Tentative
Translation

Report 2019 Appendix 1
(Attachment)

Detailed Explanation on Key Points Concerning AI Utilization Principles

August 9, 2019
The Conference toward AI Network Society

Explanatory notes

Content described from the viewpoint of **AI service providers, business users, and data providers**

<Reference>
*Content described from the viewpoint of **consumer users***

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.

Key points

A) Utilization in the proper scope and manner

B) Human intervention

C) Cooperation among stakeholders

1-A) Utilization in the proper scope and manner (1/2)

- AI service providers and business users are expected to use AI within a proper scope and manner, based on the information and explanations provided by developers, and after properly recognizing the utilization purpose, usage, nature, and capability of the AI according to its social context. Furthermore, AI service providers are expected to provide information on the following items in a timely manner:

[Information to be provided]

- Applications and usage of provided AI
- Benefits and risks according to the nature and usage mode of AI
- How to periodically check the range and method of provided AI (in particular, observation and confirmation methods for autonomous AI updates), the importance and frequency of confirmation, and the risks resulting from the absence of confirmations.
- AI software updates[1], inspections, and repairs which are implemented to improve AI functions and control risks through their utilization.

[Timing for provisions information]

- It is desirable to provide corresponding information before using AI.
- If the information cannot be provided in advance, with consideration for assumed risks based on the nature and usage mode of AI, it is desirable to have a system in place to respond to feedback from consumer users.

- AI service providers are expected to provide AI software updates and AI inspections/repairs, etc. services to improve AI functions and mitigate risks in their utilization. In particular, if it is assumed that the update affects[2] other linked AI systems, AI service providers are expected to provide information on these risks.

- Depending on the nature and usage mode of AI systems or AI services to be provided, AI service providers are expected to confirm the reliability of users in advance in cases where the use of an AI is likely to harm human lives, bodies, or property. Furthermore, after an AI service is provided, there may be a necessity for recording and saving input and output logs on the service in order to make sure that no end users misuse or make malicious use of the AI service or AI system.

1) From the time a problem is discovered to the time an update information is provided, AI service providers are expected to give information on the problem to end users in a timely and appropriate manner and alert them.
2) It is assumed that the operation of AI to which the update is applied affects the other AIs. For example, if AI software incorporated into a home electrical appliance is updated, it is assumed that a conflict will result between the judgment of a home administration robot in control of the whole house and that of other home electrical appliances incorporating AI unless they respond to the updating. (An example case in Appendix 3 *Risk of AI in Unexpected Operation* in “Report 2018”).

1-A) Utilization in the proper scope and manner (2/2)

<Reference>

- *Consumer users are recommended to use AI in a proper range and method, with consideration for information and explanations provided from developers and AI service providers, and within a social context. When using an AI, they are recommended to take the following matters into consideration.*

[Matters that should be taken into consideration]

(Before use)

- *Recognize the benefits and risks, understand its proper use, and acquire the needed knowledge and skills according to the nature of the AI and the mode of its use.*

(In use)

- *Regularly check that the users' own utilization of the AI is within a proper range and method.*
- *Make efforts towards updating AI software, inspections, and repairs in order to improve the functions of the AI and to control risks through the utilization of the AI. (However, it is desirable to keep in mind that updates may affect other linked AI systems.)*
- *Provide feedback information to the AI developers and service providers if a problem occurs or if there is a sign of a problem.*

1-B) Human intervention (1/2)

- 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.

1-B) Human intervention (2/2)

<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.*

1-C) Cooperation among stakeholders

- AI service providers, business users and data providers are expected to cooperate with related stakeholders and to work on preventive or remedial measures (including information sharing, stopping and starting of AI, elucidation of causes, and measures to prevent recurrence, etc.) in accordance with the nature, and conditions, etc. of accidents that have occurred or may occur in the future through the use of AI or damage caused by security breaches and privacy infringements, etc.

- At that time, they are expected to pay attention to the following principles.

[Examples of preventive or remedial measures to be taken jointly by the related stakeholders]

- 1) Principle of proper utilization
  - A) Utilization within the proper scope and manner
    (Provision of information for utilization within the proper scope and manner)
- 4) Principle of safety
  - A) Consideration for live, body, and property
    (Measures to be taken if AI damages human lives, bodies, or property through actuators, for example)
- 5) Principle of security
  - A) Implementation of security measures
    (Measures to be taken if security is infringed)
- 6) Principle of privacy
  - A) Respect for the privacy of end users and others
    (Measures to be taken if the privacy of others is violated)

<Reference>

- *With consideration of information that developers or AI service providers provide, consumer users are recommended to cooperate with related 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 accidents through the use of AI or damages caused by security breaches and privacy infringement, etc. that may occur in the future or have occurred.*

