Tentative
Translation

Draft AI Utilization Principles

17 July 2018

The Conference toward AI Network Society

Overview

- The acceleration of the development and utilization of AI needs efforts to promote the benefits and mitigate the risks of AI systems, as well as to gain the trust of users and the society in AI.

- Due to past efforts, the Conference toward AI Network Society announced “Draft AI R&D Guidelines for International Discussions” (hereinafter referred to as “Draft AI R&D Guidelines”) regarding matters expected to be considered in R&D activities.

- On the other hand, AI systems’ outputs or programs might continuously change as a result of learning or other methods in the process of actual use; therefore, there are not only matters that the developers are expected to consider, but there are other matters that users are expected to consider.

- With the above consideration, the Conference focused on the matters that users are expected to consider, on the basis of the scenario analysis (use case assessment) and ecosystem prospects formed with the progress of AI networking.

→ In the 2018 Report, the Conference proposes “Draft AI Utilization Principles” and compiles the points to discuss concerning the content of each principle. The Conference will continue its best figure out the final output.

Classification of Related Entities (1/7)

In the utilization of AI, as various entities are assumed to be involved, the Conference classifies the related entities based on the following policies.

■ "Draft AI R&D Guidelines" defines that "the term 'users' refers to those who use AI systems, including end users as well as providers who provide third parties with AI-network services developed by others." This means that users include people in various positions.
In addition, there may be cases where people do not directly use AI systems or AI services (hereinafter referred to simply as "AI") but use services which are based on the judgments of AI systems, i.e., indirectly use AI.

■ The Conference proposes that the term "AI services" is defined as services providing functions of AI systems and the term "AI-accompanying services" is defined as AI-systems-update services or additional learning services, etc. On the basis of the fact that developers often carry out updating of AI systems or additional learning, etc., developers are also involved in the utilization of AI.
Note: Definitions of terms such as AI, AI software, and AI systems in "Draft AI Utilization Principles" follow the definitions thereof in "Draft AI R&D Guidelines."

■ With consideration of the above, in order to study matters that users are expected to consider in the actual use of AI, it is appropriate to classify users into several types and consider the matters according to the user types. The Conference attempts to classify users objectively as described on the next page after observing how they are involved in the provision and utilization of AI.

■ The Conference will deepen the discussion on the classification of users while taking into account the following opinions from the Conference members:
- It may be necessary to distinguish between users who are aware of using AI and those who do not recognize it;
- It may be necessary to distinguish between users who own and manage AI systems or adjust the inputs and outputs of AI systems, and those who are not able to do so; and
- It may be necessary to keep in mind with regard to those who are socially vulnerable, such as children, elderly people, and people with disabilities.

Classification of Related Entities (2/7)

Users

Those who use AI systems, AI services, or AI-accompanying services.

Indirect users

Those who receive services provided by end users using AI systems or AI services.

Data providers

Those who provide data for learning or other methods of AI systems used by others.

Third parties

Those whose rights and interests are concomitantly affected due to AI used by others.

Developers

Those who conduct the R&D of AI systems

AI service providers

Users who provide AI services or AI-accompanying services on business basis to others.

- **AI-network-service providers**

  AI service providers who provide AI services or AI-accompanying services on business basis to others through information-and-communication networks.

- **Offline AI service providers**

  AI service providers other than the above AI-network-service providers.

End users

Users who use AI systems or AI services without providing AI services or AI-accompanying services on business basis to others.

- **Business users**

  End users who use AI systems or AI services on business basis.

- **Consumer users**

  End users who use AI systems or AI services (excluding business users)

