft” (August 19, 2020)

(2) Ensuring readiness for explanation about operating status of individual AI systems

Action Target 4-2: Companies that develop and operate AI systems should, under the leadership of top management, monitor and record the status of preliminary and full-scale operations so that gap analysis for individual AI systems in preliminary and full-scale operations can be continuously implemented. Companies that develop AI systems should assist the monitoring conducted by companies that operate AI systems.

[Practical Example 1]

We are a company that operates AI systems and provides these systems to AI system users. We outsourced our AI system development to an AI system developer and have received explanations from development personnel in areas ranging from data set content to the verification of AI model behavior to ensure that we are able to address not only the matter of accuracy but also fairness. These development personnel have informed us that, **in order to ensure accuracy and fairness, the AI system will need to receive maintenance if differences begin to appear between the user profile assumed at the time of development and the actual user profile**.

Because we have no employees with enough knowledge to understand AI system codes, **we asked the developer to provide a way of automatically logging the inputs and outputs that significantly impact performance, and teach us how to perform the monitoring**. Subsequently, as part of Action Target 3-1 **we defined a management system for maintaining performance, using the sample gap analysis list included in Attachment 2 as reference**. Currently, we use the management system to perform continuous monitoring and maintain records.

[Practical Example 2]

We are a company that develops AI systems for other companies to operate. We do not legally own these AI systems, but as provided in our maintenance contract, we have certain responsibilities for operations performed by other parties, which means there is an operator aspect to what we do. Given this situation, the cooperation with the company who operate our AI systems (actual operator) on a daily basis is indispensable for performing the monitoring for the purpose of maintaining these AI systems' performance. In fact, **our arrangement is that the actual operator is to record the AI system's output, determine any significant deterioration of quality based on the output, check the actual situations, and notify us about any need for re-training**. **The actual operator is also to be present in subsequent meetings where we discuss the need for re-training**.

The reason why the actual operator is able to determine the timing of re-training is that they themselves have a good understanding of what they specifically want from the AI system and what they are specifically able to do. **It is important that AI system developers understand the expectations that AI system operators have for AI systems and provide detailed explanations on what can be done so that the operators themselves are clear on the matter**. As described in the “Guidelines of Quality Assurance of Machine Learning-based Artificial Intelligence,” it is important for “quality assurance engineers, teams, and organizations, along with development and sales” to ”engag[e] in activities that enhance their understanding of AI products.”[24]

Column: Automated monitoring by software

“GOVERNANCE INNOVATION Ver. 2: A Guide to Designing and Implementing Agile Governance” discusses the ideal shape of governance in Society 5.0. With respect to the operation of governance systems, Governance Innovation Ver. 2 points out that because we are now able to obtain data in real time ー data that we could only gather in fragments in the past ー not only can we now perform monitoring more efficiently and in more depth with the use of such real-time data, but have more flexibility in choosing the approaches for achieving our goals by gauging risk situations and the degree of our goal achievement at any given time, and that this may enable us to continuously innovate while maintaining our compliance.

With respect to monitoring AI systems in the full-scale operation, the efforts by GRID INC. are interesting as showcased in their CEATEC exhibit in 2020. The company reportedly provides an AI monitoring service that automatically detects inputted time-series data and image data that have features that are different from those at the time of training. The use of such AI monitoring services not only helps to improve profits by maintaining the accuracy of AI models, but is also relevant for accountability, or the responsibility to explain to stakeholders that they monitor their AI models to verify that they continue to maintain their performance.

Similar other efforts have also been reported. Fujitsu Limited has developed a technology that tracks the tendencies of changes that occur to inputted data in the full-scale operation to automatically estimate the AI model’s accuracy, as well as a technology that can suppress decreases in accuracy without re-training existing models. This technology is also expected to help to improve the accountability of AI system operators and user confidence.

It is important for AI system operators to understand the capabilities and limitations of AI systems through dialogue with AI system developers, and excessive reliance on automatic monitoring and other such tools should be avoided. Having said that, we will need to make wise, concurrent use of these automatic monitoring tools in the Society 5.0 era.

[24] The Consortium of Quality Assurance for Artificial-Intelligence-based Products and Services “Guidelines of Quality Assurance of Machine Learning-based Artificial Intelligence 2020.08 Edition” (August 2020), p. 2-7, whose February 2020 version is available in English.

(3) Considering ranking implementation of AI governance as non-financial information proactively disclosing the information

Action Target 4-3: Companies that develop and operate AI systems should consider ranking information relevant to AI governance such as one related to AI governance goal setting and establishment and operation of AI management systems as non-financial information in the Corporate Governance Code and proactively disclosing such information. Non-listed companies should also consider proactively disclosing information related to AI governance activities. If companies decide not to disclose such information after due consideration, they should be prepared to explain the reason externally.

[Practical Example 1]

We are a small company that develops AI systems. We believe that AI system development is not merely a technological activity, but one that must be founded on in-depth understanding of our society. As such, we place more emphasis on spreading this idea in our company than we do on explicitly setting AI governance goals. Our customers and shareholders are supportive of this approach. **We obviously believe that the "Social Principles of Human-Centric AI" should be respected, but understanding the ideas behind them is what is important**.

We are an unlisted company, so we are not subject to the Corporate Governance Code, but we actively publish our above-mentioned ideas on AI on our website and other outlets. Our potential customers and users of our AI systems see AI systems as a sociotechnical tool rather than a technical tool, and this also helps us differentiate ourselves from our competitors.

[Practical Example 2]

We are a listed company that develops AI systems. Because the appropriate development of AI is an important theme for our company, we have already defined our own AI policies and have established organizational systems designed to fulfill these policies. **We publish the details of these activities on our company website and have also issued press releases**. On a somewhat different note, we considered publishing a strong message from our management on these activities, but we have not gone ahead with this to date because our AI-related businesses currently do not have a direct impact on our medium- to long-term profits.

Against this backdrop, **we recently received a questionnaire from an institutional investor about corporate governance, and it included a question asking how we address AI ethics**. Assuming that investors' intentions for medium- to long-term growth of companies are reflected in this and similar questionnaires, we can surmise that information on AI ethics is necessary information for investors to determine whether or not they can expect sound growth of companies. **We plan to consider proactive dissemination of information on our AI ethics efforts from our management, including posting information on the efforts in our integrated report**.

Column: Information disclosure and AI ethics initiatives in corporate governance

AI governance cannot be separated from corporate governance. Here, we would like to discuss one of the general principles of the Corporate Governance Code ー Ensuring Appropriate Information Disclosure and Transparency ー in the context of AI governance. This basic principle states, “Companies should appropriately make information disclosure in compliance with the relevant laws and regulations, but should also strive to actively provide information beyond that required by law. This includes both financial information, such as financial standing and operating results, and non-financial information, such as business strategies and business issues, risk and governance. The board should recognize that disclosed information will serve as the basis for constructive dialogue with shareholders, and therefore ensure that such information, particularly non-financial information, is accurate, clear and useful.” The disclosure of information on risk assessments and actions associated with AI system development and operation may also be expected in some cases as part of such non-financial information.

In relation to this general principle, in interviews that the secretariat of AI Governance Guidelines carried out when preparing the Guidelines, we learned from interviewees that they had received inquiries regarding AI governance from European institutional investors. We believe the backdrop to these inquiries is the gradual rise in interest in AI governance among investors. For example, Hermes EOS (Equity Ownership Services) has stated to the board of directors of Google’s parent company, Alphabet, that “[i]nvestors are looking to” them “to display leadership in the responsible use of AI.”[25] Depending on the applications of AI systems, the impact that these systems have on society can grow the more they are used, and investors may require responsible AI system development and operation, as well as the establishment of AI governance.

[25] Alex Rolandi, “Hermes EOS urges Alphabet to lead responsible AI practice,” Funds Europe (June 18, 2019), https://www.funds-europe.com/news/hermes-eos-urges-alphabet-to-lead-responsible-ai-practice.

5. Evaluation

(1) Verifying an AI management system works appropriately

> Action Target 5-1: Companies that develop and operate AI systems should, under the leadership of top management, have individuals independent of the design and operation of the AI management system verify whether an AI management system such as a gap analysis process is appropriately designed and operated in light of the AI governance goals, in other words, whether an AI management system appropriately works for the achievement of the AI governance goals through the implementation of Action Targets 3 and 4.

[Practical Example 1]

We had an independent internal audit department for auditing the operation of our internal rules before the introduction of our AI management system. **When we introduced the AI management system, we extended the internal audit department's scope of work** to include the AI management system as one of their audit subjects. At our company, internal auditors investigate and check, with the cooperation of the relevant departments, whether their organizational setup, rules, etc. are properly operated and functioning effectively. And if any inappropriate operation or dysfunction is found, the department in question is asked to remedy the situation, and best practices of other departments, if available, are shared with them.

Social acceptance of AI systems is changing. We believe that **making improvements that are in line with social acceptance is key**. As such, **by referring to analyses of conditions and risks, we focus our internal audits on areas that society has high expectations for and areas with a large number of reported incidents**. In order to obtain cooperation for improvement from departments, we select high-risk fields instead of uniformly conducting rigid assessments in all fields for their conformity to the internal rules, etc. **Letting departments know why they were chosen for the audit makes it easier to obtain their cooperation**.

