Tentative
Translation

Overview of 2019 Report

(incl. AI Utilization Guidelines)

August 9, 2019

The Conference toward AI Network Society

Structure of 2019 Report

“2019 Report” of the conference toward AI Network Society

- Preface
- Chapter 1: Recent Trends in AI Networking
  - 1. Domestic trends (Trends in Japan)
  - 2. Overseas trends (Trends in the world)
  - 3. Trends in international discussions
- Chapter 2: Concept of the formulation of AI Utilization Guidelines
  - 1. Background and history
  - 2. Positioning of AI Utilization Guidelines
  - 3. Overview of AI Utilization Guidelines
  - 4. Issues to be studied in the future
- Chapter 3: Future Challenges
- Conclusion

<Attachment 1> AI Utilization Guidelines

- Purpose and basic philosophies
- Classification of Related Entities
- AI Utilization Principles
- General flow of AI utilization
- Commentary on the AI Utilization Principles
- Timing to Consider the AI Utilization Principles

<Attachment 2> Comparison of AI guidelines

Recent Trends in AI Networking

Domestic trends (Trends in Japan)

➢ **Announcement of the "Social Principles of Human-Centric AI" (March 29, 2019)**

The government established the "Council for Social Principles of Human-centric AI," under the AI Strategy Expert Meeting for Strength and Promotion of the Innovation, for the purpose of formulating the basic principles for implementing and sharing AI in a better way and for usage in international discussions, and so released the "Social Principles of Human-Centric AI." The social principles consist 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.

Overseas trends (Trends in the world)

➢ **The Institute of Electrical and Electronics Engineers (IEEE) announced "Ethically Aligned Design, 1st edition" (March 25, 2019)**

The "Ethically Aligned Design" describes that the ethical and values-based design, development, and implementation of autonomous and intelligent systems(A/IS) should be guided by the following General Principles: (1) Human Rights; (2) Well-being; (3) Data Agency; (4) Effectiveness; (5) Transparency; (6) Accountability; (7) Awareness of Misuse; and (8) Competence. In addition, Ethically Aligned Design focuses on "From principle to practice" that is, IEEE launched projects of the IEEE P7000™ series of standards that explicitly focus on societal and ethical issues associated with a certain field of technology, and created an A/IS Ethics Glossary.

➢ **HLEG on AI set up by European Commission announced "Ethics Guidelines for Trustworthy AI" (April 8, 2019)**

The High-Level Expert Group (HLEG) on AI set up by the European Commission presented their "Ethics guidelines for trustworthy artificial intelligence". According to the guidelines, trustworthy AI should be lawful, ethical and robust, and it identifies four principles based on fundamental rights (respect for human autonomy, prevention of harm, fairness and explicability) and seven requirements for trustworthy AI: (1) human agency and oversight; (2) technical robustness and safety; (3) privacy and data governance; (4) transparency; (5) diversity, non-discrimination and fairness; (6) environmental and societal well-being; and (7) accountability. In addition, it has provided an assessment list aimed at operationalising the requirements.

Trends in international discussions

➢ **G7 Multistakeholder Conference on Artificial Intelligence（Canada, December 6, 2018）**

Several key AI experts and G7 focal point organizations collaboratively drafted discussion papers for each topic: (1) AI for Society; (2) Unleashing Innovation; (3) Accountability in AI; and (4) the Future of Work, and directed breakout sessions during the multistakeholder conference. Japan, together with Canada, was in charge of (3) Accountability in AI.

➢ **OECD and partner countries adopted the "Recommendation of the Council on Artificial Intelligence" (May 22, 2019)**

The OECD's 36 member countries, along with six partner countries, signed up to the OECD Principles on Artificial intelligence at the Organisation's annual Ministerial Council Meeting. The recommendations, which have been created by the AI Expert Group at the OECD (AIGO) that has met four times since September 2018, identify five principles for responsible stewardship of trustworthy AI, namely: (1) inclusive and sustainable growth and well-being; (2) human-centred values and fairness; (3) transparency and explainability; (4) robustness and safety; and (5) accountability. They also include recommended national policy priorities for trustworthy AI. The recommendation consists of the high level principles and specific measures to be taken will be considered continuously at the CDEP meeting after the formulation of the recommendation.

➢ **G20 Ibaraki-Tsukuba Ministerial Meeting on Trade and Digital Economy (June 8 and 9, 2019)**

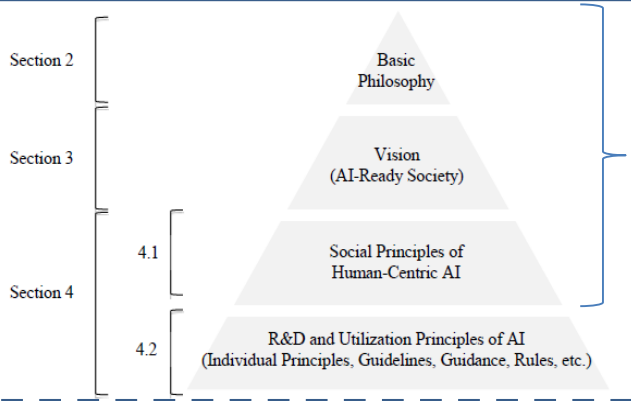
In order to foster the development and utilization of AI, the G20 adopted their first ministerial statement which includes AI principles ("G20 AI principles") based on the "human-centered" idea. The principle is drawn from the "OECD Recommendation of the Council on Artificial Intelligence" and agreed as an annex to the Ministerial Statement.