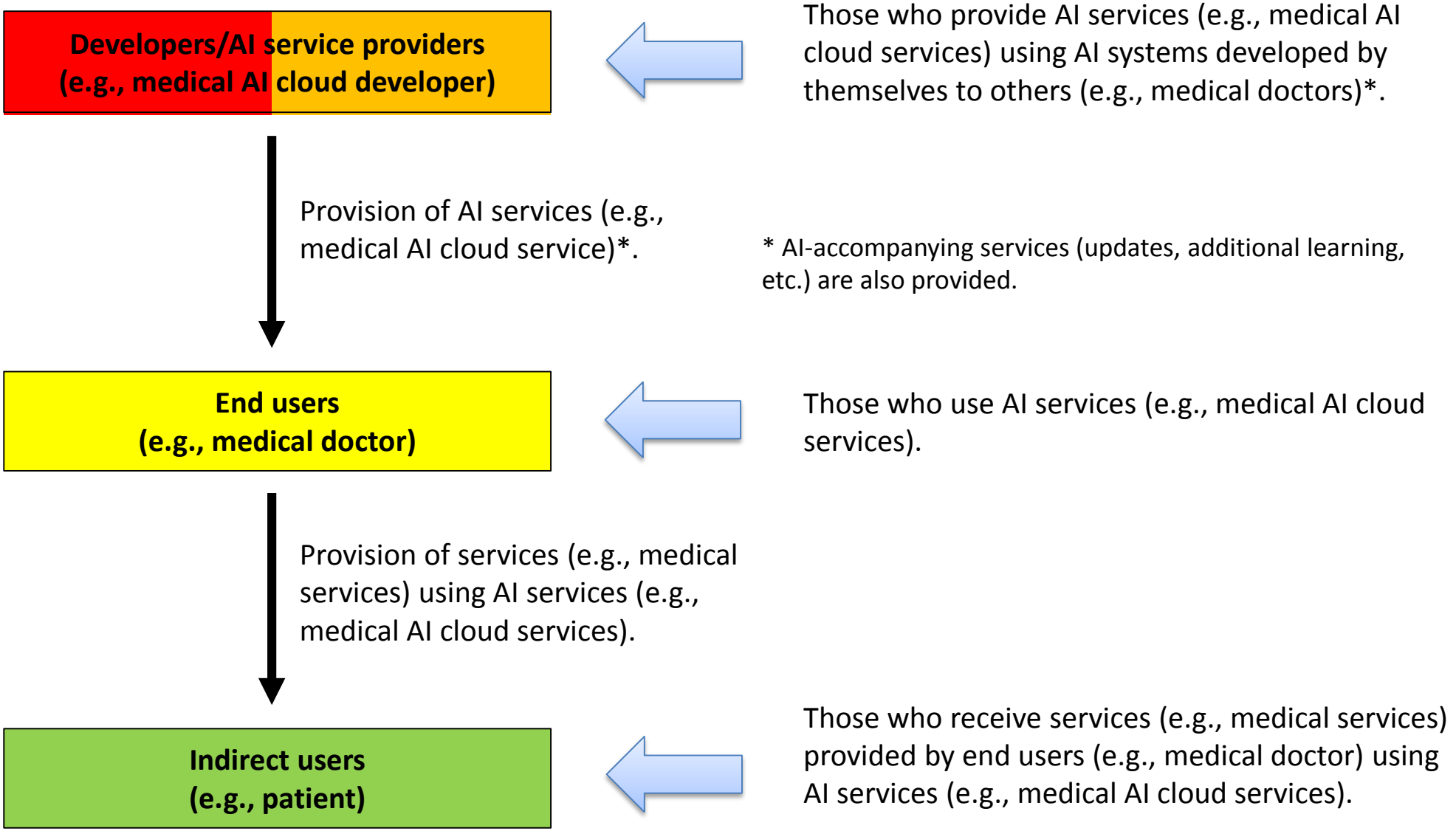
(Note) Each name is tentative.

(Note) One individual or enterprise may be included in multiple entities.

- AI systems: Systems that incorporate AI software as a component.
- AI services: Services that provide the functions of AI systems
- AI-accompanying services: AI-systems-update services or additional learning services, etc.