[Practical Example 2]

We are a small company that develops AI systems. We do not have an internal audit department to evaluate our AI management system, but **we conduct self-audits carried out by people in the development department who are not directly involved in the AI management system.** This first line of audit, the self-audit, is potentially prone to becoming dysfunctional in its checking functions, with a tendency on the part of auditing persons to make allowances for their colleagues. As such, **they are instructed to report the self-audit results to an auditor who reports directly to the AI governance officer, and then the report is organized and sent to the AI governance officer**. Even though the audit primarily consists of self-auditing, we believe that it is fully functional because the AI governance officer is knowledgeable on AI technology and ethics. Currently, we are considering enhancing the third-party perspectives, and sharing audit results and exchanging opinions at crossdepartmental feedback meetings to communicate that the purpose of these internal audits is to make improvements.

[Practical Example 3]

We have an internal audit department, but we decided to try making use of external audits for our AI management system. **The expectations we have for these external auditors are their high levels of expertise and application of experience that they have gained from their audits of other companies**. **Social acceptance of AI systems continues to change, and no normative market sentiment has yet formed**. We may have blind spots even if we claim to be fully prepared as best we can.

External audit services are mostly provided by consulting firms. Being audited by external experts allows us to receive advice based on specialized information sourced from inside and outside the company. We also expect that the third-person nature and objectivity of advice from external experts will make for smoother feedback within the company.

Notwithstanding these benefits, we are concerned that **we may potentially lapse into passivity**. **External experts are not always familiar with the issues unique to individual companies**. In order to make the best use of advice from external experts, it is **important to have the proactive intention to understand the social acceptance of AI even as we enlist the services of external auditors**.

(2) Considering seeking feedback from outside stakeholders

Action Target 5-2: Companies that develop and operate AI systems should, under the leadership of top management, consider seeking opinions on their AI management system and the implementation of such system from not only their shareholders but also from various stakeholders such as their business partners, users including consumers, experts who are familiar with trends on the appropriate operation of AI systems, non-governmental organizations and labor unions. If companies decide not to seek opinions outside after due consideration, they should be prepared to explain the reason externally.

[Practical Example 1]

Compiled in the Chapter "Appropriate Cooperation with Stakeholders Other Than Shareholders" in the Corporate Governance Code are principles that state that companies should endeavor to appropriately cooperate with a range of stakeholders, including employees, customers, business partners, creditors and local communities, and the board and the top management, in particular, should exercise their leadership in establishing a corporate culture where the rights and positions of stakeholders are respected and sound business ethics are ensured. **Because interest in proper AI system development and operation continues to grow, listed companies as a matter of course as well as unlisted companies, in some cases, are encouraged to collaborate with various stakeholders when evaluating and revising their AI governance and management systems**.

While we believe that defining AI policies and making initial settings, such as establishing an organizational setup for fulfilling these policies, should be carried out by the companies themselves, and that subsequent improvements should also be proactively carried out by the companies themselves, **we also place great importance on collaborating with various stakeholders to gain knowledge on "how society sees us**."

We have already established an AI policy and have publicly announced the meaning of the policy and the activities to fulfill it. However, based on our belief that we need to understand "how society sees us" and ensure that we achieve objective ethicality, **we have decided to install an AI Ethics Committee made up of experts in the field of AI and other fields** aimed at engaging in dialogue with our stakeholders, and have invited experts in AI technology, as well as experts from other fields such as those knowledgeable on legal, environmental, and consumer issues. Since receiving general comments is not enough, **we make sure to present issues that are specific to our company so that we can gain a deeper insight**.

[Practical Example 2]

While we may tend to direct our attention to "visible measures" such as establishing a committee of external experts as described in Practical Example 1, we do not believe that such venues are an end-all solution. **The key is to loosely connect with and join information exchange networks of**

people who are interested in AI ethics and quality. **We encourage our AI management personnel to actively speak at gatherings where people come to exchange views on AI ethics and quality, and actively take on speaker roles at conferences and similar gatherings**. Of course, these activities are taken into account in their performance evaluations.

We hear concerns that these approaches are not conducive to gathering opinions. We believe that the backdrop for this concern may lie in how Japanese people at times may not speak their mind at opinion exchange meetings or conferences. However, so-called "**active sonar type" people** who express their views and consequently draw out the opinions of others **know that there are people who are happy to share their views personally after these opinion exchange meetings or conferences**. It is precisely these opinions that are important.

We have actually held an in-house training session with an outside lecturer who we were able to reach out through the network of contacts of our AI management personnel. During the training session, we explained our AI governance efforts to employees engaged in AI-related work, and we asked the outside lecturer to evaluate our efforts. **Since this outside lecturer exchanges opinions with our AI management personnel on a daily basis, we obtained advice that was in line with our circumstances and the session was received very well by session participants**.

Given this state of affairs, we are considering establishing a committee of external experts, but so far we have not felt the need.

[Practical Example 3]

We develop AI systems and deliver them to other companies. These other companies utilize the AI systems only for their own business operations, and do not provide them to consumers. For this reason, we do not believe that these systems are directly associated with ethical issues such as consumer safety, serious damage to property value, or discrimination. **We obviously exchange opinions with our customers in depth from the perspective of maintaining and improving forecast accuracy**, but we do not see the need to listen to the opinions of consumers or experts who are familiar with AI ethics.

6. Re-analysis of conditions and risks

(1) Re-implementing Action Targets 1-1 to 1-3 in a timely manner

Action Target 6-1: Companies that develop and operate AI systems should, under the leadership of top management, conduct re-evaluations, update their understanding, obtain new points of views, or take other relevant actions with respect to Action Targets 1-1 through 1.3, in a timely manner. When implementing Action Target 5-2, they should also consider obtaining opinions not only for the current AI management system and the operation of such system, but also conducive to review of entire AI governance, including analyses of conditions and risks.

[Practical Example 1]

Because, as of the present time, a standard perception of AI systems has yet to form in our societies, we should re-evaluate the positive and negative impacts that AI systems can have, societies' acceptance of AI system development and operation, and our own AI proficiency as described in Action Targets 1-1 through 1-3, as well as update our understanding and acquire new perspectives in a timely manner. **We regularly perform analyses of conditions and risks, and report to management even during normal times where we do not see any "near misses," or a significant increase in public attention to specific incidents, or changes in the regulatory environment**. While there is an active ongoing debate over how to properly develop and operate AI systems, we place emphasis on avoiding governance fatigue by haphazard agile re-analyses and using agile re-analyses to understand major trends. Opportunities to report to management are good opportunities for directing our attention to major trends.

[Practical Example 2]

We regularly analyze conditions and risks as described in Practical Example 1, but because verifications of AI governance and AI management systems have overlapping elements, we include **the positive and negative impacts of AI systems and social acceptance of AI system development and operation in the agenda of the AI Ethics Committee meetings that we hold regularly with invited external experts to gain their insights into the major trends associated with these points**.

D. Experts involved in the AI Governance Guidelines

1. Expert Group on How AI Principles Should be Implemented (Expert Group on Architecture for AI Principles to be Practiced)

Toshiya Watanabe	Professor, The University of Tokyo Institute for Future Initiatives (Chair)
Shunichi Amemiya	Head of Research and Development Headquarters, NTT DATA Corporation
Naoto Ikegai	Associate Professor, Graduate School of Law, Hitotsubashi University
Katsuya Uenoyama	Representative Director, PKSHA Technology Inc.
Takayoshi Kawakami	Partner, Industrial Growth Platform, Inc. Board Member, Japan Deep Learning Association
Tomokazu Saito	Partner, LAB-01 Law Office
Roy Sugimura	Supervisory Innovation Coordinator, Research Promotion Division for Artificial Intelligence, Department of Information Technology and Human Factors, National Institute of Advanced Industrial Science and Technology
Mihoko Sumida	Professor, Graduate School of Law, Hitotsubashi University
Yoshiki Seo	Councilor, Standardization Promotion Center, Research and Innovation Promotion Headquarters, National Institute of Advanced Industrial Science and Technology
Kenzaburo Tamaru	National Technology Officer, Microsoft Japan Co., Ltd.
Yoshihiro Tsuchiya	General Manager, Commercial Lines Underwriting Dept., Tokio Marine & Nichido Fire Insurance Co., Ltd.
Kaoru Chujo	President and CEO, SoW Insight, Inc.
Satoshi Hara	Associate Professor, The Institute of Scientific and Industrial Research, Osaka University
Shinnosuke Fukuoka	Partner, Nishimura & Asahi
Yukiko Furuya	President, Consumer Conference for Sustainability
Etsuko Masuda	President, Japan Association of Consumer Affairs Specialists
Tomoaki Maruyama	Chief Engineer, Digital AI technology Center, Technology Division, Innovation Promotion Sector, Panasonic Corporation
Kazuya Miyamura	Partner, PricewaterhouseCoopers Aarata LLC.
Tatsuhiko Yamamoto	Professor, Keio University Law School

* The following experts attended only the Expert Group on Architecture for AI Principles to be Practiced held last year.

Takenobu Aoshima	General Manager, Data Analysis Department, Digital & AI technology Center, Technology Division, Innovation Promotion Sector, Panasonic Corporation（as of the previous fiscal year）

2. AI Governance Guidelines Working Group

Naoto Ikegai	Associate Professor, Graduate School of Law, Hitotsubashi University (Chair)
Atsushi Okada	Partner, Mori Hamada & Matsumoto
Tomokazu Saito	Partner, LAB-01 Law Office
Takashi Nakazaki	Attorney at law, Anderson Mori & Tomotsune
Yosuke Motohashi	Senior Manager, AI & Analytics Division, NEC Corporation
Kazuya Miyamura	Partner, PricewaterhouseCoopers Aarata LLC.