2) Principle of Data Quality

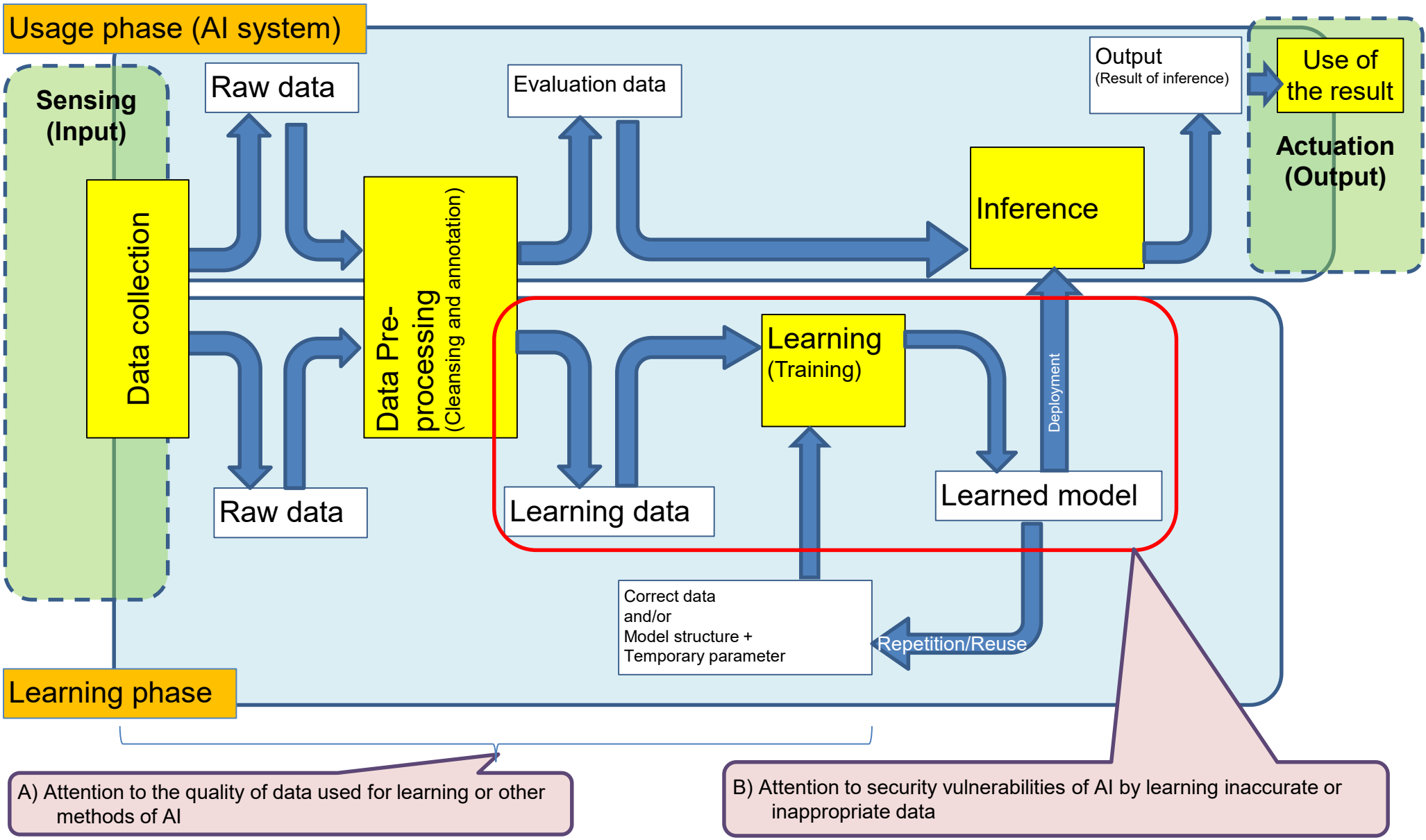
Users and data providers should pay attention to the quality of data used for learning and other methods of AI systems.

Key points

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

Each point concerning the 2) Principle of Data Quality in the Flow of Learning and Use Focusing on Machine Learning



2-A) Attention to the quality of data used for learning or other methods of AI(1/2)

- AI service providers, business users, and data providers are expected to pay attention to the quality of data (e.g. data accuracy and integrity) used for learning or other AI methods, with consideration for the characteristics of the AI to be used and its usage. In machine learning, in particular, it is possible to ensure data quality via the following measures.

[Example measures for collecting data]

- Check if the collected data is suitable for the purpose of using AI.
- Use data released by highly trusted sources.
- Confirm the provenance of the data.
- Be aware of rights embedded into data when collecting it.

[Example measures for data pre-processing]

- Exclude the data which humans cannot understand or identify from those for learning[1].
- Actively adopt the data for learning if it is easily misrecognized by machines (learner)[2].
- Be careful not to cause an error when annotating (labeling) (especially for supervised learning).
- Create a data set while being conscious of the data format used (input) at the utilization phase.
- Acquire and store logs on how pre-processing is performed (the provenance of data pre-processing).

[Example measures for learning]

- Perform transfer learning[3] using an existing learning model.
- Perform learning after specific data augmentation[4] in order to improve the accuracy of learning.
- Properly define the temporal range of data to be learned when AI learns transient time-series data.

1) For example, in the case of image recognition, where the target object cannot be identified even with human eyes.
2) For example, in the case of image recognition, where the target object is at the edge.
3) “Transfer learning”, which is a technique for machine learning, is a technique applying a model learned in a specific domain to another domain. The merit is that there is a possibility that high precision results can be obtained with a small amount of data.
4) “Data augmentation” is a technique taken to ensure data accuracy by enhancing generalization performance (performance for unknown data) when specific learning data is scarce. Generalization performance may be improved by expanding data used for learning (applying inversion, enlargement, or reduction, for example, in the case of image data) and using each as data based on another source.

2-A) Attention to the quality of data used for learning or other methods of AI(2/2)

- It is assumed that the accuracy[1] of an AI's judgment can become impaired or decline afterwards. Therefore, AI service providers, business users, and data providers are expected to define reference levels concerning accuracy in advance based on the assumed magnitude and frequency of occurrence of the infringement of rights, the technology level available, and the cost[2] to maintain accuracy, etc. If accuracy falls below a reference level, they are expected to put the AI through relearning with consideration for data quality.