(Note) Main items are listed.

Positioning of “AI Utilization Guidelines”

Reference when private sectors, etc. formulate their own principles

Structure of “Social Principles of Human-Centric AI [1]”



・・・ Basic principles to which society (especially state legislative and administrative bodies) should pay attention

・・・ Developers and business operators of AI should establish and comply with the AI development and utilization principles

Each developer or business operator is encouraged to establish AI development and utilization principles.

Practical guidance to be referred to is needed.

Contribution to the international discussions

Consensus on AI principles is being reached. Hereafter, the focus of the discussion will be on **how to implement the AI principles**. Japan will continue to contribute to the international discussions, and foster sharing the recognitions.

(Examples of the discussion on how to implement AI principles)

[European Commission]

- **Assessment list** in the “Ethics Guidelines for Trustworthy AI”
- A revised version of the list, taking into account the feedback gathered through the piloting phase, will be presented to the European Commission in early 2020.

[OECD]

- Practical measures to realize “Recommendation of the Council on Artificial Intelligence” (=**Practical Guidance**)
- Discussion at OECD/CDEP will start from summer of 2019.

[1] https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf

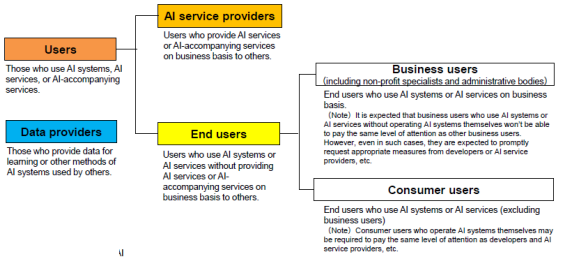
Structure of “AI Utilization Guidelines”

Part 1: Perspective of AI Utilization Principles

1. Purpose

2. Basic philosophies

3. Classification of related entities



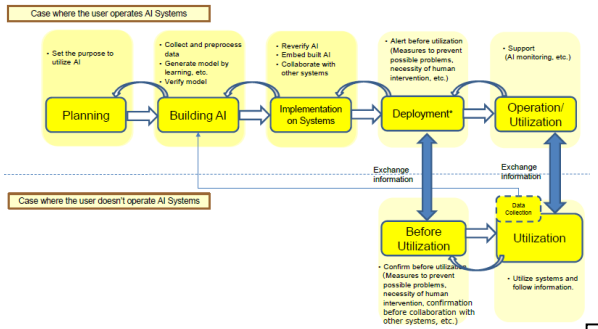
4. AI Utilization Principles

10 principles

1. Proper Utilization
2. Data quality
3. Collaboration
4. Safety
5. Security
6. Privacy
7. Human dignity and individual autonomy
8. Fairness
9. Transparency
10. Accountability

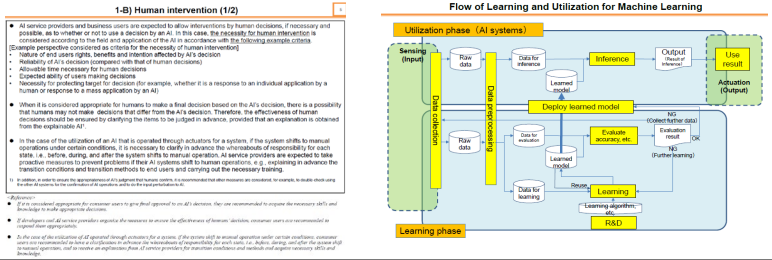
Part 2: Comments of AI Utilization Principles

5. General flow of AI utilization



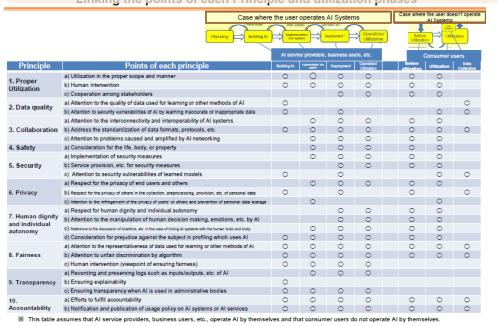
6. Commentary of the AI Utilization Principles

Commentary on points of the content of each principle



7. Timing to Consider the AI Utilization Principles

Phases in which each principle and its points should considered



Purpose and Basic Philosophies

Purpose

To **facilitate AI utilization and social implementation** by way of increasing the benefits, and mitigating the risks of AI, as well as fostering trust in AI through the sound progress of AI networks

Basic Philosophies

- To achieve a **human-centred society**
- To respect the **diversity** of users and advance **inclusion** of people with diverse backgrounds
- To achieve a **sustainable society** that can solve various problems faced by individuals, local communities, countries, and the international community
- To ensure an appropriate **balance between the benefits and risks** of AI
- To **realize appropriate role assignment among stakeholders** with consideration for **ability and knowledge** on AI that each user is expected to have
- To **share the Guidelines and their best practices** internationally among stakeholders
- To constantly **review** the Guidelines and **flexibly revise** them as necessary

Classification of Related Entities

Developers

Those who conduct the R&D of AI systems

Users

Those who use AI systems, AI services, or AI-accompanying 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 affected due to AI systems or AI services used by others.