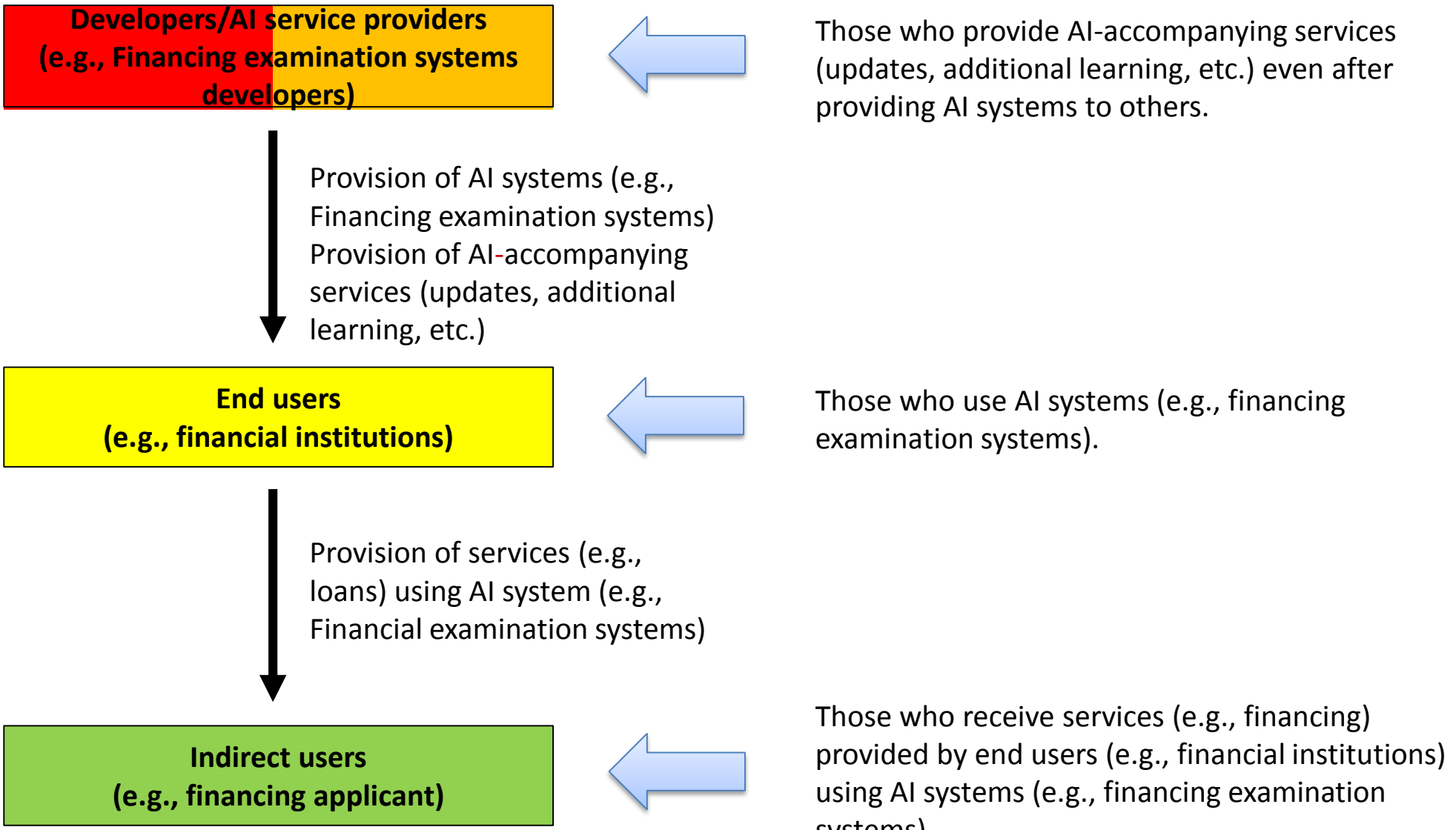
Classification of Related Entities (3/7)