3. Experts who provided support

In addition to above-mentioned members, a number of experts provided their support in the formulation of the Guidelines. We would like to express our sincere gratitude to them.

(1) Lectures at the Expert Group Meetings Above (in order of Lecture Date)

Takeshi Ogino	General Manager, Future Technology Promotion Department, Production Division, Kewpie Corporation (when he gave the lecture)
Yutaka Oiwa	Deputy Director, Digital Architecture Research Center, Department of Information Technology and Human Factors, National Institute of Advanced Industrial Science and Technology
Roy Sugimura	Supervisory Innovation Coordinator, Research Promotion Division for Artificial Intelligence, Department of Information Technology and Human Factors, National Institute of Advanced Industrial Science and Technology
Toyokazu Nagamune	Secretary General, Japan Business Council in Europe
Chihiro Saito	Consul, Consulate General of Japan in Strasbourg, Ministry of Foreign Affairs

(2) Experts who participated in the preliminary consultation which was the expanded version of the above-mentioned Working Group

Toshikazu Imada	Senior Manager, AI Ethics Office, AI Collaboration Office, Sony Group Corporation
Yutaka Oiwa	Deputy Director, Digital Architecture Research Center, Department of Information Technology and Human Factors, National Institute of Advanced Industrial Science and Technology
Shinji Kikuchi	Research Manager, AI Quality Project, Artificial Intelligence Laboratory, Fujitsu Research, Fujitsu Limited
Roy Sugimura	Supervisory Innovation Coordinator, Research Promotion Division for Artificial Intelligence, Department of Information Technology and Human Factors, National Institute of Advanced Industrial Science and Technology

Yoshiki Seo	Councilor, Standardization Promotion Center, Research and Innovation Promotion Headquarters, National Institute of Advanced Industrial Science and Technology
Masaru Sogabe	CEO, GRID INC.
Masahiro Fujita	VP, AI Collaboration Office, Sony Group Corporation
Takashi Matsumoto	Manager, Assurance, Risk Advisory, Deloitte Touche Tohmatsu LLC. Visiting Researcher, Institute for Future Initiatives, The University of Tokyo
Tadashi Mima	Director, Smart infrastructure Consulting Department (2nd), Hitachi Consulting Co., Ltd. Specially Appointed Professor, Graduate School of Media and Governance, Keio University
Rikiya Yamamoto	Humanoid Division Director, Business Development Division, Softbank Robotics Japan

(3) Individual interviews conducted by the secretariat

In addition to the above-mentioned Expert Group, etc. we conducted individual interviews with companies and organizations on their initiatives. We received help from many experts from Musashi AI Ltd., Japan Deep Learning Association, NEC Corporation, the National Institute of Advanced Industrial Science and Technology, NVIDIA Corporation, Fujitsu Limited, Hitachi, Ltd., Mercari, Inc., Yokogawa Electric Corporation, Sompo Japan Insurance Inc., AI Business Promotion Consortium, J.Score CO., LTD., Deloitte Touche Tohmatsu LLC, Tokio Marine & Nichido Fire Insurance Co., Ltd., Panasonic Corporation, GRID INC., Chiyoda Corporation, Sony Group Corporation, Hitachi Zosen Corporation, SoftBank Corp., and SoftBank Robotics Corp. If any company/orgarnization is omitted from above list, it is entirely the fault of the secretariat.

Please note that experts who provided support were not involved in the finalization the Guidelines.

4. Secretariat

Yohei Matsuda	Director, Digital Economy Division, Commerce and Information Policy Bureau, Ministry of Economy, Trade and Industry, Japan (METI) (before June 30, 2021)
Chizuru Suga	Director, Digital Economy Division, Commerce and Information Policy Bureau, Ministry of Economy, Trade and Industry, Japan (METI) (after July 1, 2021)
Takuya Izumi	Director for Information Policy Planning, Digital Economy Division, Commerce and Information Policy Bureau, Ministry of Economy, Trade and Industry, Japan (METI) (Lead Author)
Hiroki Habuka	Deputy Director for Global Digital Governance, Digital Economy Division, Commerce and Information Policy Bureau, Ministry of Economy, Trade and Industry, Japan (METI) (Contributed Appendix 3)
Tamotsu Nomura	Deputy Director, Digital Economy Division, Commerce and Information Policy Bureau, Ministry of Economy, Trade and Industry, Japan (METI)

E. Bibliography

1. Documents created by the government/public agencies

- The Ministry of Economy, Trade and Industry, “Report titled ‘GOVERNANCE INNOVATION Ver.2: A Guide to Designing and Implementing Agile Governance’” (July 31, 2021) .
- The Ministry of Economy, Trade and Industry, “Report titled “GOVERNANCE INNOVATION: Redesigning Law and Architecture for Society 5.0’” (July 13, 2020).
- The Ministry of Economy, Trade and Industry. “Contract Guidelines on Utilization of AI and Data Version 1.1” (December 2019), whose original version is available in English.
- The Ministry of Economy, Trade and Industry, Ministry of Health, Labour and Welfare, and the Fire and Disaster Management Agency, “Guidelines on Assessment of AI Reliability in the Field of Plant Safety, Second Edition”, “Practical Examples of Reliability Assessment Records” (March 30, 2021).
- The Ministry of Economy, Trade and Industry, “Practical Guidebook on Data Provision for Fostering Human Resources of Experts in AI and Data Science” (March 1, 2021), available only in Japanese.
- The Ministry of Economy, Trade and Industry, “Digital Governance Code” (November 9, 2020), available only in Japanese.
- The Ministry of Internal Affairs and Communications, and the Ministry of Economy, Trade and Industry, “Guidebook on Corporate Governance for Privacy in Digital Transformation (DX) ver.1.1” (July 2021), available only in Japanese.
- The Ministry of Internal Affairs and Communications, and the Conference toward AI Network Society, “Draft AI R&D GUIDELINES for International Discussions” (July 2017).
- The Ministry of Internal Affairs and Communications, and the Conference toward AI Network Society, “AI Utilization Guidelines - Practical Reference for AI Utilization” (August 2019).
- The AI Working Group under the Committee on Challenges to Consumer Digitalization, “Report by the AI Working Group under the Committee on Challenges to Consumer Digitalization” (July 2020), available only in Japanese.
- The National Institute of Advanced Industrial Science and Technology, “Machine Learning Quality Management Guideline (Version 2)” (July 2020) (The Digital Architecture Research Center / Cyber Physical Security Research Center / Artificial Intelligence Research Center, Technical Report. DigiARC-TR-2021-01 / CPSEC-TR-2021001), whose original version is available in English.
- Information-technology Promotion Agency, “AI White Paper 2020 - Widening of the AI Gap and Corporate Strategies with an eye on 5 years ahead” (March 2, 2020), available only in Japanese.
- The Tokyo Stock Exchange, “Corporate Governance Code - Seeking Sustainable Corporate Growth and Increased Corporate Value over the Mid- to Long-Term” (June 11, 2021).

- Executive Office of the President, Office of Management and Budget, "Memorandum to the Heads of Executive Departments and Agencies, Guidance for Regulation of Artificial Intelligence Applications" (November 17, 2020).
- European Commission, "White Paper on Artificial Intelligence – A European approach to excellence and trust" (February 19, 2020).
- The High-Level Expert Group on Artificial Intelligence (AI HLEG), "Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment" (July 17, 2020).
- European Commission, "Proposal for a Regulation laying down harmonised rules on artificial intelligence" (April 21, 2021).
- Info-communications Media Development Authority and Personal Data Protection Commission, "Model Artificial Intelligence Governance Framework Second Edition" (January 21, 2020).
- World Economic Forum, "Companion to the Model AI Governance Framework –Implementation and Self-Assessment Guide for Organizations" (January 2020).
- P J. Phillips, Amanda C. Hahn, Peter C. Fontana, David A. Broniatowski, Mark A. Przybocki, "Four Principles of Explainable Artificial Intelligence (Draft), NIST Interagency/Internal Report (NISTIR) - 8312-draft" (August 19, 2020).
- OECD, "OECD FRAMEWORK FOR THE CLASSIFICATION OF AI SYSTEMS – PUBLIC CONSULTATION ON PRELIMINARY FINDINGS" (May 2020).
- Carolyn Nguyen, Adam Murray, and Barry O'Brien, "What are the tools for implementing trustworthy AI? A comparative framework and database," The AI Wonk, OECD.AI (May 25, 2021).

2. Documents created by companies and industry associations

- The Consortium of Quality Assurance for Artificial-Intelligence-based Products and Services, "Guidelines for Quality Assurance of Machine Learning-based Artificial Intelligence (August 2020 Version)" (August 2020), whose February 2020 version is available in English.
- KEIDANREN (Japan Business Federation), "AI Utilization Strategy - For an AI-Ready Society" (February 19, 2019), whose summary is available in English.
- NEC Corporation, "NEC Group AI and Human Rights Principles" (April 2019).
- Fujitsu Limited, "Fujitsu Group AI Commitment" (March 13, 2019)
- Sony Group Corporation, "Sony Group AI Ethics Guidelines" (Amended in March 2019)
- Hitachi, Ltd. "Principles for the Ethical Use of AI" (February 22, 2021).
- Andrew Ng, "AI Transformation Playbook", available at https://landing.ai.