AI service providers

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

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
(including non-profit specialists and administrative bodies)

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

(Note) It is expected that business users who use AI systems or AI services without operating AI systems themselves won't be able to pay the same level of attention as other business users. However, even in such cases, they are expected to promptly request appropriate measures from developers or AI service providers, etc.

Consumer users

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

(Note) Consumer users who operate AI systems themselves may be required to pay the same level of attention as developers and AI service providers, etc.

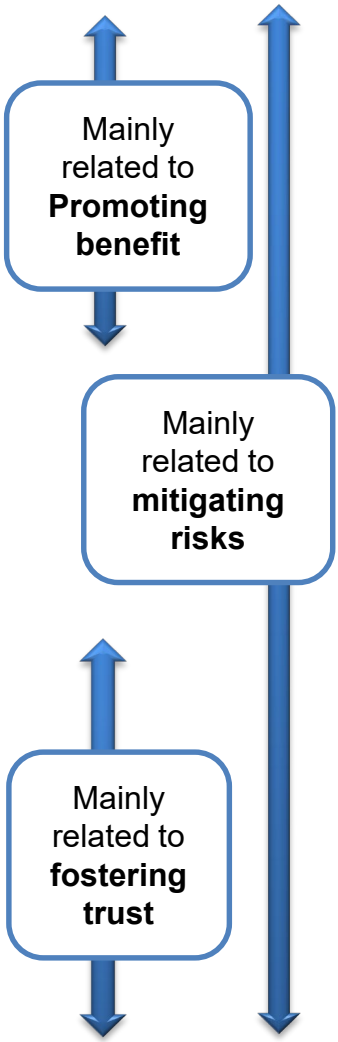
- AI systems : Systems that incorporate AI software as a component.
- AI services : Services that provide the functions of AI systems
- AI-accompanying services : AI-systems-update services or additional learning services, etc.

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

AI Utilization Principles

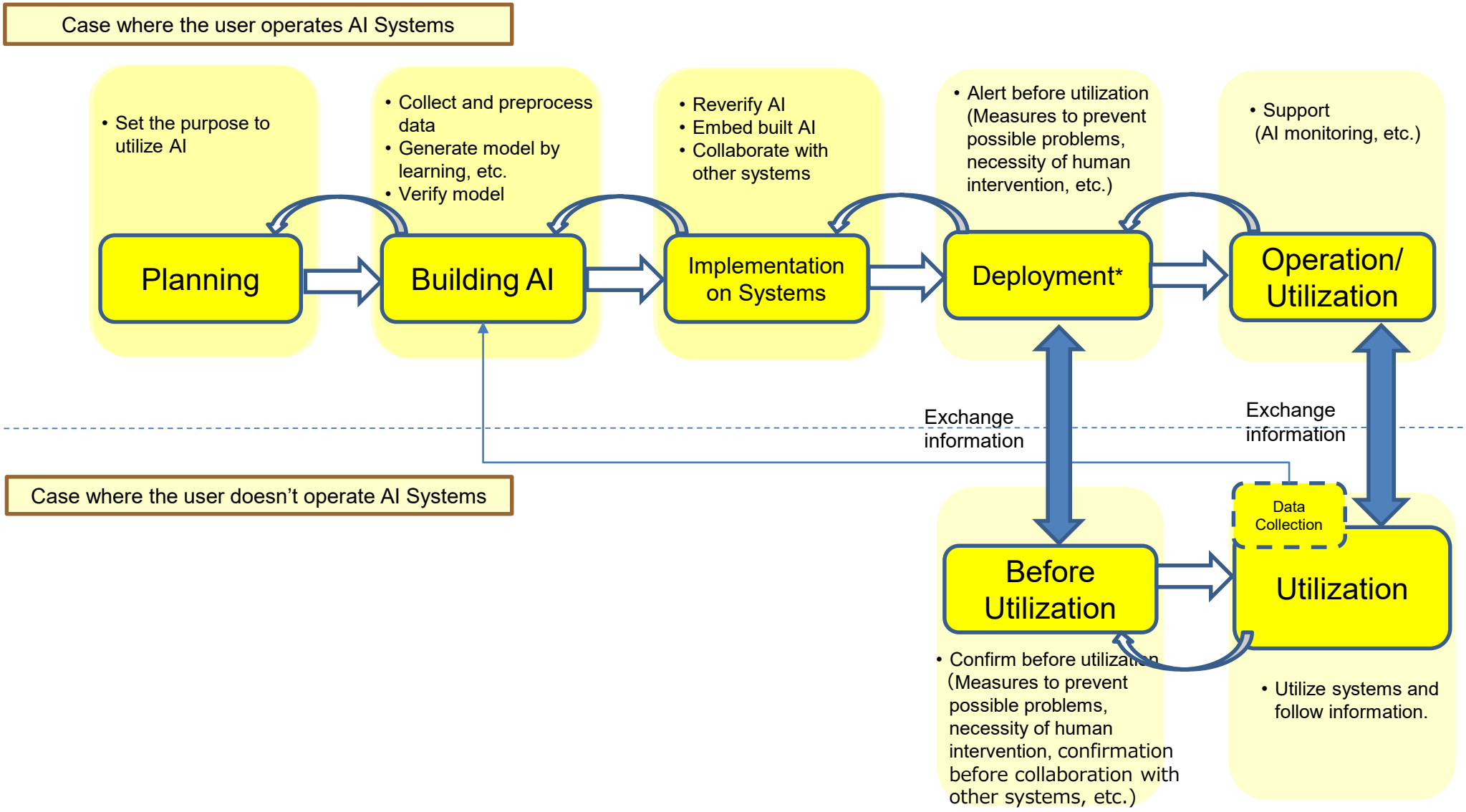
Compiled AI Utilization Principles that are expected to be referred to by AI service providers, business users, etc. and are to be shared internationally