Example (1) <Medical AI cloud>



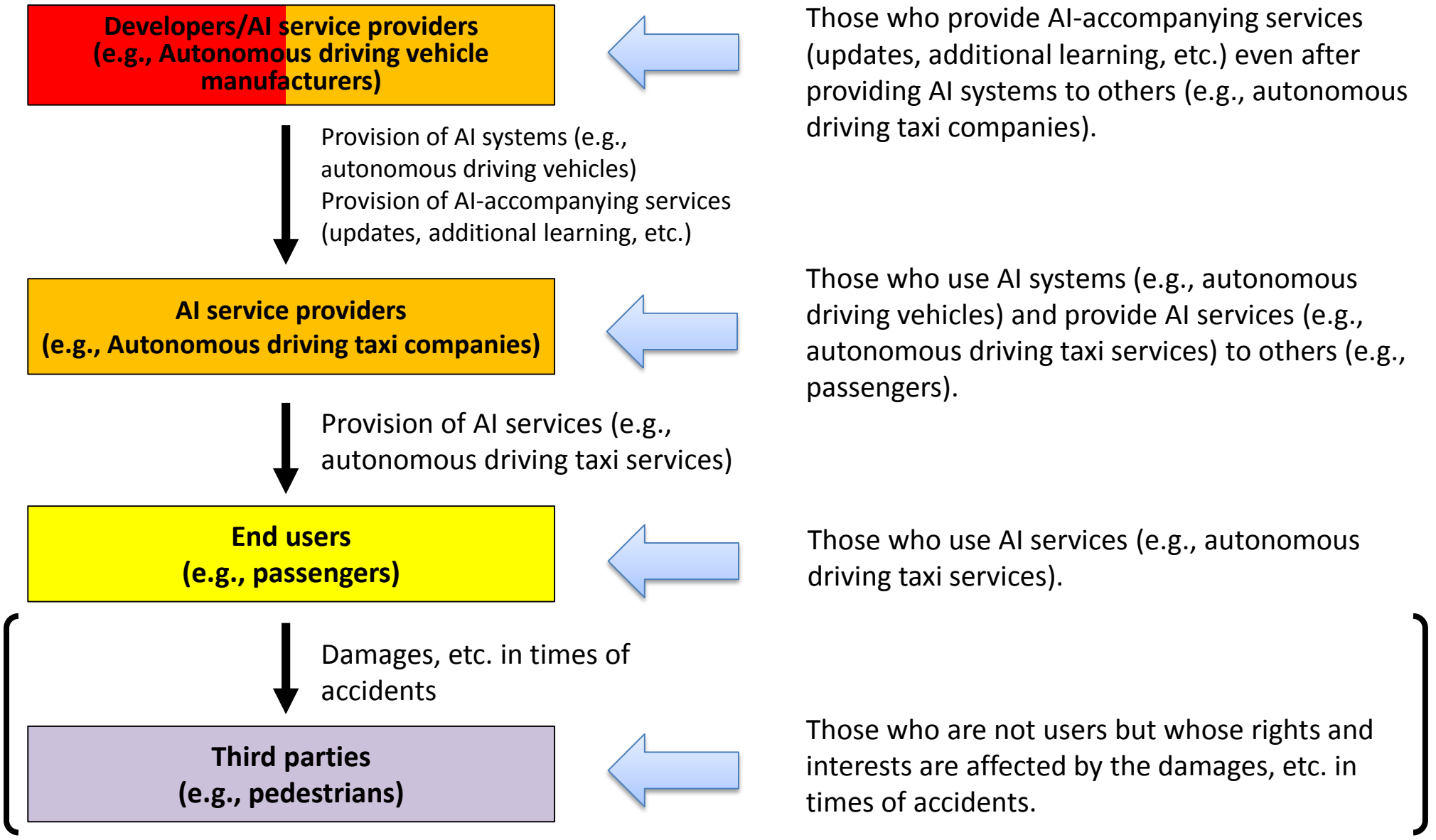
Classification of Related Entities (4/7)

Example (2) <Financing Examination>



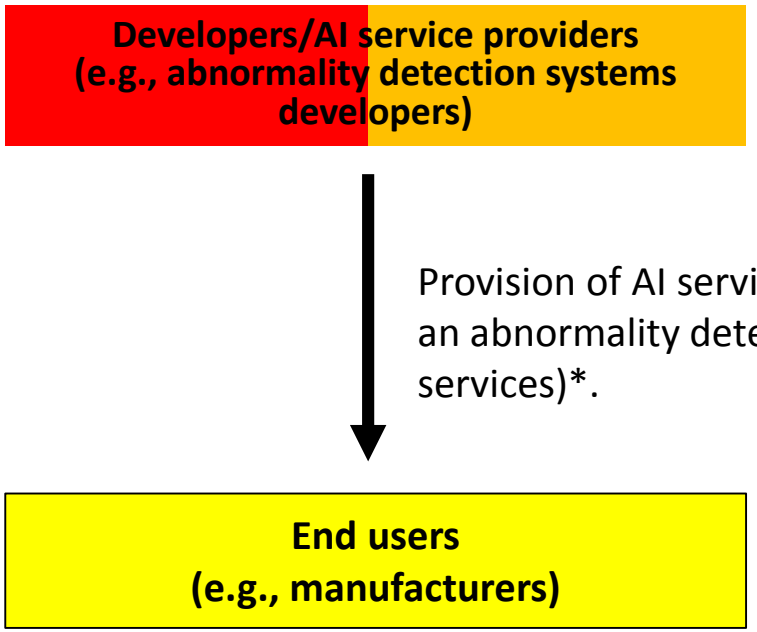
Classification of Related Entities (5/7)