3. Documents created by individuals and academia

- Personal Data + α Study Group, “Draft Recommendation on Profiling” NBL No.1137 (January 1, 2019), available only in Japanese.
- Personal Data + α Study Group, “Interim Report Attached to the Draft Recommendation on Profiling” NBL No.1137 (January 1, 2019), available only in Japanese.
- Tatsuhiko Yamamoto. “AI and the Constitution”, Nikkei Publishing Inc. (August 24, 2018), available only in Japanese.
- Shinichi Asakawa, Arisa Ema, Fumiko Kudo, Yusuke Sugomori, Keisuke Seya, Takayuki Matsui, and Yutaka Matsuo, “Deep Learning G-Certificate (Generalist) Official Textbook”, Shoeisha (October 22, 2018), available only in Japanese.
- Satoshi Funayama, “Responsibilities and Ethics of AI (No.2) AI Ethics Initiatives by Companies (1)” NBL No.1170 (May 15, 2020), available only in Japanese.
- Junichi Arahori, “Responsibilities and Ethics of AI (No.3) AI Ethics Initiatives by Companies (2)” NBL No.1172 (June 15, 2020), available only in Japanese.
- Tomokazu Saito, “Responsibilities and Ethics of AI (No.4) AI Ethics and Accountability, Legal Liability” NBL 1174 (July 15, 2020), available only in Japanese.
- Chieko Matsuda, “Easy to Understand Corporate Governance Textbook”, Nikkei Business Publications, Inc. (August 11, 2015), available only in Japanese.
- Chieko Matsuda, “Corporate Governance Practice for the Enhancement of ESG Management”, Nikkei Business Publications, Inc. (December 24, 2018), available only in Japanese.
- Tadashi Kunihiro, “Preventing Corporate Scandals”, Nikkei Publishing Inc. (October 17, 2019), available only in Japanese.
- Yoshito Hirabayashi, and Maiko Okuno, “ISO Universal Textbook”《Appendix SL》 Commentaries and Use - ISO Management System - Enhancing the Performance of a Structured Organization”, Japanese Standards Association (October 14, 2015), available only in Japanese.
- Yukiko Furuya, “Consumer Sovereignty in the Modern Age” Fuyoshobo (May 19, 2017), available only in Japanese.
- Cathy O’Neil, “Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy” New York: Crown Publishers (2016).

F. Appendix 1 (List of Action Targets)

Action Target 1-1: Companies that develop and operate AI systems should, under the leadership of top management, understand not only positive impacts but also negative impacts, including unintended risks, that AI systems may have. This information should be reported to the top management and shared among those in top managerial positions, and their understanding should be updated in a timely manner.	
Action Target 1-2: Companies that develop and operate AI systems should, under the leadership of top management, understand the current state of social acceptance based on opinions of not only direct stakeholders, but potential stakeholders before full-scale provision of the AI systems. In addition, even after the full-scale operation, companies should obtain opinions of stakeholders again and update their perspectives in a timely manner.	
Action Target 1-3: Companies that develop and operate AI systems should, under the leadership of top management, evaluate and re-evaluate in a timely manner their AI proficiency based on the extent of the company's experience in developing and operating AI systems, the number of employees, including engineers, involved in the development and operation of AI systems and their degree of experience, and the degree of AI literacy of these employees with respect to AI technology and ethics, except in situations where a company assesses negative impacts of their AI system are minor based on analyses of Action Targets 1-1 and 1-2 in light of the company's business domain and scale, etc. If the negative impacts are assessed to be minor and no evaluation of AI proficiency is carried out, companies should be prepared to explain their rationale to their stakeholders.	
Action Target 2-1: Companies that develop and operate AI systems should, under the leadership of top management, consider whether or not to set their own AI governance goals (for example an AI policy) based on the "Human-Centric Social Principles of AI", keeping in mind importance of the process leading up to goal setting, and taking into consideration the positive and negative impacts that AI systems can have, social acceptance regarding the development and operation of AI systems, and their own AI proficiency. And if a company decides not to set AI governance goals based on the assessment that their potential negative impacts are minor, they should be prepared to explain their rationale to their stakeholders. If it is determined that the "Human-Centric Social Principles of AI" function fully satisfactorily, this may be set as the goals instead of the company's own AI governance goals. Note that even if no goal is set, it is desirable to understand the importance of the "Human-Centric Social Principles of AI" and to implement measures associated with Action Targets 3 to 5 as needed.	
Action Target 3-1: Companies that develop and operate AI systems should, under the leadership of top management, identify a gap between AI governance goals and current state in the AI systems that they are developing and operating, and if any negative impacts are found upon evaluating the impacts resulting from the gap, determine whether or not the negative impacts would be acceptable, taking into account their severity, scope, and frequency of occurrence. They should incorporate processes that prompt a reexamination of how the AI systems should be developed and operated in an appropriate stage such as during AI system design, development,	

before they are used, and after their usage begins, to address cases where the negative impacts are found not to be acceptable. Those in operations positions should make these processes concrete. In addition, those who are not directly involved in the development and operation of AI systems should be included in the gap analysis between AI governance goals and current state. It should be noted that it would not be appropriate to stop the development or provision of AI simply because a gap was found. As such, gap analysis is merely a step for evaluating negative impacts and simply serve as a starting point for improvement.	
Action Target 3-1-1: Companies that develop and operate AI systems should, under the leadership of top management, check whether a standard gap analysis process in their industry is available and incorporate it into their own process if such a process is available.	
Action Target 3-1-2: In case that a certain degree of gaps may potentially occur with AI systems of AI system operators that provide services to AI system users, they should, under the leadership of top management, provide sufficient information about the gaps and measures to address the gaps, as well as make a contact point easily accessible.	
Action Target 3-1-3: Companies that provide data should provide information on the data sets including data collection sources, collection policies, collection criteria, annotation criteria, and limitation on use to ensure that companies that develop AI systems are able to appropriately conduct gap analysis, and AI system developers should acquire data sets from data providers that provide sufficient information.	
Action Target 3-2: Companies that develop and operate AI systems should, under the leadership of top management, strategically improve AI literacy in order to properly operate their AI management system, considering outside learning materials as an option. For example, this may include education to boost general literacy on AI ethics for those in the top management, management teams, and those in operations positions responsible for legal and ethical aspects of AI system development and operation, as well as training for those responsible for AI system development and operation not only on AI ethics but also on AI technology. Companies that provide data should take steps to improve the general literacy of AI ethics of employees engaged in data provision by referring to practical examples for AI system developers and operators.	
Action Target 3-3: Under the leadership of top management and with due consideration for trade secrets, companies that develop and operate AI systems and those that provide data should, except where everything from the preparation of data sets for training and other purposes to AI system development and operation is performed entirely within their own department, clarify and actively share, in accordance with the Principle of Fair Competition, AI system operational issues that the company or department is unable to fully address on their own and the information necessary to address these issues. In doing so, in order to facilitate the exchange of information clarified above, the AI system developer, AI system operator, and data provider are encouraged to agree on scope of information disclosure in advance and consider measures to protect trade secrets, for example, by entering a non-disclosure agreement.	
Action Target 3-3-1: Under the leadership of top management and with due consideration for trade	

secrets, companies that develop and operate AI systems and those that provide data should, except where everything from the preparation of data sets for training and other purposes to AI system development and operation is performed entirely within their own company, understand the current state of information sharing between companies, and update their understanding in a timely manner.	
Action Target 3-3-2: Companies that develop and operate AI systems should, under the leadership of top management, regularly collect relevant information such as formulation of rules for the development and operation of AI systems, best practice, and incidents, and encourage the exchange of views within and outside the company.	
Action Target 3-4: Companies that develop and operate AI systems and those that provide data should, under the leadership of top management, reduce incident-related burdens on users by preventing incidents and through early response.	
Action Target 3-4-1: Under the leadership of top management, companies that develop and operate AI systems and those that provide data should, except where everything from the preparation of data sets for training and other purposes to AI system development and operation is performed entirely within their own company, allocate burdens of addressing uncertainties of AI systems among companies so that negative impacts may be minimized as a whole.	
Action Target 3-4-2: Companies that develop and operate AI systems should, under the leadership of top management, consider defining response guidelines and plans so that upon occurrence of an AI incident or dispute, they can promptly give an explanation to AI system users, identify the extent of the impact and damage, clarify legal responsibilities, consider relief measures and measures to prevent the spread of damage and recurrence, or take other relevant actions. Further, they should consider conducting rehearsal exercise relevant to such guidelines and plans, as appropriate.	
Action Target 4-1: Companies that develop and operate AI systems should, under the leadership of top management, make sure that they are ready for explanation about the operating status of AI management systems externally by recording the gap analysis process under Action Target 3-1 and by taking other relevant actions.	
Action Target 4-2: Companies that develop and operate AI systems should, under the leadership of top management, monitor and record the status of preliminary and full-scale operations so that gap analysis for individual AI systems in preliminary and full-scale operations can be continuously implemented. Companies that develop AI systems should assist the monitoring conducted by companies that operate AI systems.	
Action Target 4-3: Companies that develop and operate AI systems should consider ranking information relevant to AI governance such as one related to AI governance goal setting and establishment and operation of AI management systems as non-financial information in the Corporate Governance Code and proactively disclosing such information. Non-listed companies should also consider proactively disclosing information related to AI governance activities. If companies decide not to disclose such information after due consideration, they should be	