Principle of	Description
1. 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. Data quality	Users and data providers should pay attention to the quality of data used for learning or other methods of AI systems.
3. Collaboration	AI service providers, business users, and data providers should pay attention to the collaboration of AI systems or AI services. Users should take into consideration that risks might occur and even be amplified when AI systems are to be networked.
4. Safety	Users should take into consideration that AI systems or AI services in use will not harm the life, body, or property of users or third parties through the actuators or other devices.
5. Security	Users and data providers should pay attention to the security of AI systems or AI services.
6. 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. Human dignity and individual autonomy	Users should respect human dignity and individual autonomy in the utilization of AI systems or AI services.
8. Fairness[1]	AI service providers, business users, and data providers should pay attention to the possibility of bias inherent in the judgements of AI systems or AI services, and take into consideration that individuals and groups will not be unfairly discriminated against by their judgments.
9. Transparency[2]	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. Accountability[3]	Users should make efforts to fulfill their accountability to the stakeholders.



1) It should be noted that there are multiple definitions and criteria for “fairness.”
2) 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.
3) “Accountability” means the possibility to take appropriate measures, such as to proven an explanation of the meaning and reason for the judgment, along with compensation as needed, after clarifying with the person responsible, in order to gain the understanding of the person who is affected by the result of the judgment.

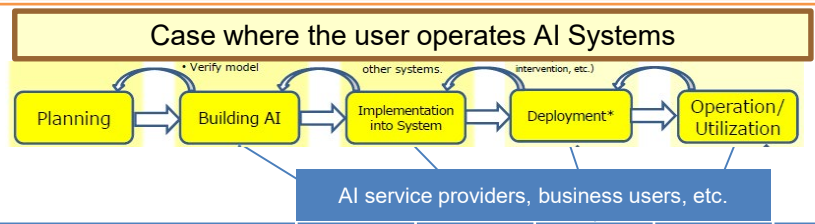
General flow of AI utilization



*Deployment: To make AI software/systems available

Note: This general flow of AI utilization describes a typical case in order to clarify the phase in which each principle is to be considered. Keep in mind that there are various cases of AI utilization, such as in the case that development and operation are performed simultaneously (ex. DevOps).

Linking the points of each Principle and utilization phases



Principle	Points of each principle	Building AI	Implementation into system	Deployment	Operation/ Utilization		Before utilization	Utilization	Data Collection
1. Proper Utilization	A) Utilization in the proper scope and manner	○	○	○	○		○	○	
	B) Human intervention	○	○	○	○		○	○	
	C) Cooperation among stakeholders			○	○		○	○	
2. Data quality	A) Attention to the quality of data used for learning or other methods of AI	○							○
	B) Attention to security vulnerabilities of AI by learning inaccurate or inappropriate data	○		○				○	○
3. Collaboration	A) Attention to the interconnectivity and interoperability of AI systems		○	○	○		○	○	
	B) Address the standardization of data formats, protocols, etc.	○	○	○	○		○	○	○
	C) Attention to problems caused and amplified by AI networking		○	○	○		○	○	
4. Safety	A) Consideration for the life, body, or property		○	○	○		○	○	
5. Security	A) Implementation of security measures		○	○	○		○	○	
	B) Service provision, etc. for security measures			○	○		○	○	
	C) Attention to security vulnerabilities of learned models	○		○				○	○
6. Privacy	A) Respect for the privacy of end users and others		○	○	○		○	○	
	B) Respect for the privacy of others in the collection, preprocessing, provision, etc. of personal data	○		○			○		○
	C) Attention to the infringement of the privacy of users' or others and prevention of personal data leakage		○					○	
7. Human dignity and individual autonomy	A) Respect for human dignity and individual autonomy			○	○		○	○	
	B) Attention to the manipulation of human decision making, emotions, etc. by AI			○	○		○	○	
	C) Reference to the discussion of bioethics, etc. in the case of linking AI systems with a human brain and body		○	○	○		○	○	
	D) Consideration for prejudice against the subject in profiling which uses AI	○	○	○	○		○	○	
8. Fairness	A) Attention to the representativeness of data used for learning or other methods of AI	○	○	○	○		○	○	○
	B) Attention to unfair discrimination by learning algorithm	○	○	○	○		○	○	○
	C) Human intervention (viewpoint of ensuring fairness)	○	○	○	○				
9. Transparency	A) Recording and preserving logs such as inputs/outputs, etc. of AI		○	○	○				
	B) Ensuring explainability	○							
	C) Ensuring transparency when AI is used in administrative bodies	○	○	○	○				
10. Accountability	A) Efforts to fulfill accountability	○	○	○	○		○	○	○
	B) Notification and publication of usage policy on AI systems or AI services	○	○	○	○		○	○	○