Example (3) <Autonomous driving taxi>



Classification of Related Entities (6/7)

Example (4) <Detection of abnormality in production lines>



Those who provide AI services (e.g., an abnormality detection services) using AI systems developed by themselves to others (e.g., manufactures)*.

* AI-accompanying services (updates, additional learning, etc.) are also provided.

Those who use AI services (e.g., an abnormality detection services).

Classification of Related Entities (7/7)

Example (5) <AI assistant (AI speaker)>

Developers/AI service providers (e.g., AI assistant developers)

Those who provide AI services (e.g., an AI assistant services) using AI systems developed by themselves to others (e.g., individuals using AI-assistants-compliant-AI speakers)*.

Provisions of AI services (e.g., AI assistant services)*.

* AI-accompanying services (updates, additional learning, etc.) are also provided.

End users (e.g., individuals using AI-assistants-compliant-AI speakers)

Those who use AI services (e.g., an AI assistant services).

Impact on privacy, etc.

Third parties (E.g., family members living together with end users, friends visiting them, etc.)

Those who are not users but whose rights and interests such as privacy are affected.

Draft AI Utilization Principles (1/2)

- The Conference assessed use cases and ecosystem prospects of AI services and identified the benefits and the risks of AI systems. To promote the benefits from AI systems and to mitigate the risks as well as to foster the trust in AI systems, the Conference proposes “Draft AI Utilization Principles” which will be expected to be considered in the utilization of AI.
  - ⇒ Services utilizing AI systems will be provided beyond national borders via information-and-communication networks; therefore, as with “Draft AI R&D Guidelines,” “Draft AI Utilization Principles” are prepared for international discussions and are proposed as non-regulatory and non-binding principles, so-called “Soft Law.”
- The Conference also compiled the points to discuss concerning the content of each principle. The Conference will deepen the discussion on the points and clarify the details of the principles depending on the classification of “Users” (e.g., AI service providers and End users).

1) Principle of proper utilization
2) Principle of data quality
3) Principle of collaboration
4) Principle of safety
5) Principle of security
6) Principle of privacy
7) Principles of human dignity and individual autonomy
8) Principle of fairness
9) Principle of transparency
10) Principle of accountability

... Mainly related to **promoting benefits**

... Mainly related to **mitigating risks**

... Mainly related to **fostering trust**

Draft AI Utilization Principles (2/2)

1) **Principle of proper utilization**
   Users should make efforts to utilize AI systems or AI services in a proper scope and manner, under the proper assignment of roles between humans and AI systems, or among users.
2) **Principle of data quality**
   Users and data providers should pay attention to the quality of data used for learning or other methods of AI systems.
3) **Principle of collaboration**
   AI service providers, business users, and data providers should pay attention to the collaboration of AI systems or AI services.
   Users should take it into consideration that risks might occur and even be amplified when AI systems are to be networked.
4) **Principle of safety**
   Users should take into consideration that AI systems or AI services in use will not harm the life, body, or property of users, indirect users or third parties through the actuators or other devices.
5) **Principle of security**
   Users and data providers should pay attention to the security of AI systems or AI services.
6) **Principle of privacy**
   Users and data providers should take into consideration that the utilization of AI systems or AI services will not infringe on the privacy of users' or others.
7) **Principles of human dignity and individual autonomy**
   Users should respect human dignity and individual autonomy in the utilization of AI systems or AI services.
8) **Principle of fairness**
   AI service providers, business users, and data providers should take into consideration that individuals will not be discriminated unfairly by the judgments of AI systems or AI services.
9) **Principle of transparency**
   AI service providers and business users should pay attention to the verifiability of inputs/outputs of AI systems or AI services and the explainability of their judgments.
10) **Principle of accountability**
    AI service providers and business users should make efforts to fulfill their accountability to the stakeholders including consumer users and indirect users.