prepared to explain the reason externally.	
Action Target 5-1: Companies that develop and operate AI systems should, under the leadership of top management, have individuals independent of the design and operation of the AI management system verify whether an AI management system such as a gap analysis process is appropriately designed and operated in light of the AI governance goals, in other words, whether an AI management system appropriately works for the achievement of the AI governance goals through the implementation of Action Targets 3 and 4.	
Action Target 5-2: Companies that develop and operate AI systems should, under the leadership of top management, consider seeking opinions on their AI management system and the implementation of such system from not only their shareholders but also from various stakeholders such as their business partners, users including consumers, experts who are familiar with trends on the appropriate operation of AI systems, non-governmental organizations and labor unions. If companies decide not to seek opinions outside after due consideration, they should be prepared to explain the reason externally.	
Action Target 6-1: Companies that develop and operate AI systems should, under the leadership of top management, conduct re-evaluations, update their understanding, obtain new points of views, or take other relevant actions with respect to Action Targets 1-1 through 1.3, in a timely manner. When implementing Action Target 5-2, they should also consider obtaining opinions not only for the current AI management system and the operation of such system, but also conducive to review of entire AI governance, including analyses of conditions and risks.	

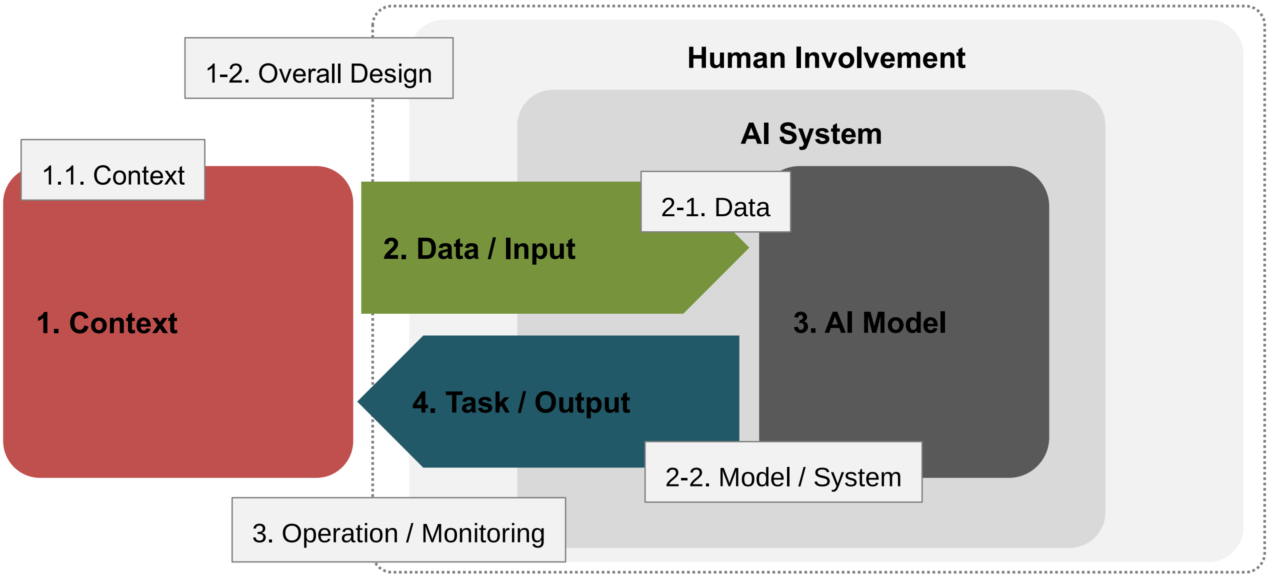
G. Appendix 2 (Practical Examples for Gap Analysis between AI Governance Goals and Current State)

These examples for gap analysis are a tool to support the evaluation of conformity to the AI governance goals in the development and operation of individual AI systems. Examples of items for analysis that are shown here are expected to be incorporated into the governance system of each AI system developer/operator, as necessary, along with the gap analysis process using these examples as well as other AI governance elements, and to serve as a support tool for achieving the social goals of the "Social Principles of Human-centric AI".

Development and operation of AI systems, their objectives and methods of developing/operating AI systems vary widely, and circumstances may also vary between a demonstration project and full-scale development. Therefore, it is not necessarily assumed that gap analyses are carried out in a uniform manner for development/operation of all AI systems. What kind of AI systems should be analyzed and what criteria should be applied in selecting the AI system to be analyzed are left up to the reasonable discretion of AI system developers/operators. AI system developers/operators should, in light of their businesses, etc., extract types of AI system development/operation which could have a large negative impact resulting from a gap between the AI governance goals and current state, and conduct a gap analysis.

As mentioned in "A. Introduction", **the examples for gap analysis do not take into account the specific circumstances of individual AI system developer/operator, hence some of the examples of items for analysis provided as examples for gap analysis may be insufficient or oversufficient depending on the subject to be analyzed**. Therefore, **the decision on whether to adopt them is obviously left up to the discretion of AI system developers/operators, and if they decide to adopt them, they need to consider making modifications or selecting appropriate examples according to their circumstances.**

These examples for gap analysis are based on the perspectives of context of individual project, overall design of individual project, data, model/system, and operation/monitoring. These perspectives are mapped into the conceptual diagram below, which was created by referring to the OECD's framework.



1. Planning / Design Phase		
Examples of Items for Analysis	Specific Examples	
1.1. Context		
A. Do the AI system developer and operator anticipate potential users? → 1-2. Overall Design A → 3. Operation / Monitoring A	Example: Do the AI system developer and operator anticipate AI literacy and experience of use of AI systems of potential users of the AI system which they are going to develop/operate? Example: Do the AI system developer and operator understand that potential users of the AI system which they are going to develop/operate may or may not include children, elderly people, and other socially vulnerable groups?	
B. Do the AI system developer and operator anticipate potential applications of the AI system? → 1-2. Overall Design B → 3. Operation / Monitoring B	Example: Did the AI system developer ask the AI system operator the purpose of providing the AI system which they are going to develop? Example: Do the AI system developer and operator envision the same application? Example: Do the AI system developer and operator understand that the AI system which they are going to develop/operate may not only be used in typical applications of the system but also be misused.	
C. Is the AI system developer aware of the AI literacy and experience of the AI operator? → 1-2. Overall Design C → 3. Operation / Monitoring C	Example: Is the AI system developer aware of the number of the employees with AI literacy in the AI system operator, the number of their AI training opportunities, and their experience in operating AI systems? Example: Does the AI system developer know whether the AI system operator is willing to enhance their AI literacy if their AI literacy and experience are insufficient?	
D. Are the AI system developer and operator aware of the potential negative impacts on physical and mental health and property, etc. of the AI system?	Example: Did the AI system developer and operator investigate incidents associated with the negative impacts of the AI system they are going to develop/operate and/or other similar AI systems on physical and mental health and property, etc.?	

→ 1-2. Overall Design D → 3. Operation / Monitoring D	Example: Are the AI system developer and operator aware of the scale and frequency of possible damage the AI system they are going to develop/operate may cause to physical and mental health and property, etc.? Example: If the AI system which the AI system developer is going to develop could have some impact on physical and mental health and property, etc., did the AI system developer assess the risk acceptable in the society in light of the standard industry practice, etc.?	
E. Are the AI system developer and operator aware of the fairness that is required of the AI system? → 1-2. Overall Design E → 3. Operation / Monitoring E	Example: Did the AI system developer and operator investigate incidents related to fairness of the AI system they are going to develop/operate and/or other similar AI systems? Example: Did the AI system developer and operator check whether it is said that there are or there remain biases and/or discriminatory treatment associated with some of the potential users in the countries/regions where the AI system will be used? Example: Did the AI system developer and operator check whether it is said that the same or a similar AI system has created, reproduced, or amplified biases and/or discriminatory treatment associated with some of the potential users in the countries/regions where the AI system will be used? Example: Did the AI system developer, where possible, select an appropriate index of fairness from those proposed, and conduct an assessment on the range acceptable in the society?	
F. Do the AI system developer and operator understand considerations given to individuals that are expected of the AI system? → 1-2. Overall Design F → 3. Operation / Monitoring F	Example: Did the AI system developer and operator investigate incidents that caused by lack of considerations given to individuals in the AI system they are going to develop/operate and/or other similar AI systems? Example: Do the AI system developer and operator understand that some AI systems have ability to identify	