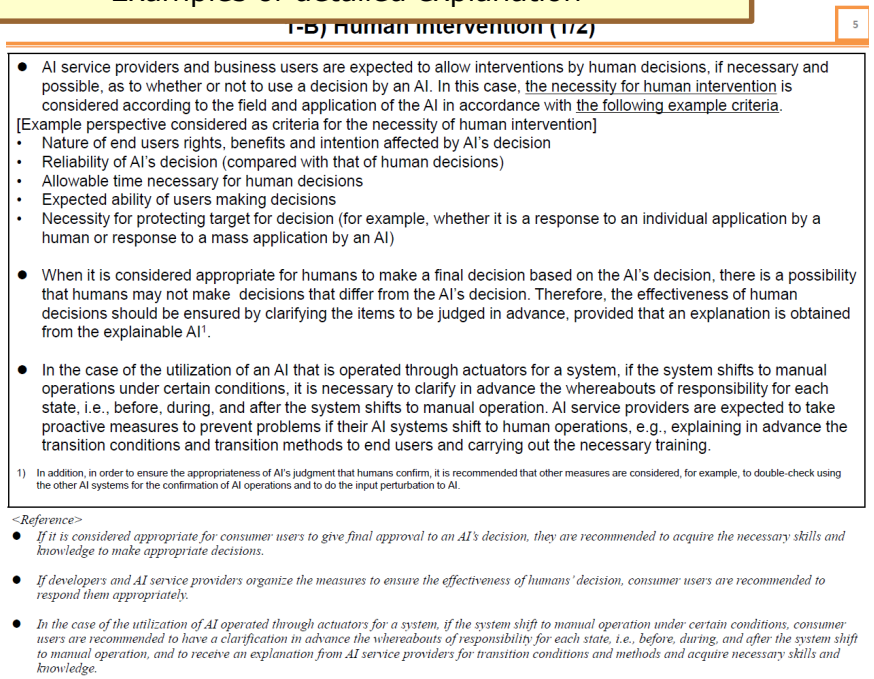
※ This table assumes that AI service providers, business users, etc., operate AI by themselves and that consumer users do not operate AI by themselves.

Commentary on “AI Utilization Principles”

Concept of organization of detailed explanation

- The detailed explanation for each point, of each principle, is organized as a matter to be noted from the following perspectives（the lower left text）：
  - AI service providers, business users and data providers
  - Consumer users
- The explanation is supplemented by a concept diagram on “machine learning”, which is used as major AI technologies in recent years, on the assumption that “the AI Utilization Guidelines” are reviewed regularly (the lower right figure)

Examples of detailed explanation



1-B) Human intervention (1/2)

5

- AI service providers and business users are expected to allow interventions by human decisions, if necessary and possible, as to whether or not to use a decision by an AI. In this case, the necessity for human intervention is considered according to the field and application of the AI in accordance with the following example criteria.

[Example perspective considered as criteria for the necessity of human intervention]

- Nature of end users rights, benefits and intention affected by AI's decision
- Reliability of AI's decision (compared with that of human decisions)
- Allowable time necessary for human decisions
- Expected ability of users making decisions
- Necessity for protecting target for decision (for example, whether it is a response to an individual application by a human or response to a mass application by an AI)

- When it is considered appropriate for humans to make a final decision based on the AI's decision, there is a possibility that humans may not make decisions that differ from the AI's decision. Therefore, the effectiveness of human decisions should be ensured by clarifying the items to be judged in advance, provided that an explanation is obtained from the explainable AI[1].