Points of the Content of Each Principle (1/11)

1) Principle of proper utilization

Users should make efforts to utilize AI systems or AI services in a proper scope and manner, under the proper assignment of roles between humans and AI systems, or among users.

[Main points to discuss]

A) Utilization in the proper scope and manner

On the basis of the provision of information and explanation from developers, etc. and with consideration of social contexts and circumstances, users may be expected to use AI in the proper scope and manner. In addition, users may be expected to recognize benefits and risks, understand proper uses, acquire necessary knowledge and skills and so on before using AI, according to the characteristics, usage situations, etc. of AI. Furthermore, users may be expected to check regularly whether they use AI in an appropriate scope and manner.

B) Proper balance of benefits and risks of AI

AI service providers and business users may be expected to take into consideration proper balance between benefits and risks of AI, including the consideration of the active use of AI for productivity and work efficiency improvements, after appropriately assessing risks of AI.

C) Updates of AI software and inspections/repairs, etc. of AI

Through the process of utilization, users may be expected to make efforts to update AI software and perform inspections, repairs, etc. of AI in order to improve the function of AI and to mitigate risks.

Points of the Content of Each Principle (2/11)

D) Human Intervention

Regarding the judgment made by AI, in cases where it is necessary and possible (e.g., medical care using AI), humans may be expected to make decisions as to whether to use the judgments of AI, how to use it etc. In those cases, what can be considered as criteria for the necessity of human intervention?

In the utilization of AI that operates through actuators, etc., in the case where it is planned to shift to human operation under certain conditions, what kind of matters are expected to be paid attention to?

[Points of view as criteria (example)]

- The nature of the rights and interests of indirect users, et al., and their intents, affected by the judgments of AI.
- The degree of reliability of the judgment of AI (compared with reliability of human judgment).
- Allowable time necessary for human judgment
- Ability expected to be possessed by users

E) Role assignments among users

With consideration of the volume of capabilities and knowledge on AI that each user is expected to have and ease of implementing necessary measures, users may be expected to play such roles as seems to be appropriate and also to bear the responsibility.

F) Cooperation among stakeholders

Users and data providers may be expected to cooperate with stakeholders and to work on preventive or remedial measures (including information sharing, stopping and restoration of AI, elucidation of causes, measures to prevent recurrence, etc.) in accordance with the nature, conditions, etc. of damages caused by accidents, security breaches, privacy infringement, etc. that may occur in the future or have occurred through the use of AI.

What is expected reasonable from a users point of view to ensure the above effectiveness?

Points of the Content of Each Principle (3/11)

2) Principle of data quality

Users and data providers should pay attention to the quality of data used for learning or other methods of AI systems.

[Main points to discuss]

A) Attention to the quality of data used for learning or other methods of AI

Users and data providers may be expected to pay attention to the quality of data (e.g., the accuracy and completeness of data) used for learning or other methods of AI, with consideration of the characteristics of AI to be used and its usage. If the accuracy of the judgment of AI is impaired or declines, it may be expected to let AI systems learn again with paying attention to the quality of data.

In what cases and to what extent are users and data providers expected to pay attention to the quality of data used for learning or other methods?

B) Attention to security vulnerabilities of AI by learning inaccurate or inappropriate data

Users and data providers may be expected to pay attention to the risk that the security of AI might become vulnerable by learning inaccurate or inappropriate data.

Points of the Content of Each Principle (4/11)

3) Principle of collaboration

AI service providers, business users, and data providers should pay attention to the collaboration of AI systems or AI services. Users should take it into consideration that risks might occur and even be amplified when AI systems are to be networked.

[Main points to discuss]

A) Attention to the interconnectivity and interoperability of AI systems
AI-network-service providers may be expected to pay attention to the interconnectivity and interoperability of AI with consideration of the characteristics of AI to be used and its usage, in order to promote the benefits of AI through the sound progress of AI networking.

B) Address to the standardization of data formats, protocols, etc.
AI service providers and business users may be expected to address the standardization of data formats, protocols, etc. in order to promote cooperation among AI systems and between AI systems and other systems, etc. Also, data providers may be expected to address the standardization of data formats.