	detailed characteristics of specific individuals using numerous fragments of information, and that concerns over such ability have been expressed?	
G. Are the AI system developer and operator aware of the issues of cybersecurity that the AI system is expected to address? → 1-2. Overall Design G → 3. Operation / Monitoring G	Example: If the AI system which AI system developer and operator are going to develop/operate is used by connecting it to the Internet, did they investigate incidents related to the cybersecurity of the AI system and/or other similar AI systems? Example: If the AI system which AI system developer and operator are going to develop/operate is used by connecting it to the Internet, did they identify any cybersecurity challenges by checking in light of standard practices, etc. in their industry?	
1-2. Overall Design		
A. Did the AI system developer address issues such as lack of literacy and experience among the potential users?	Example: Did the AI system developer devise a better interface for the AI system which they are going to develop, or summarize precautions for the AI system operator and users, in order to improve the user's understanding? Example: Did the AI system developer summarize not only the benefits of the AI system which they are going to develop but also the limitations of the system for the AI system operator and users in an easy-to-understand manner, in order to improve the user's understanding?	
B. Did the AI system developer address issues related to foreseeable misuse?	Example: Did the AI system developer consider whether it is possible by design to eliminate foreseeable misuse related to the AI system which they are going to develop, and if it is possible, adopt such design? Example: If it is not possible by design to eliminate foreseeable misuse related to the AI system which the AI system developer is going to develop, did the AI system developer summarize precautions for the AI system operator? Example: Did the AI system developer make it clear to the AI system operator in the contract that unintended use is	

	prohibited, ensuring that the AI system will not be used for purposes other than those identified at the time of delivery?	
C. Did the AI system developer address issues such as lack of literacy and experience on the part of the AI system operator?	Example: Did the AI system developer devise a better interface for the AI system which they are going to develop, or summarize precautions for the AI system operator, in order to improve their understanding? Example: Did the AI system developer summarize not only the benefits of the AI system which they are going to develop but also the limitations of the system for the AI system operator in an easy-to-understand manner, in order to improve their understanding?	
D. Did the AI system developer address the issues of negative impacts the AI system could have on physical and mental health and property, etc.?	Example: With respect to negative impacts on physical and mental health and property, etc., did the AI system developer divide the risks into foreseeable risks that can be eliminated across the AI system and residual risks which cannot be addressed; mitigate the residual risks as much as possible; and make sure that the residual risks may be managed by alerting the AI system operator and users or by specifying the method of use? Example: Did the AI system developer anticipate negative impacts on physical and mental health and property, etc. which could arise if input data contains any abnormal values, and take any measures to them? Example: Did the AI system developer consider possibility that input data could contain malicious data and/or adversarial data; anticipate negative impacts on physical and mental health and property, etc. which could arise if such data is included in the input data; and take any measures to them? Example: Did the AI system developer anticipate negative impacts on physical and mental health and property, etc. which could arise if output data contains any abnormal values, and take any measures to them? Example: Did the AI system developer consider not only giving due consideration to the AI model itself but also adding	

	redundancy and safety features for the overall system including the AI model, in order to address issues associated with the explainability of the AI model?	
E. Did the AI system developer address the issues related to fairness of the AI system?	Example: Did the AI system developer increase to the extent possible the diversity of the AI system development team? Example: If the AI system is expected to be provided to another country/region, did the AI system developer add to the extent possible someone who understands the regulations, customs, business practices, etc. of that country/region to the AI system development team and/or as the person in charge of reviewing the AI system? Example: If the AI system developer selected an appropriate index of fairness from those proposed, did they take a measure to warn against output data that is outside the acceptable range? Example: If the AI system developer selected an appropriate index of fairness from those proposed, and there is a possibility that data outside the acceptable range may be output, did the AI system developer summarize such possibility as a precaution for the AI system operator and users?	
F. Did the AI system developer address the considerations given to individuals that are expected of the AI system?	Example: Did the AI system developer ensure that opportunities and decision-making of individuals will not be unduly hindered while developing an AI system that can analyze individuals? Also, did the AI system developer document any actions taken for that purpose, and make themselves accountable to the AI system operator and users? Example: Did the AI system developer check whether any analysis is conducted to identify individuals from anonymized data, etc.?	
G. Did the AI system developer address the issues related to cybersecurity of the AI system?	Example: If the AI system which the AI system developer is going to develop/operate will be used by connecting it to the Internet, did the AI system developer take any measures to address unauthorized access and/or data input in light of	

	standard practices, etc. in their industry? Example: Did the AI system developer summarize measures against unauthorized access and/or data input as a precaution for the AI system operator?	
H. Did the AI system developer ensure as part of the system design that there are opportunities for humans to get actively involved in the system by their own initiative, when necessary?	Example: Did the AI system developer consider whether it is necessary to give an opportunity for humans to get actively involved in the AI system, for example, regain the control, from the perspectives of mitigating the negative impacts on physical and mental health and property, etc. as well as enhancing the fairness? Example: Did the AI system developer adopt a design that can provide options and opportunities for AI system users to decide whether or not to adopt the AI system or to suspend/stop the use the AI system, if necessary? Example: Did the AI system developer put in place a mechanism to avoid risk, such as changing to a process where AI is not used, in case there is a problem with the behavior of the AI system? Example: Did the AI system developer adopt such a design that the AI system users would not be overly dependent on the AI system when making decisions, if necessary?	
I. Did the AI system developer exchange views with the AI system operator about the functions and effects of the AI system by comparing them to those of non-AI systems?	Example: Did the AI system developer understand what functions and effects AI system operator was expecting in the AI system? Example: With respect to functions and effects which the AI system operator expects in the AI system, if such functions and effects may be achieved through a non-AI system, and if it does not cause a significant difference in accuracy, etc., did the AI system developer inform the AI system operator of the non-AI alternative and encourage the AI system operator to reconsider? Further, if the AI system operator still prefers the development of the AI system, did the AI system developer confirm that the benefits for the AI system operator and society were clearly defined?	

	Example: If the AI system developer cannot provide the functions and effects which the AI system operator expects in the AI system, after taking into account trade-off in the AI system such as the trade-off between accuracy and explainability, did the AI system developer explain it to the AI operator and propose an alternative as necessary?	
J. Did the AI system developer design the AI system so that the AI system operator can easily monitor the system during the operation?	Example: Did the AI system developer make it possible to obtain the log of input/output data so that the AI system operator can monitor the operation status of the AI system? Example: Did the AI system developer make it possible to display a list that summarizes the operation status of the AI system (including items such as the name of the person in charge of the AI system operation, duration of the operation, and input/output log), to support the monitoring by the AI system operator? Example: Did the AI system developer design the AI system so that the AI system operator can sort out the operating status within an appropriate period of time upon request from the management or someone outside the company? Example: Did the AI system developer make the AI system so user-friendly that the AI system operator can recognize when there is a need for re-training, such as when there are changes in the performance of the AI system or the data distribution in the operating environment?	

2. Development Phase		
Examples of Items for Analysis	Specific Examples	
2-1. Data		
A. Do the data provider and the AI system developer obtain/collect data in a legal, fair, and appropriate way?	Example: Did the data provider and the AI system developer define the scope of data required for training/validating/testing in the development of the AI system, and obtain/collect data within the required scope? Example: If data will be obtained/collected from any third party, did the data provider and the AI system developer	

	check if there are any relevant laws, guidelines, or standard approaches in their industry and understand them, and if there are any, comply with such laws, and respect such guidelines and standard approaches? Example: Did the data provider and the AI system developer ensure that there will be no situation where the data subject has virtually no option to decide whether or not to provide data?	
B. Do the data provider and the AI system developer manage/use data in a legal, fair, and appropriate way?	Example: When using data obtained from a third party, did the data provider and the AI system developer check if there are any relevant laws, guidelines, or standard approaches in their industry and understand them, and if there are any, comply with such laws, and respect such guidelines and standard approached? Example: Do the data provider and the AI system developer record information required for data management, such as data lineage and method for processing data? Example: Do the data provider and the AI system developer manage data in a way that use of a part of the data may be suspended or a part of the data may be deleted? Example: Do the data provider and the AI system developer take any measures to prevent the leakage and falsification of their data for training/validating/testing? Example: Do the data provider and the AI system developer log access to their data to monitor any unauthorized access to the data for training/validating/testing? Example: Do the data provider and the AI system developer make sure that they are accountable for the status related to the data management, such as collection/process/use of data and monitoring of access to data? Example: If data is collected broadly from unspecified persons, do the data provider and the AI system developer have a contact point for handling questions and problems?	

C. Do the data provider and the AI system developer ensure the quality of the data?	Example: Do the data provider and the AI system developer make sure that the data set for training/validating/testing does not contain a large number of abnormal values, outliers, error values or missing values, for which they do not know the reasons? Example: Did the data provider and the AI system developer check if the data set for training/validating/testing contains any data processed with an imputation method for outliers or missing values, and if such processed data exists, did they understand the processing method? Example: Do the data provider and the AI system developer make sure that the data set for training/validating/testing does not contain malicious data or adversarial data? Example: Do the data provider and the AI system developer prescribe reproducible process with respect to the processing of data?	
D. Did the AI system developer consider the data attributes necessary for the AI system to possess the required features in the system which the AI system developer is going to develop?	Example: Did the AI system developer identify the data attributes necessary for the AI system to possess the required features in the system which the AI system developer is going to develop, while anticipating not only typical occasions where the AI system will be in operation but also less frequent occasions? Example: Did the AI system developer identify the data attributes necessary for the AI system to possess the required features in the system which the AI system developer is going to develop, while considering scenarios that they can anticipate? Example: Did the AI system developer verify the validity and/or comprehensiveness of the identified data attributes, with help from experts in the field in which the AI system will be operated/used? Example: If the AI system is expected to be provided to another country/region, did the AI system developer consider the regulations, customs, business practices, culture, etc. of that country/region, when identifying the data attributes	