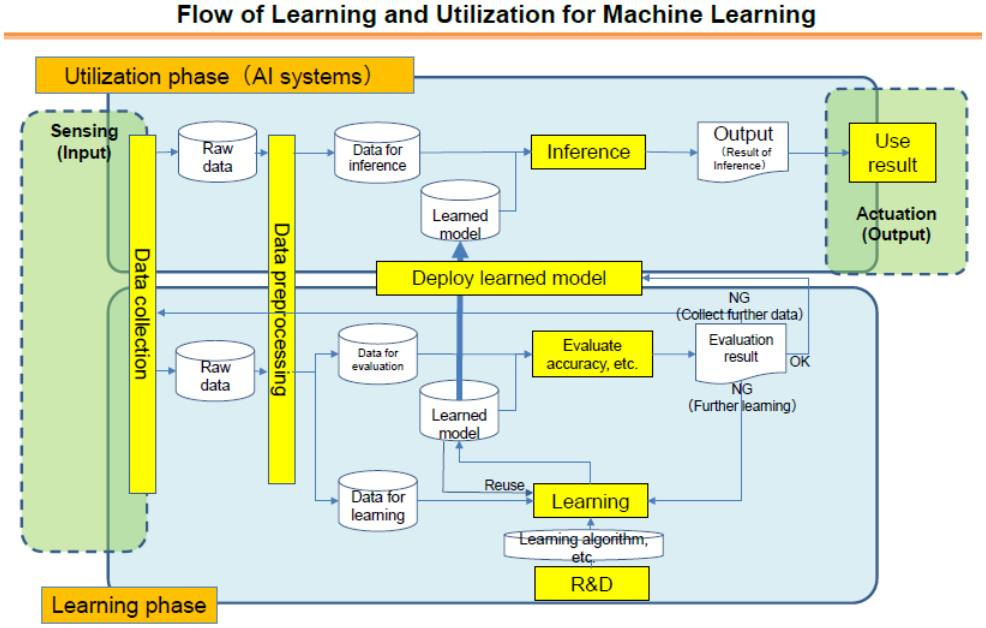
- In the case of the utilization of an AI that is operated through actuators for a system, if the system shifts to manual operations under certain conditions, it is necessary to clarify in advance the whereabouts of responsibility for each state, i.e., before, during, and after the system shifts to manual operation. AI service providers are expected to take proactive measures to prevent problems if their AI systems shift to human operations, e.g., explaining in advance the transition conditions and transition methods to end users and carrying out the necessary training.

1) In addition, in order to ensure the appropriateness of AI's judgment that humans confirm, it is recommended that other measures are considered, for example, to double-check using the other AI systems for the confirmation of AI operations and to do the input perturbation to AI.

<Reference>

- *If it is considered appropriate for consumer users to give final approval to an AI's decision, they are recommended to acquire the necessary skills and knowledge to make appropriate decisions.*
- *If developers and AI service providers organize the measures to ensure the effectiveness of humans' decision, consumer users are recommended to respond them appropriately.*
- *In the case of the utilization of AI operated through actuators for a system, if the system shift to manual operation under certain conditions, consumer users are recommended to have a clarification in advance the whereabouts of responsibility for each state, i.e., before, during, and after the system shift to manual operation, and to receive an explanation from AI service providers for transition conditions and methods and acquire necessary skills and knowledge.*

Detailed explanation example (Focusing on notes on AI service providers, business users and data providers (In addition, notes on consumer users as reference))



Illustrated flow example

Future Challenges

	issues	overview
1. Matters related to the sound development of AI networking		
(1)	Dissemination and development of “AI R&D/Utilization Guidelines”	Holding symposiums to disseminate “AI R&D/Utilization Guidelines”; Dissemination of detailed explanations to realize the principles in international frameworks, etc.
(2)	Following-up for discussions regarding AI development/ utilization principles/guidelines	Following-up and continuously reviewing international discussions on AI development / utilization principles/guidelines.
(3)	Issues related to environmental improvements addressed by related stakeholders	Cooperation among stakeholders, sharing of best practices, and studies on the ideal state of legal systems.
(4)	Securement of the smooth collaboration of AI systems or AI services	Studies on the range of related information expected to be shared among stakeholders and how to share it.
(5)	Securement of a competitive ecosystem	Keeping watch on the trends of related markets.
(6)	Protection of the interests of users	Studies on the manner of voluntary information provision from developers to users and on the ideal state of protecting users (e.g., by insurance).
2. Matters related to the evaluation of the socioeconomic impact of AI networking		
(1)	Scenario analysis on the socioeconomic impact of AI networking	Ongoing implementation of scenario analysis and international sharing.
(2)	Establishment of evaluation indicators for the impact of the progress of AI networking, and for richness and happiness	Study on setting indicators.
(3)	Fostering social acceptability on the utilization of AI systems	Keeping watch on the degree of social acceptance for the utilization of AI systems.
3. Matters related to issues over a human who is in a society under the ongoing progress of AI networking		
(1)	Study on the ideal state of relationship between humans and AI systems	Studies on the ideal state of role assignments of professionals (doctors, lawyers, accountants, etc.) and AI systems.
(2)	Study on the ideal state of relationship among stakeholders	Studies on the ideal state of responsibility sharing in case of AI’s risks becoming apparent.
(3)	Safety net development	Keeping watch on labor market trends and prevention of the unfair redistribution of income accompanying the progress of AI networking.