C) Attention to problems caused and amplified by AI networking
Although it is expected that collaboration of AI promotes the benefits, users may be expected to pay attention to the possibility that risks (e.g. the risk of loss of control by interconnecting or collaborating their AI systems with other AI systems, etc. through the Internet or other network) might be caused or amplified by AI networking.
[Problems (examples) over risks that might become realized and amplified by AI networking]
- Risks that one AI system's trouble, etc. spreads to the entire system.
- Risks of failures in the cooperation and adjustment between AI systems.
- Risks of failures in verifying the judgment and the decision making of AI (risks of failure to analyze the interactions between AI systems because the interactions become complicated).
- Risks that the influence of a small number of AI becomes too strong (risks of enterprises and individuals suffering disadvantage by the judgment of a few AI systems).
- Risks of the infringement of privacy as a result of information sharing across fields and the concentration of information to one specific AI.
- Risks of unexpected actions of AI.

Points of the Content of Each Principle (5/11)

4) Principle of safety

Users should take into consideration that AI systems or AI services in use will not harm the life, body, or property of users, indirect users or third parties through the actuators or other devices.

[Main points to discuss]

A) Consideration for the life, body, or property

In the case of using AI in fields where AI systems might harm the life, body, or property, such as the fields of medical care and autonomous driving, with consideration of the nature, conditions, etc. of assumed damages, users may be expected to take into consideration that AI will not harm the life, body, or property through the actuators or other devices, by inspecting and repairing AI, updating AI software, etc. as necessary.

In addition, users may be expected to consider in advance measures to be taken, in case AI might harm the life, body, or property through the actuators or other devices.

Points of the Content of Each Principle (6/11)

5) Principle of Security

Users and data providers should pay attention to the security of AI systems or AI services.

[Main points to discuss]

A) Implementation of security measures

Users may be expected to pay attention to the security of AI and take reasonable measures in light of the technology level at that time.

In addition, users may be expected to consider measures to be taken against security breaches of AI in advance.

B) Service provision, etc. for security measures

AI service providers may be expected, with regard to their AI services, to provide services for security measures to end users and share incident information with end users.

C) Attention to security vulnerabilities of AI by learning inaccurate or inappropriate data

Users and data providers may be expected to pay attention to the risk that AI's security might become vulnerable by learning inaccurate or inappropriate data. [as referred to in *supra* 2) Principle of data quality—Main point B)]

Points of the Content of Each Principle (7/11)

6) Principle of privacy

Users and data providers should take into consideration that the utilization of AI systems or AI services will not infringe on the privacy of users' or others.

[Main points to discuss]

A) Respect for the privacy of others
With consideration of social contexts and reasonable expectations of people in the utilization of AI, users may be expected to respect the privacy of others in the utilization of AI.
In addition, users may be expected to consider measures to be taken against privacy infringement caused by AI in advance.

B) Respect for the privacy of others in the collection, analysis, provision, etc. of personal data
Users and data providers may be expected to respect the privacy of others in the collection, analysis, provision, etc. of personal data used for learning or other methods of AI.

C) Consideration for the privacy, etc. of the subject of profiling which uses AI
In the case of profiling by using AI in fields where the judgments of AI might have significant influences on individual rights and interests, such as the fields of personnel evaluation, recruitment, and financing, AI service providers and business users may be expected to pay due consideration to the privacy, etc. of the subject of profiling.

D) Attention to the infringement of the privacy of users' or others
Consumer users may be expected to pay attention not to give information that is highly confidential (including information on others as well as information on users' themselves) to AI carelessly, by excessively empathizing with AI such as pet robots, or by other causes.

E) Prevention of personal data leakage
AI service providers, business users, and data providers may be expected to take appropriate measures so that personal data should not be provided by the judgments of AI to third parties without consent of the person.

Points of the Content of Each Principle (8/11)

7) Principles of human dignity and individual autonomy

Users should respect human dignity and individual autonomy in the utilization of AI systems or AI services.

[Main points to discuss]

A) Respect for human dignity and individual autonomy
With consideration of social contexts in the utilization of AI, users may be expected to respect human dignity and individual autonomy.