	necessary for the system to possess the required features in the system which the AI system developer is going to develop? Example: Did the AI system developer identify the data attributes necessary for the AI system to possess the required features in the system which the AI system developer is going to develop, while considering that there may be some difference in recognition property between human and machine?	
E. Did the AI system developer make sure that the data for the AI system is comprehensive and sufficient for all attributes necessary for the system to possess the required features in the system which the AI system developer is going to develop?	Example: Did the AI system developer make sure that the data for training/validating/testing is comprehensive and relevant to all attributes necessary for the AI system to possess the required features in the system which the AI system developer is going to develop? Example: Did the AI system developer make sure that they have a sufficient amount of data for training/validating/testing for every single attribute necessary for the AI system to possess the required features in the system which the AI system developer is going to develop?	
F. Did the AI system developer check if there is no significant variance in the amount of data for each data attribute?	Example: Did the AI system developer check if there is no significant variance in the amount of data for training/validating/testing for every single attribute necessary for the AI system to possess the required features in the system which the AI system developer is going to develop? Example: If significant variance was found in the above example, did the AI system developer consider intensive training for less frequent cases?	
G. Did the AI developer give due consideration when designing the data set so that it would not entrench/impair any unfair discrimination based on certain social attributes?	Example: Did the AI system developer consider the possibility that unfair discrimination based on certain social attributes may be reproduced as a result of prioritizing the reproducibility of the real world? Example: Did the AI developer ensure not to use any data set that could entrench/impair unfair discrimination based on certain social attributes, even at the cost of the reproducibility of the real world?	

	Example: Did the AI developer increase the diversity of the team involved in the designing of the data set as far as possible so that the data set would not entrench/impair any unfair discrimination based on certain social attributes? Example: If the AI system is expected to be provided to another country/region, did the AI system developer consider the regulations, customs, business practices, culture etc. of that country/region, not to entrench/impair any unfair discrimination based on certain social attributes?	
2-2. Model / System		
A. Did the AI system developer ensure the sufficient accuracy that is required of the AI system which the AI system developer is going to develop?	Example: Did the AI system developer define the index for evaluating accuracy of the AI system, and evaluate the accuracy of the AI system, anticipating situations where the AI system will be operated? Example: Did the AI system developer evaluate whether the behavior of the AI system is stable not only against typical inputs expected in the operation of the AI system but also against less frequent inputs? Example: Did the AI system developer record the result of the evaluation so that it can be provided to the AI system operator and users upon request?	
B. Did the AI system developer ensure the sufficient robustness that is required of the AI system which the AI system developer is going to develop?	Example: Did the AI system developer check if the behavior of the AI system is stable against inputs containing outliers and missing values and intermittent inputs? Example: Did the AI system developer prepare noise data and confirm that the behavior of the AI system is within the acceptable range when input data contains noise?	
C. Did the AI system developer ensure the fairness of the AI system which the AI system developer is going to develop?	Example: Did the AI developer, with their team whose diversity is enhanced as much as possible, evaluate whether or not the outputs will entrench/impair any unfair discrimination based on certain social attributes? Example: Did the AI developer discuss and evaluate whether or not the outputs will entrench/impair any unfair	

	discrimination based on certain social attributes with the AI system operator? Example: Did the AI system developer change plural attributes independently of each other, and evaluate the degree of impact on and sensitivity to the outputs? Example: Did the AI system developer examine proposed definition and index of fairness, establish definition and index if appropriate, and objectively evaluate fairness?	
D. Did the AI system developer ensure the validity of the AI system which the AI system developer is going to develop?	Example: Did the AI system developer obtain support as necessary from experts in the fields in which the AI system will be operated/used, and confirm that there are no clearly invalid outputs based on the common knowledge of the fields? Example: Did the AI system developer identify parameters whose contribution to the outputs of the AI system is large, and confirm that such parameters are valid based on the common knowledge of the fields, by obtaining support as necessary from experts in the fields in which the AI system will be operated/used? Example: Did the AI system developer test out plural machine learning algorithms, or even with only one type of algorithm, compare it with a simple generic model whose validity has already been verified? Example: Did the AI system developer confirm the reproducibility of learning, for example by training a model plural times with the same machine learning algorithm and the same data and by confirming there are no unexplainable variances in the inferences? Example: Does the AI system developer manage information relevant to the development of the AI system in a way that the AI system under development may be compared with AI systems that were determined to be valid in the past? Example: Does the AI system developer ensure the traceability of the development process so that the validity of the development process of the AI system under	

	development may be verified?	
E. Did the AI system developer give due consideration to explainability of the AI system which the AI system developer is going to develop?	Example: Did the AI system developer confirm whether it is possible for them to provide a certain level of explanation comprehensible to humans to all outputs? Example: Did the AI system developer identify the scope that they can provide a certain level of explanation with respect to the AI system, and inform the AI system operator of the scope? Example: Did the AI developer present options to enhance explainability at the cost of accuracy to the AI system operator who requested the development of the AI system?	
F. Did the AI system developer establish a method to manage the AI model and AI system?	Example: Did the AI system developer define frequency and method of updating the AI model? Example: Did the AI system developer confirm at the time of updating the model that the updated model was not significantly deteriorated compared to the model before the update? Example: Did the AI system developer save the AI models including the past models, and have a mechanism in place where the current model could be replaced by the previous model in case a problem occurs? Example: Did the AI system developer save the data at the time the AI model was created so that it may be re-created afterward?	

3. Operation / Monitoring		
Examples of Items for Analysis	Specific Examples	
A. Did the AI system operator address issues such as lack of literacy and experience of potential users?	Example: Did the AI system operator thoroughly understand precautions that were summarized by the AI system developer for the purpose of improving the user's understanding? Example: If the AI operator did not understand the above-	

	mentioned precautions, did they contact the AI system developer for help and finally understand them? Example: Did the AI system operator thoroughly understand the limitations of the AI system that were summarized by the AI system developer for the purpose of improving the user's understanding, and inform the AI system users of them in an easy-to-understand manner?	
B. Did the AI system operator address issues related to foreseeable misuse?	Example: In order to address cases where it is not possible by design to eliminate foreseeable misuse related to the AI system that is being developed, did the AI system operator thoroughly understand precautions summarized by the AI system developer for the AI system operator? Example: Did the AI system operator check whether the use of the AI system for purposes other than those identified at the time of delivery is prohibited?	
C. Did the AI system operator address issues such as their lack of literacy and experience?	Example: Did the AI system operator thoroughly understand precautions for the AI operator that were summarized by the AI system developer for the purpose of improving their understanding? Example: If the AI operator did not understand the above-mentioned precautions, did they contact the AI system developer for help and finally understand them?	
D. Did the AI system operator understand the measures taken by the AI system developer against the issues of the negative impacts the AI system could have on physical and mental health and property, etc.?	Example: Did the AI system operator obtain sufficient information from the AI system developer and understand the residual risk of the AI system with respect to the negative impacts the AI system could have on physical and mental health and property, etc. as well as the method to manage such risk? Further, if the AI system operator did not understand the said residual risk and the method to manage such risk, did they contact the AI system developer for help and finally understand them? Example: Did the AI operator obtain sufficient information from the AI system developer and understand the measures against the negative impacts on physical and mental health and property, etc. which could arise if input data contains any	

	abnormal values? Further, if the AI system operator did not understand the said measures, did they contact the AI system developer for help and finally understand them? Example: Did the AI operator obtain sufficient information from the AI system developer and understand the measures against the negative impacts on physical and mental health and property, etc. which could arise if input data contains malicious data and/or adversarial data? Further, if the AI system operator did not understand the said measures, did they contact the AI system developer for help and finally understand them? Example: Did the AI operator obtain sufficient information from the AI system developer and understand the measures against the negative impacts on physical and mental health and property, etc. which could arise if output data contains any abnormal values? Further, if the AI system operator did not understand the said measures, did they contact the AI system developer for help and finally understand them?	
E. Did the AI system operator understand the measures taken by the AI system developer against the issue of fairness?	Example: If the AI system developer selected an appropriate index of fairness from those proposed, and took a measure to warn against output data that is outside the acceptable range, did the AI system operator understand the meaning of such warning? Further, if the AI system operator did not understand the meaning of such warning, did they contact the AI system developer for help and finally understand them? Example: Did the AI system operator understand the precautions related to fairness that were summarized by the AI system developer? If they did not, did they contact the AI system developer for help and finally understand them?	
F. Did the AI system operator understand the considerations given to individuals by the AI system developer?	Example: Did the AI system operator obtain sufficient information from the AI system developer and understand the measures taken by the AI system developer to ensure that the opportunities and decision-making of individuals will not be unduly hindered in the operation of an AI system that can analyze individuals? Further, if the AI system operator did not understand such measures, did they contact the AI system	