References

Commentary on “AI Utilization Principles” (example)

4. Principle of Safety

AI service providers should take the following measures with consideration for potential damage, etc.:

(A) Build a mechanism to ensure the safety of the entire system when implementing AI (fail-safe)

(B) Inspect, repair and update AI

8. Principle of Fairness

AI service providers should take the following actions according to the social context, etc.:

(C) Ensure representativeness of learning data

(D) Remove biases inherent in learning data

(E) Eliminate biases of those who label learning data

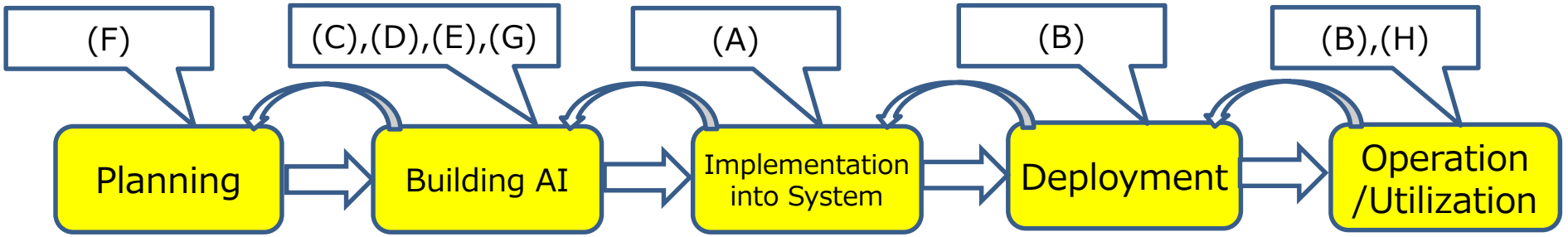
(F) Clarify matters to be fair and design an algorithm satisfying them

9. Principle of Transparency

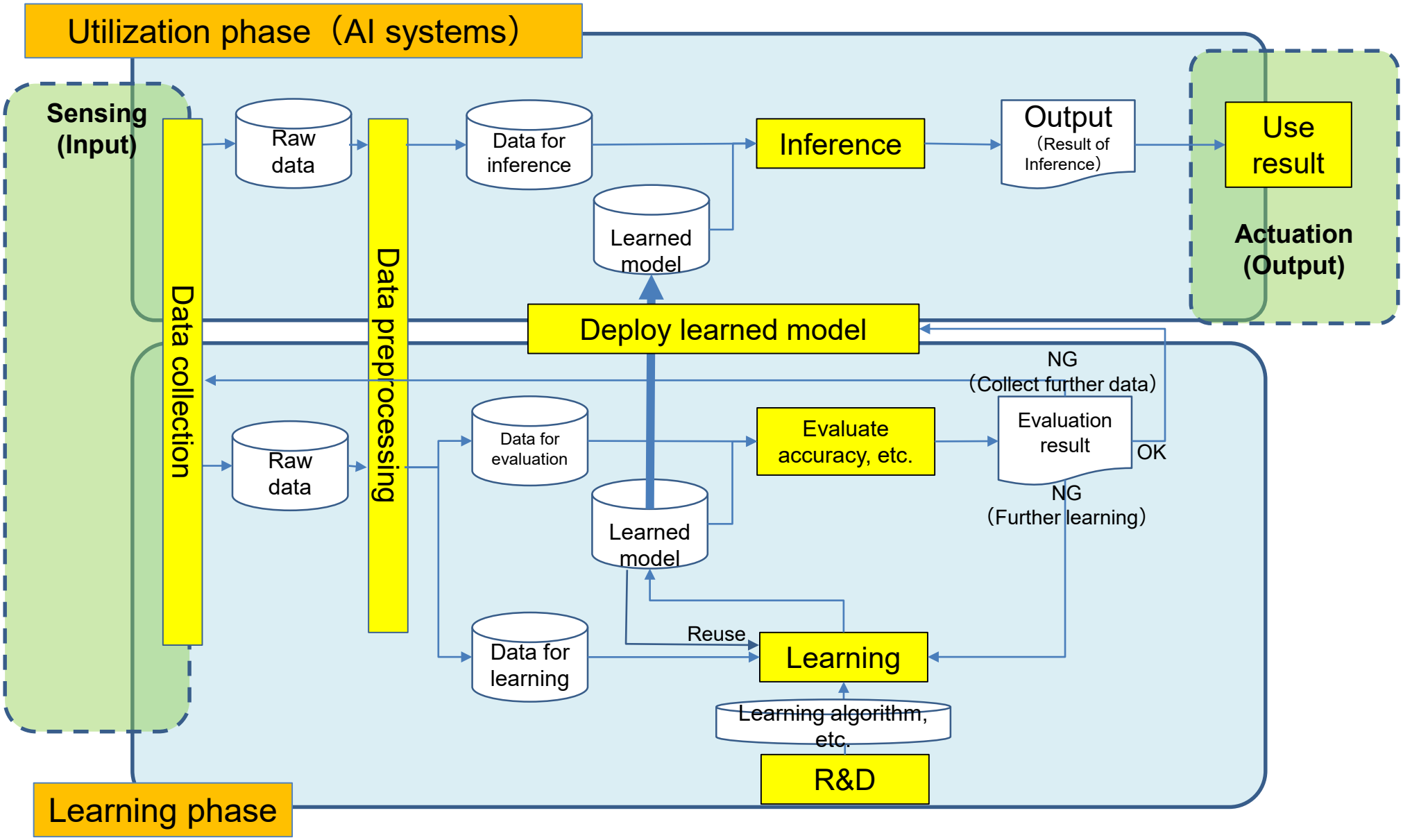
AI service providers should take the following measures in light of AI usage, etc.:

(G) Manage the provenance of data used for learning, etc.