B) Attention to the manipulation of human decision making, emotions, etc. by AI
Users may be expected to pay attention to the risks of the manipulation of human decision making and emotions by AI and risks of excessive dependence on AI.
It is crucial to consider who takes what measures against such risks.

C) Reference to the discussion of bioethics, etc. in the case of linking AI systems with the human brain and body
When linking AI with the human brain and body, users may be required to particularly take into consideration that human dignity and individual autonomy will not be violated, in light of discussions on bioethics, etc.

Points of the Content of Each Principle (9/11)

8) Principle of fairness

AI service providers, business users, and data providers should take into consideration that individuals will not be discriminated unfairly by the judgment of AI systems or AI services.

[Main points to discuss]

A) Attention to the representativeness of data used for learning or other methods of AI

AI service providers, business users, and data providers may be expected to pay attention to the representativeness of data used for learning or other methods of AI and the social bias inherent in the data so that individuals should not be unfairly discriminated against due to their race, religion, gender, etc. as a result of the judgment of AI.

In light of the characteristics of the technologies to be used and their usage, in what cases and to what extent is attention expected to be paid to the representativeness of data used for learning or other methods and the social bias inherent in the data?

Note: The representativeness of data refers to the fact that data sampled and used do not distort the propensity of the population of data.

B) Attention to unfair discrimination by algorithm

AI service providers and business users may be expected to pay attention to the possibility that individuals may be unfairly discriminated against due to their race, religion, gender, etc. by the algorithm of AI.

C) Human intervention

Regarding the judgment made by AI, AI service providers and business users may be expected to make judgments as to whether to use the judgments of AI, how to use it, or other matters, with consideration of social contexts and reasonable expectations of people in the utilization of AI, so that individuals should not be unfairly discriminated against due to their race, religion, gender, etc.

In light of the characteristics of the technologies to be used and their usage, in what cases and to what extent is human intervention expected?

Points of the Content of Each Principle (10/11)

9) Principle of transparency

AI service providers and business users should pay attention to the verifiability of inputs/outputs of AI systems or AI services and the explainability of their judgments.

Note: This principle is not intended to ask for the disclosure of algorithm, source code, or learning data. In interpreting this principle, privacy of individuals and trade secrets of enterprises are also taken into account.

[Main points to discuss]

A) Recording and preserving the inputs/outputs of AI

In order to ensure the verifiability of the input and output of AI, AI service providers and business users may be expected to record and preserve the inputs and outputs.

In light of the characteristics of the technologies to be used and their usage, in what cases and to what extent are the inputs and outputs expected to be recorded and preserved? For example, in the case of using AI in fields where AI systems might harm the life, body, or property, such as the field of autonomous driving, the inputs and outputs of AI may be expected to be recorded and preserved to the extent whch is necessary for investigating the causes of accidents and preventing the recurrence of such accidents.

B) Ensuring explainability

AI service providers and business users may be expected to ensure explainability on the judgments of AI.

In light of the characteristics of the technologies to be used and their usage, in what cases and to what extent is explainability expected to be ensured? Especially in the case of using AI in fields where the judgments of AI might have significant influences on individual rights and interests, such as the fields of medical care, personnel evaluation and recruitment and financing, explainability on the judgments of AI may be expected to be ensured. (For example, we have to pay attention to the current situation where deep learning has high prediction accuracy, but it is difficult to explain its judgment.)

Points of the Content of Each Principle (11/11)

10) Principle of accountability

AI service providers and business users should make efforts to fulfill their accountability to the stakeholders including consumer users and indirect users.

[Main points to discuss]

A) Efforts to fulfill accountability

In light of the characteristics of AI to be used and its purpose, etc., AI service providers and business users may be expected to make efforts to establish appropriate accountability to consumer users, indirect users, and third parties affected by the use of AI, to gain enough trust in AI from people and society.

B) Notification and publication of usage policy on AI systems or AI services

AI service providers and business users may be expected to notify or announce the usage policy on AI (the fact that they provide AI services, the scope and manner of proper AI utilization, the risks associated with the utilization, and the establishment of a consultation desk) in order to enable consumer users and indirect users to recognize properly the usage of AI.

In light of the characteristics of the technologies to be used and their usage, we have to focus on which cases will lead to the usage policy is expected to be notified or announced as well as what content is expected to be included in the usage policy.