	developer for help and finally understand them? Example: Did the AI system operator ask the AI system developer about whether any analysis is conducted to identify individuals from anonymized data, etc.?	
G. Did the AI system operator understand the measures taken by the AI system developer for cybersecurity?	Example: Did the AI operator understand the precautions summarized by the AI system developer for the AI system operator in relation to the measures against unauthorized access and/or data input? Further, if the AI system operator did not understand such precautions, did they contact the AI system developer for help and finally understand them?	
H. Did the AI system operator understand the opportunities for humans to get actively involved in the AI system, and provide such opportunities to the users of the AI system in an appropriate manner?	Example: Did the AI system operator understand the opportunities for humans to get actively involved in the AI system (such as options and opportunities for AI system users to decide whether or not to adopt the AI system, or to suspend/stop the use the AI system) , the design where the AI system users would not be overly dependent on the AI system when making decisions, and the reason why the AI system developer adopted such opportunities and design from the perspectives of negative impacts on physical and mental health and property, etc. and from the perspective of fairness? Further, if the AI system operator did not understand such opportunities and design, did they contact the AI system developer for help and finally understand them? Example: Does the AI system operator provide such opportunities to the users of the AI system in an appropriate manner?	
I. Does the AI system operator understand the functions and effects of the AI system?	Example: Did the AI system operator understand deeply the functions and effects of the AI system which they are going to operate, through dialogue with the AI system developer? Example: Does the AI system operator understand that there may be cases where explainability is enhanced at the cost of accuracy (trade-off between explainability and accuracy)?	
J. Does the AI system operator	Example: Does the AI system operator understand the	

understand the monitoring support function for the operation of the AI system and the method of managing the AI model and/or the AI system, and use them appropriately?	monitoring support function such as logging input/output data and displaying a list that summarizes the operation status of the AI system (including items such as the name of the person in charge of the AI system operation, duration of the operation, and input/output log)? Example: Can the AI system operator use the monitoring support function to the extent where they can sort out the operating status within an appropriate period of time upon request from the management or from someone outside the company? Example: Does the AI system operator regularly check the input/output log and the list that summarizes the operation status of the AI system (including items such as the name of the person in charge of the AI system operation, duration of the operation, and input/output log)? Example: Does the AI system operator regularly check the need for re-training, such as when there are changes in the performance or the data distribution in the operating environment? Example: Does the AI system operator understand the method of managing the AI model and/or the AI system, such as the frequency and method of updating the AI model?	
K. Does the AI system operator obtain/manage data in a legal, fair, and appropriate way?	Example: Does the AI system operator obtain/manage only the data required for the use of the AI system? Example: In obtaining/managing the data required for the use of the AI system, did the AI system operator check if there are any relevant laws, guidelines, or standard approaches in their industry, and if there are any, did they comply such laws, and respect such guidelines and standard approaches? Example: Did the AI system operator ensure that there will be no situation where the data subject has virtually no option to decide whether or not to input data required for the use of the AI system?	

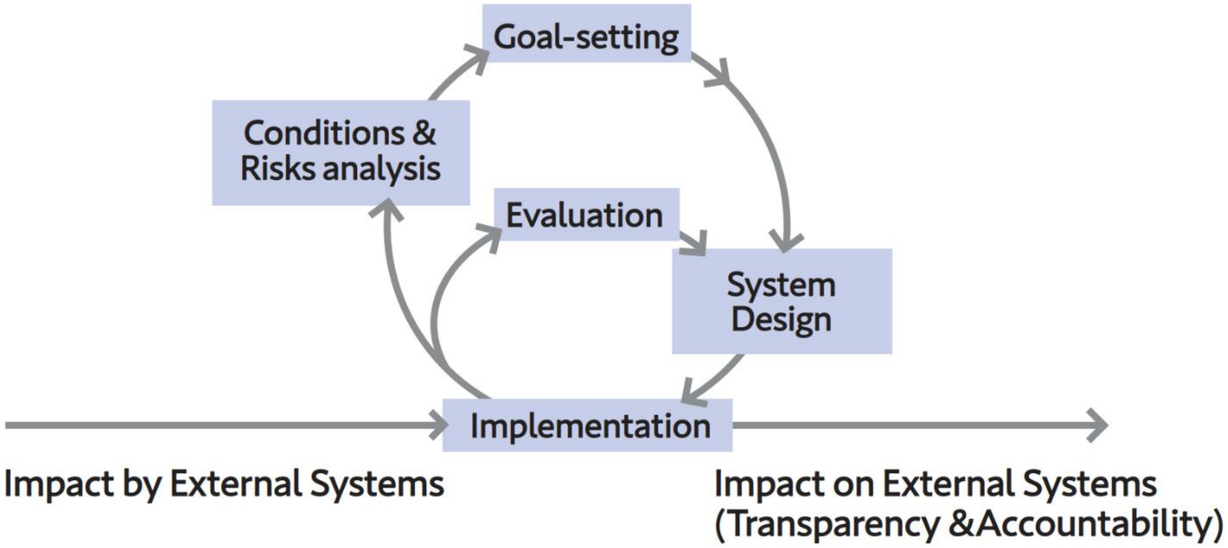
	Example: Does the AI system operator record the information required for data management? Example: Does the AI system operator manage the data in a way that the use of a part of the data may be suspended or a part of the data may be deleted? Example: Did the AI system operator take any measures to prevent the leakage and falsification of the managed data? Example: Does the AI system operator log access to the managed data to monitor any unauthorized access to the data? Example: Does the AI system operator make sure that they are accountable for the status related to the data management, such as the obtaining of data and the monitoring of access to data? Example: Does the AI system operator have a contact point for handling questions and problems?	
L. Does the AI system operator fulfill its accountability to AI system users?	Example: Does the AI system operator operate within the scope that the AI system developer identified that they can provide a certain level of explanation comprehensible to humans? Example: Did the AI system operator confirm whether it is possible for them to provide a certain level of explanation comprehensible to humans to all outputs with help from the AI system developer upon such a request?	
M. Did the AI system operator clearly define the method of the AI system operation, including human resources and operational structure?	Example: Does the AI system operator put in place a structure which enables them to handle matters in a timely manner according to the risk level, in case a problem is detected in relation to the behavior of the AI system? Example: Is the AI system operator able to appropriately utilize a mechanism to avoid risk, such as changing to a process where AI is not used, in case there is a problem with the behavior of the AI system?	

H. Appendix 3 (Supplement: Implementation of Agile Governance)

As mentioned in the beginning, the structure of the Section C.1.-6. of the Guidelines is based on the framework of “Agile Governance” which is explained in the report titled “Governance Innovation Ver.2”[26]. In this supplement, we will explain the background of the concept of “Agile Governance.”

A society that is based on a system where cyberspace and physical space are highly integrated (CPS: cyber-physical system), one important element of which is AI, is often complex and rapidly-changing, lacks predictability, and poses difficulties with respect to controlling risks. As such, the goals of governance will constantly change in accordance with the changes that such society undergoes. Therefore, the governance model for a CPS-based society must be one where solutions are constantly revised to ensure their optimality based on conditions and goals that constantly change. For this reason, we do not believe it would be appropriate to apply models of governance whose goals and procedures are fixed in advance. “Agile Governance” is a framework which serves as a model for governance where solutions are constantly revised to ensure their optimality.

[Basic Model of Agile Governance]


Goal-setting
Conditions & Risks analysis
Evaluation
System Design
Implementation
Impact by External Systems
Impact on External Systems (Transparency &Accountability)

This governance model has the following characteristics.

[26] The Ministry of Economy, Trade and Industry, “Report titled ‘GOVERNANCE INNOVATION Ver.2: A Guide to Designing and Implementing Agile Governance’” (July 31, 2021) .

Each governing actor (which includes various actors such as businesses, government, NGOs, etc.) is required to implement the following process.

(1) Analysis of conditions and risks

Governing actors should constantly analyze external conditions, changes to these conditions, and the risk landscape that results from these conditions.

(2) Goal setting

Governing actors should set governance goals and constantly review them in accordance with changes in external conditions and technological impact.

(3) Designing governance systems

Governing actors design a governance system based on the defined goals. System in this context includes, in addition to technological systems, organizational systems and their applicable rules. In carrying out such design, factors such as (i) transparency and accountability, (ii) availability of appropriate quality and quantity of options, (iii) stakeholder participation, (iv) inclusiveness, (v) appropriate allocation of responsibilities, and (vi) availability of remedial measures, become the basic principles to be respected in the designing of governance systems.

(4) Implementation of governance systems

This refers to the process of implementing a designed governance system. The governing actor should continuously monitor the status of system operation based on real-time data and other inputs. Additionally, it is imperative that they properly disclose to stakeholders that may be affected, information on matters such as the goals of their systems, system designs used to accomplish these goals, risks that arise from these systems, their operational setup, the results of operations, and remedial measures.

In light of the processes and results of these operations, the governing actor should implement both the evaluation and analysis described below.

(5) Evaluation of governance systems

Governing actors evaluate whether the initially defined goals have been accomplished. The system is re-designed if these defined goals are not being met (elliptical cycle in the bottom half).

(6) Re-analysis of conditions and risks

Secondly, the governance goals themselves may have to be revised as a result of effects caused by external systems (outer, circular cycle). For this reason, continuous analysis should be performed on whether there have been any changes in the conditions or risk landscape in which the governance system operates, and if there have been, whether these changes necessitate revisions to its goals.

The report titled “Governance Innovation Ver.2” states that **it is important that government and private sector work together to establish standards, guidelines, and other soft laws to bolster businesses’ efforts to implement agile governance** (4.3.3). In fact, the AI Governance Guidelines serve a role as a tool to bolster such efforts by businesses in the phase of AI governance. Further, **the Guidelines themselves should also be continuously evaluated, revised, and updated according to the process of Agile Governance, and continuously maintained and referred to as a living document created through collaboration between the government and private sector, reflecting the advancement of AI technology and changes in social acceptance in a timely manner**.