(H) Save the related logs (input/output, etc.)



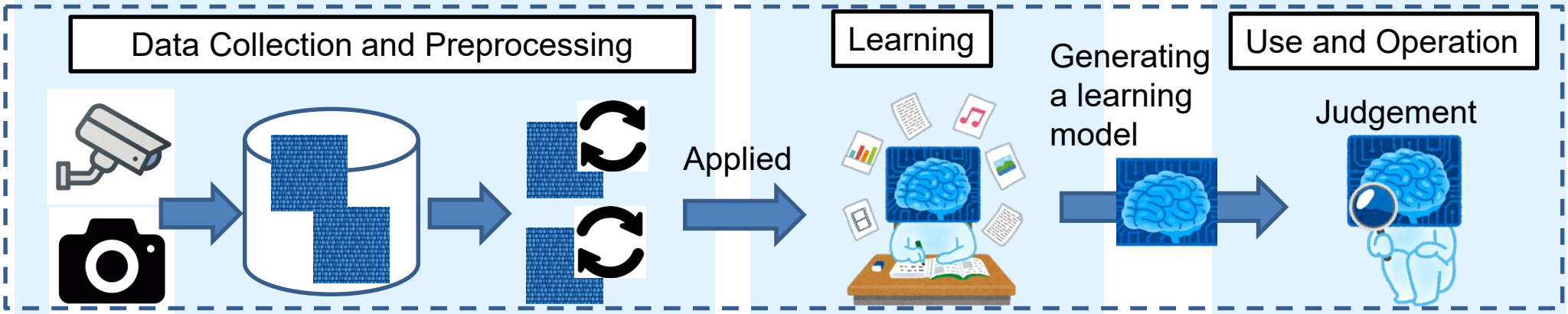
Flow of Learning and Utilization for Machine Learning



Model cases for the application of “AI Utilization Guidelines” (1/2)

Indicate specific measures to implement the principles according to the process of AI utilization

Ex. : A system that judges whether there is a possibility of a crime being committed through human image input



[Fairness]
Ensure data representativeness:
Not to target persons living in specific regions

[Fairness]
Eliminate bias in annotating data:
Not to label data with personal intentions and impressions

[Transparency]
Data provenance

[Privacy]
Respect for the privacy of persons who were captured on the images

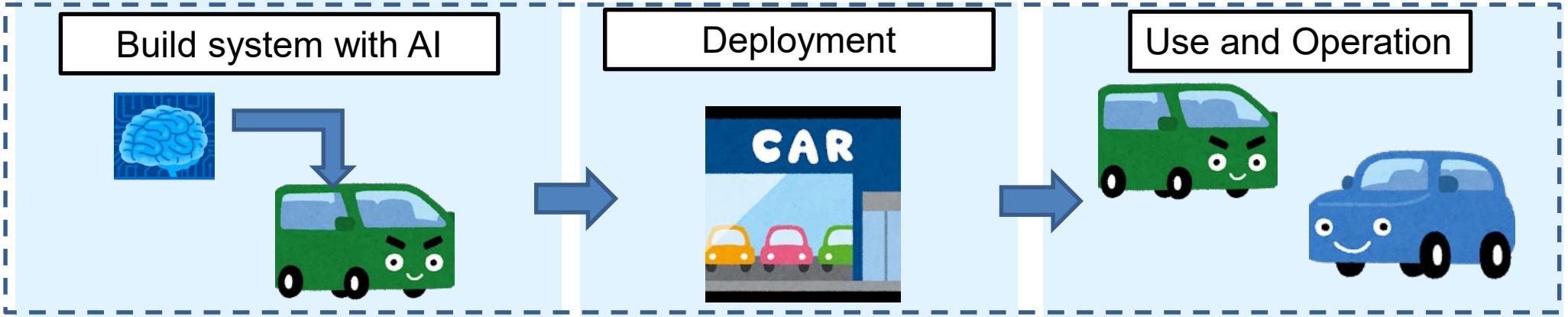
[Fairness]
Attention to unfair discrimination by algorithm:
Not to discriminate in the computing process

[Proper utilization]
Human intervention:
Consider end user's right to be influenced by AI judgements

Model cases for the application of “AI Utilization Guidelines” (2/2)

Indicate specific measures to implement the principles according to the process of AI utilization

Ex. Autonomous driving



[Safety]
Ensure safety across the entire system (**Fail-safe**)

[Security]
Take reasonable measures corresponding to the current technology level to prevent system hacking

[Safety/Security]
Share information about measures to be taken when infringement occurs

[Proper Utilization]
Human intervention:
Share conditions on switching control from AI to human

[Safety/Security]
Provide updates (information) for systems with **AI**

[Transparency/Accountability]
Ensure explainability, and **fulfill accountability**, when accident occurs

[Collaboration]

- Negotiate and coordinate among autonomous vehicles, and **support data format / protocol**
- Address risks that a problem in one AI system spreads to the entire system

Examples of Trade-offs