- If it is planned to use data provided by consumer users, they are expected to provide consumer users with information on the means and format of data provision in advance, taking into consideration the characteristics and usage of the AI.

1) The term "accuracy" includes checks on whether AI is making right judgments, for example, checks on whether AI is not using a violent expression or making hate speech.
2) For example, since AI based on machine learning is an inductive approach, the corresponding AI alone cannot in principle guarantee 100% accuracy.

<Reference>

- *If consumer users plan to collect data by themselves and make AI learn the collected data, it is recommended that the data format should be based on the information provided by developers and AI service providers.*

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

- AI service providers, business users, and data providers are expected to pay attention to the risk that AI security might become vulnerable by learning inaccurate or inappropriate data. They are also expected to inform consumer users in advance of the existence of such risks.

[Examples risks]

- A risk that can mislead a learning model away from accurate judgments as a result of insufficient learning from deliberately inputting data with a slight change that cannot be noticed by humans into the learning model (e.g., adversarial example attack).
- A risk of making learning (models) fail by mixing incorrectly labeled data in supervised-learning.

<Reference>

- *Consumer users are recommended to pay attention to the risk of vulnerability to AI security by learning inaccurate or inappropriate data with consideration for information from developers, AI service providers, and data providers.*

- *Furthermore, if they have doubts in security when using the AI, they are recommended to report them to developers, AI service providers, and data providers.*

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 into consideration that risks might occur and even be amplified when AI systems are to be networked.

Key points

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

3-A) Attention to the interconnectivity and interoperability of AI systems

- AI service providers are expected to pay attention to the interconnectivity and interoperability of AI, with consideration for 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.

3-B) Address the standardization of data formats, protocols, etc.

- AI service providers and business users are expected to comply with the following standards in order to promote collaboration among AIs and between AIs and other systems: Data format (with syntax and semantics[1]) for AI input and output, and connection methods for collaboration (especially protocols in each layer when using a network for collaboration).

- Data providers are also expected to comply with data format standards (with syntax and semantics[1]) to promote collaboration among Ais, and between AIs and other systems.

1) Even if the syntax of data is shown, the collaboration will not work correctly unless the meaning is shown.

<Reference>

- *If consumer users plan to collect data by themselves and make AI learn from the collected data, it is recommended that the data format should be based on the information provided by developers and AI service providers.*

3-C) Attention to problems caused and amplified by AI networking

- Although it is expected that benefits will be promoted through collaboration between AIs, AI service providers, business users, and data providers are expected to pay attention to the possibility that risks might be caused and amplified by AI networking. Therefore, AI service providers, business users, and data providers are expected to analyze possible risks with consideration for information from developers, while sharing the risks with the cooperating parties, organizing preventive measures and countermeasures for problems, if any, and providing necessary information to consumer users.

[Example risks which may be caused and amplified by AI networking]

- Risks that individual AI system's problems, etc. spread through the entire system.
- Risks of failure in the cooperation and adjustment between AI systems.
- Risks of failure in verifying the judgment and the decision making of an AI (risk of failure to analyze the interactions between AI systems because the interactions become complicated).
- Risks that the influence of a small number of AIs become too strong (risks of enterprises and individuals suffering disadvantages because of judgements made by a few AI systems).
- Risk that competition and control in the market do not work due to multiple AIs making the same judgment or taking the same action.
- Risks of the infringement of privacy as a result of information sharing across fields and the concentration of information in one specific AI.
- Risks of unexpected actions by AIs.

<Reference>

- *Although it is expected that benefits will be enhanced through the interaction of AI systems, consumer users are recommended to pay attention that risks (e.g. loss of control by interconnecting or collaborating their AIs with other AIs, etc. through the Internet or other network) might be caused and amplified by AI networking. Furthermore, if information on preventive measures in advance or countermeasures for problems that have occurred are provided from the developers and AI service providers, they are recommended to be careful when using AIs.*

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 or third parties through actuators or other devices.

Key points

A) Consideration for the life, body, and property

4-A) Consideration for life, body or property (1/2)

- In cases where AI is used in fields where AI may harm human life, body, or property, AI service providers and business users are expected to take into consideration so that AI will not harm them through actuators or other devices by taking the following measures as necessary, based on information from the developers, and with consideration of the nature, and conditions, etc. of the assumed damage.

[Example countermeasures]

- Perform AI inspections, repairs, and AI software updates[1], and encourage consumer users to carry them out.
- Provide a fail-safe[2] design, for example, by constructing a mechanism that can ensure the safety of entire systems even if an unexpected operation is caused by the AI[3].

- Furthermore, AI service providers and business users are expected to organize in advance the measures to be taken if an AI damages a human life, body, or property through actuators or other devices. Also, they are expected to provide necessary information on such countermeasures to consumer users.

[Example measures in cases of harm]

- Initial actions (to be taken according to necessary procedures depending on the urgency of the affected systems or AI etc.)
  - Recovery by rolling back[4] the system or using an alternative system
  - System shutdown (by kill switch): If possible.
  - Network disconnection: If possible.
  - Confirmation of harmful content
  - Report to related parties
- Compensation, etc. (use of insurance for providing compensation)
- Establishment of a third-party organization and a cause investigation, analysis, and recommendation by the organization in the event of the occurrence of severe damage.

1) From the time of discovering a problem to the time of providing an update information, AI service providers is expected to give information on the problem to end users in a timely and appropriate manner and alert them.
2) Guide users in a safe direction so that damage will not occur if a failure occurs due to operation errors or malfunctions.
3) In situations where it is difficult to guarantee safety technically with AI alone, it is also possible to ensure safety of the system incorporating the AI by functions of the other systems, and demonstrate the safety of the AI through the operation experience of the system.
4) In the event of a failure, AI systems must restore to the previous (saved) state.

4-A) Consideration for life, body or property (2/2)

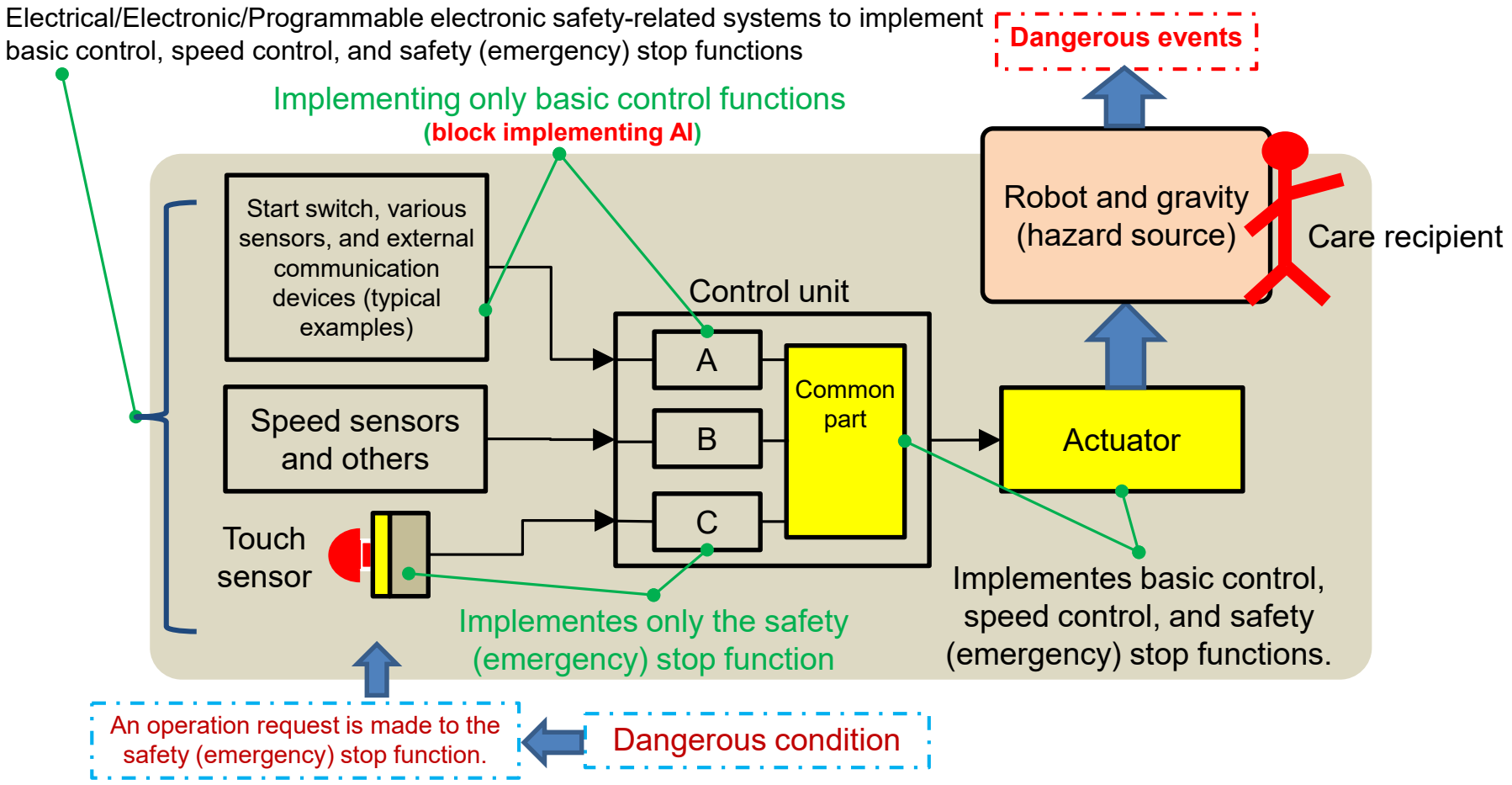
<Reference>

- *In case where AI is used in fields where AI may harm a human life, body, or property, consumer users are recommended to take into consideration that AI will not harm the human life, body, or property through the actuators or other devices by taking measures (e.g. checking AI, updating AI software, etc.) as necessary, based on information from the developers and AI service providers and with consideration of the nature, and conditions, etc. of assumed damages.*

- *Furthermore, if developers and AI service providers provide information on measures to be taken to prevent AI damage to human life, body, and property through actuators or other devices, consumer users are recommended to keep the information in mind when using the AI.*

[Reference] Example of a Mechanism for Ensuring Safety Outside AI

Electrical/Electronic/Programmable electronic safety-related systems of care robots with collision hazards and fall hazards (multiple protection layers)



Citation: Presentation material of Yoshinobu Sato, Technical Advisor, Nabtesco Corporation, at the meeting of Committee on AI Governance (4th)

→ Ensuring safety, including not only the AI mounting block (A) but also other systems (B and C) in the control unit.

5) Principle of Security

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

Key issues

A) Implementation of security measures

B) Service provision, etc. for security measures

C) Attention to security vulnerabilities of learned models

5-A) Implementation of Security Measures

- AI service providers and business users are expected to pay attention to the security of AI and take reasonable measures corresponding to the technology level at that time to ensure the confidentiality, integrity and availability (CIA) of AI systems.

- Furthermore, they are expected to organize measures to be taken against security infringements in advance, taking into consideration the usage and characteristics of the AI and the magnitude of the influence of the infringement.

[Example measures against security infringements]

- Initial actions (to be taken according to necessary procedures depending on the urgency of the affected systems or AI etc.)
  - Recovery by rolling back[1] the system or using an alternative system
  - System shutdown (by kill switch): If possible
  - Network disconnection: If possible
  - Content confirmation of security infringement
  - Report to the related parties
- Compensation, etc. (use of insurance for making compensation smoothly)
- Establishment of a third-party organization and a cause investigation, analysis, and recommendation by the organization in the event of the occurrence of severe damage.

1) In the event of a failure, AI systems must restore to the previous (saved) state.

<Reference>

- *If consumer users are supposed to implement security measures (on their side), they are recommended to pay attention to the security of the AI and take necessary measures based on the provision of information from developers and AI service providers.*

5-B) Service provisions, etc. for security measures

- AI service providers are expected, with regard to their AI services, to provide end users with services for security measures and to share past accident and incident information.

- Furthermore, AI service providers and business users are expected to provide consumer users with the necessary information on measures in cases of security infringements.

<Reference>

- *If developers and AI service providers provide information on measures to be taken against security infringement, consumer users are recommended to pay attention to this information when using the AI.*

- *Furthermore, if there are security concerns when using the AI, they are recommended to report them to developers, AI service providers, and data providers.*

5-C) Attention to security vulnerabilities of learned models

- AI service providers, business users, and data providers are expected to pay attention to the risk that the learning models in AI might be vulnerable in their generation and management. They are also expected to inform consumer users in advance the existence of such risks.

[Example risks]

- A risk that can mislead a learning model away from accurate judgments as a result of insufficient learning by deliberately inputting data with a slight change that cannot be noticed by humans into the learning model (e.g., adversarial example attack).
- A risk of making learning (models) fail by mixing incorrect labeled data in supervised-learning.
- A risk that learning models can be easily replicated.
- A risk that can reverse-engineer data used for learning from learning models.

<Reference>

- *Consumer users are recommended to pay attention to the risk that AI might become vulnerable in the generation and management of learning models, with consideration of information provided by developers, AI service providers, and data providers.*

- *Furthermore, if there are security concerns when using the AI, they are recommended to report them to developers, AI service providers, and data providers.*

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.

Note) In Japan, it is a precondition to comply with the Act on the Protection of Personal Information.

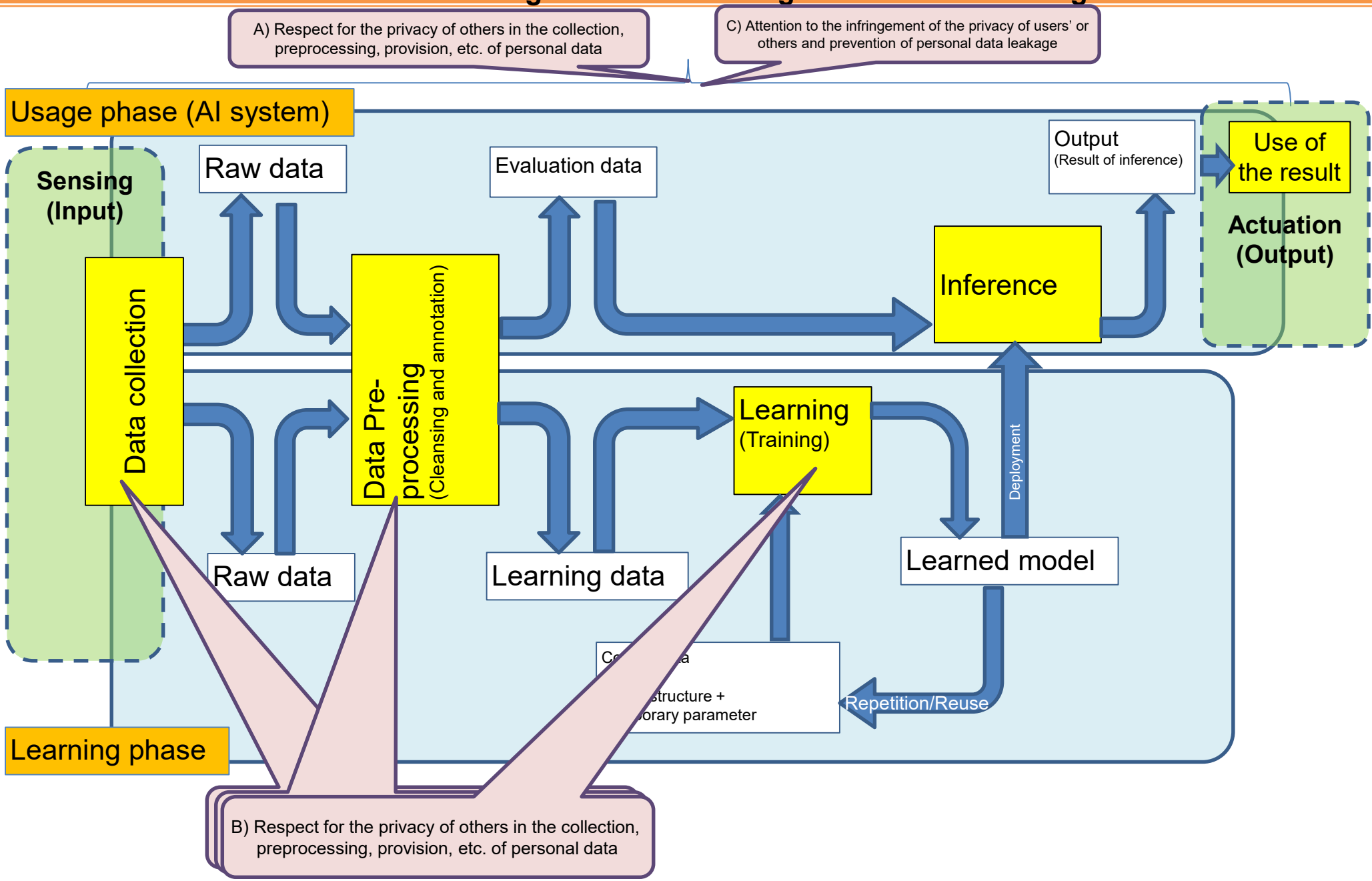
Key points

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

Each point concerning 6) Principle of Privacy in the Flow of Learning and Use Focusing on Machine Learning



6-A) Respect for the privacy of end users and others

- AI service providers and business users should respect the privacy of end users and third parties in the utilization of AI, based on the social context and reasonable expectations of people in its utilization.

- Furthermore, they are expected to consider measures to be taken against the privacy infringement of end users and third parties caused by AI.

- Besides, they are expected to provide the necessary information on such countermeasures to end users and third parties.

[Example actions to take in the case of privacy infringement]

- Deleting information that infringes on the privacy of end users and third parties, and updating AI algorithms if the information is incorrectly obtained.
- Requesting the information storage destination to delete information that infringes on the privacy of end users and third parties and updating the AI algorithm if the information is distributed.

<Reference>

- *Consumer users should respect the privacy of third parties in the utilization of AI, based on the social context and reasonable expectations of people in the utilization of AI.*

- *Furthermore, if developers and AI service providers provide information on measures to be taken against privacy infringement of third parties, they are recommended to pay attention to the information when using the AI.*

6-B) Respect for the privacy of others in the collection, preprocessing, and provision, etc. of personal data

- AI service providers, business users, and data providers should respect the privacy of end users and third parties in the collection, preprocessing, and provision[1,2] etc. of personal data used for AI learning and in the provision of learning models generated through them.

1) With regard to the handling of personal data provided to others, the deletion of the personal data, for example, is expected.
2) AI service providers, business users, and data providers are required to understand by whom and how the data they provide is used if the data contains personal information.

<Reference>

- *Consumer users should respect the privacy of third parties in the collection of data if they plan to collect data on their own for AI learning.*

6-C) Attention to the infringement of the privacy of users' or others and prevention of personal data leakage

- AI service providers, business users, and data providers are expected to take appropriate measures, including the prevention of unconsented data being made available to third parties, in their systems so that personal data is not provided under the judgement of AI to third parties without the consent of those persons.

<Reference>

- *Consumer users are recommended to be careful not to give particularly confidential information (e.g., others' personal information as well as their own information) unnecessarily to AI as a result of being overly emotional toward AI, including pet robots.*

7) Principle of Human Dignity and Individual Autonomy

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

Key points

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

7-A) Respect for human dignity and individual autonomy

- AI service providers and business users are expected to respect human dignity and individual autonomy based on the social context in the AI utilization[1].

1) For example, to recognize that AI supports human activities, on the assumption of the heterogeneity of humans and AI. The heterogeneity of humans and AI means that humans and AI have different natures. By this assumption, it can be recognized that AI should not be treated like a human (i.e., respect human dignity and individual autonomy).

<Reference>

- *Consumer users are recommended to respect human dignity and personal autonomy based on the social context in the AI utilization.*

7-B) Attention to the manipulation of human decision making, emotions, etc. by AI

- AI service providers and business users are expected to take necessary measures with consideration for the possibility[1] that consumer users' decisions or emotions are manipulated by AI, and the risk of their overdependence on AI.

[Examples of measures against decision-making and emotional manipulation]

- Alerting users when providing services
- Taking measures on the development side of computer systems including AI systems
- Support for sharing the above risks in education and other sites

1) What AI makes decision and manipulates consumer users' emotion is not always risk-taking, for example, in case of nudging by AI (i.e., supporting for a rational choice). Therefore, the term "possibility" is used here. In addition, in case of nudging by AI, AI service providers and business users are expected to refer to the Principle of user assistance (make it possible to provide selection opportunities to the users) in the "AI R&D Guidelines".

<Reference>

- *Consumer users are recommended to recognize the possibility that their decisions or emotions are manipulated by AI and the risk of over-dependence on AI, with consideration for information from developers and AI service providers[2].*

2) This includes not only information obtained directly from the developer and AI service provider, but also information obtained at educational sites, etc.

7-C) Reference to the discussion of bioethics, etc. in the case of linking AI systems with a human brain or body

- If AI is linked to a human brain and/or body, especially in pursuit of human enhancement (in pursuit of enhancements or improvements in the capabilities of humans that transcends maintaining or recovering health), AI service providers and business users are expected to particularly take into consideration that human dignity and autonomy are not violated, in light of the discussion of bioethics and information from developers about the surrounding technologies .

- Furthermore, they are expected to provide information on the function and peripheral technology of AI to be provided to consumer users.

<Reference>

- *If consumer users use AI that links to the human brain and body, they are recommended to pay attention to the possibility of the AI affecting the autonomy of humans and use the AI with consideration of information on functions and peripheral technologies of the AI from developers and AI service providers.*

7-D) Consideration for prejudice against the subject in profiling which uses AI

- In the case of profiling by using AI in fields that might have a significant influence on individual's rights or interests, AI service providers and business users are expected to carefully consider[1,2] all disadvantages that may occur to the target individuals.

[Examples that would cause a disadvantage in profiling]

- Incorrect decisions are made by providing profiling results that are different from the facts.
- Only specific characteristics of the target individual are used in profiling, resulting in the target individual being underrated.
- An adverse decision can be made on the target individual if a part of his/her profiling results is the same as the characteristics of a particular group and an unfavorable decision is made on the group.
- As a result of profiling, a treatment that impairs the rights and interests of target individuals or groups can occur, which may promote unfair discrimination against individuals or groups.
- Negative decisions can be made in the process of predicting (extrapolating) an uncertain future based on profiling results.
- Anonymous individuals are identified as a result of matching profiling results based on information about anonymous individuals with those based on information on particular individuals.

1) Article 22 of the EU's General Data Protection Regulation guarantees that the data subject shall have the right not to be subject to a decision based solely on automated processing.
2) Refer to B) human Intervention under 1) Principle of proper utilization.

<Reference>

- *Consumer users are recommended to be aware of the proper use of their information and, if necessary, check with AI service providers and business users considering that profiling might take place by AI.*

8) Principle of Fairness

AI service providers, business users, and data providers should pay attention to the possibility of bias[1] 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.

Note: there are multiple criteria for “fairness” such as group fairness and individual fairness.

Key issues

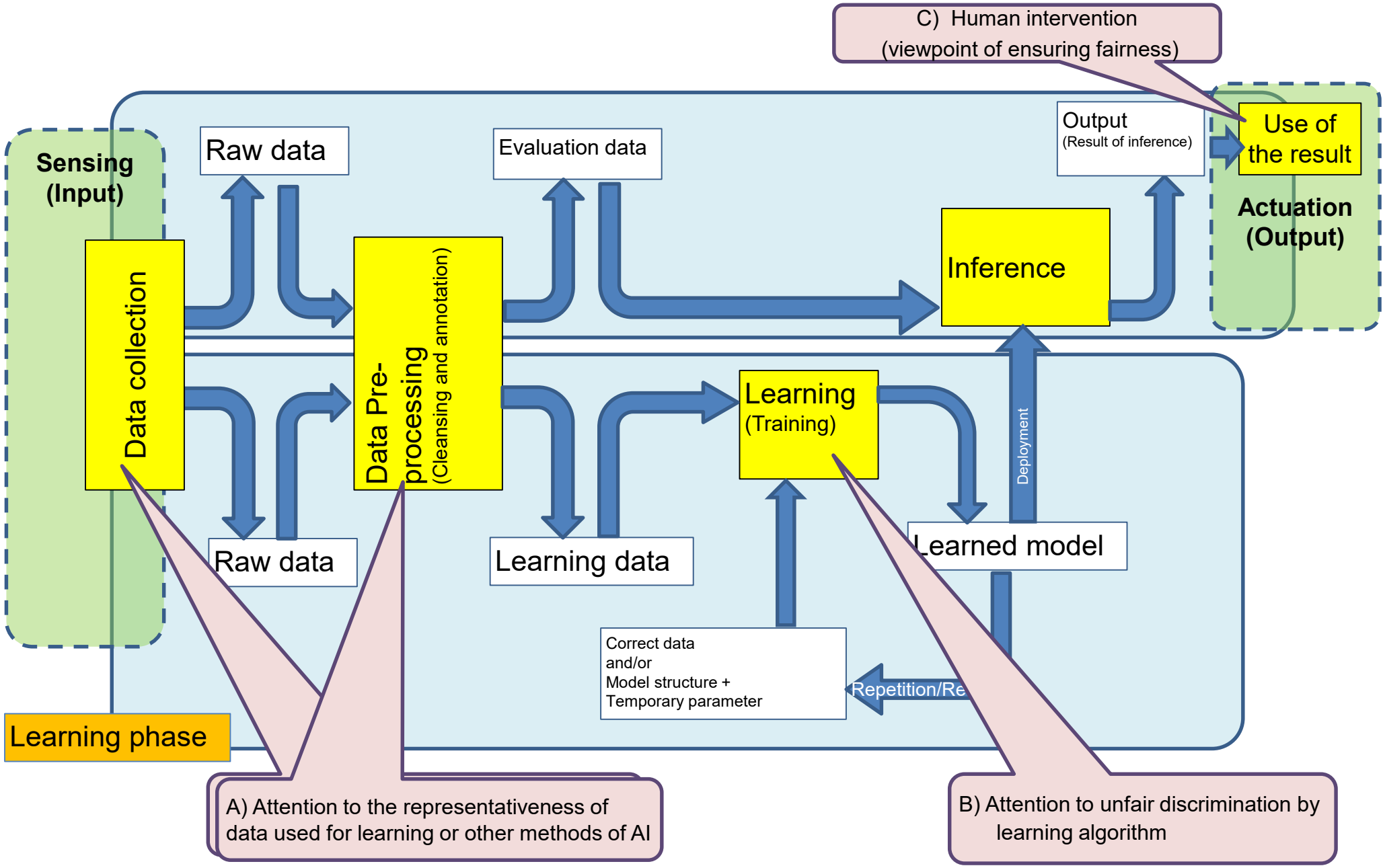
A) Attention to the representativeness of data used for learning or other methods of AI

B) Attention to unfair discrimination by learning algorithm

C) Human intervention (viewpoint of ensuring fairness)

1) The term “bias” has various possible interpretations as follows and is used as all-inclusive term in the Guidelines:
- Statistic terms (Sampling bias, deviation, etc.)
- Psychological terms (Cognitive bias (due to delusion, including social bias due to conventional wisdom etc. for each group), Emotional bias (due to human emotion and opportunity) etc.)

Each point concerning 8) Principle of Fairness in the Flow of Learning and Use Focusing on Machine Learning



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

- AI service providers, business users, and data providers are expected to pay attention to the representativeness[1] of data used for AI learning or other methods and the social bias inherent in the data according to the social context in utilizing AI, with consideration for how the result of AI judgements may be determined by learning data.

[Example matters to be taken into consideration from the viewpoint of fairness at AI learning]

- Take into consideration that AI can be biased due to the failure to ensure representativeness for the data even if the AI learning algorithm is designed to prevent an unfair judgment.
- Take into consideration that AI can be biased[2] as a result of using data embedding social bias even if sensitive information[3] is not embedded into the data[4].
- Take into consideration[5] that the learning data may be affected by the bias of the person who labels (intentionally or unintentionally) because the label for the learning data is often created and granted manually during the preprocessing phase (in the case of supervised learning).
- Respect personal privacy embedded in data in the case of collecting enormous amounts of data, including personal data, to satisfy the representativeness of the data.

1) The “representativeness” of data means the state in which data extracted as a sample and subjected to utilization does not distort the nature of the statistical population.
2) For example, in the case of implementing a gender-independent credit screening with the algorithm for the screening of each individual’s income as an attribute, bias by gender will occur as a result if there is a substantal difference in ratio of males and females with respect to high incomes.
3) Information on personal attributes, such as gender and race of the target person, which should be excluded from the viewpoint of fairness.
4) There is a consideration of criteria for ensuring fairness in case of learning by using data including sensitive information. The examples are described in [Example criteria for fairness] in 8-B).
5) There may be countermeasure such as developing common criteria for labeling.

<Reference>

- *If consumer users have any doubts in the decision made by AI, they are recommended to contact developers, AI service providers, and business users as required.*

8-B) Attention to unfair discrimination by learning algorithm

- AI service providers and business users are expected to pay attention to the possibility of bias inherent in AI judgements due to the algorithm used in it. In machine learning in particular, the majority tends to be more adopted, and the minority is less likely to be done (bandwagon effect). For example, there are the following measures to avoid this effect.

[Example measures for not causing bias by machine learning algorithm]

- Clarify sensitive attributes (i.e., personal attributes, such as gender and race of the target person, which should be excluded from the viewpoint of fairness) depending on the social context in the AI's utilization[1].
- Clarify the content of fairness to be ensured with respect to sensitive attributes depending on the following examples of criteria.
- Add constraints that satisfy the above fairness to machine learning algorithms.
- However, adding the above constraints may affect the accuracy of machine learning.

[Example criteria of fairness][2]

<Group fairness>

- Remove sensitive attributes and make predictions based only on non-sensitive attributes (unawareness).
- Ensure the same prediction result between multiple groups with different values of the sensitive attributes (demographic parity).
- Adjust the ratio of errors in the prediction result to the actual result so that the ratio will not depend on the value of the sensitive attributes (equalized odds).

<Individual fairness>

- Give the same prediction result to each person having the same attribute value other than the sensitive attribute.
- Give similar prediction results to each individual whose attribute values are similar (Fairness through awareness)

1) For example, gender is a sensitive attribute if gender is regarded as a problem in an entrance examination.
2) A method of making the average score of men and women identical in example 1).

<Reference>

- *If consumer users have any doubts in the decision made by AI, they are recommended to contact developers, AI service providers, and business users as required.*

8-C) Human intervention (viewpoint of ensuring fairness)

- AI service providers and business users are expected to intervene with human judgment on whether to adopt, or how to use the judgement of AI, as well as with consideration for the social context and the reasonable expectations of people, when utilizing an AI to ensure the fairness of the judgement results from it.

- They are expected to consider the necessity of human intervention based on the following example criteria from a viewpoint of fairness while referring to the content of [1- B)].

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

- The case in which the future statistical prediction is difficult (with high uncertainty). [2]
- The case in which a justifiable reason is necessary for making a decision. [3]
- The case in which there is a discrimination based on race, belief, or gender is assumed because the learning data contains a bias for minorities.

1) It is premised that the social bias inherent in data used in AI learning can affect the fairness of decision made by an AI.
2) For example, time-varying variables such as employee skills and productivity should be used for personnel affairs. In addition, it is difficult to predict future statistically because information that cannot be recorded is of no use.
3) For example, at a personnel evaluation, it is expected that the reason for the evaluation can be explained to employees.

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 algorithms, source codes, or learning data. In interpreting this principle, the privacy of individuals and trade secrets of enterprises are also taken into account.

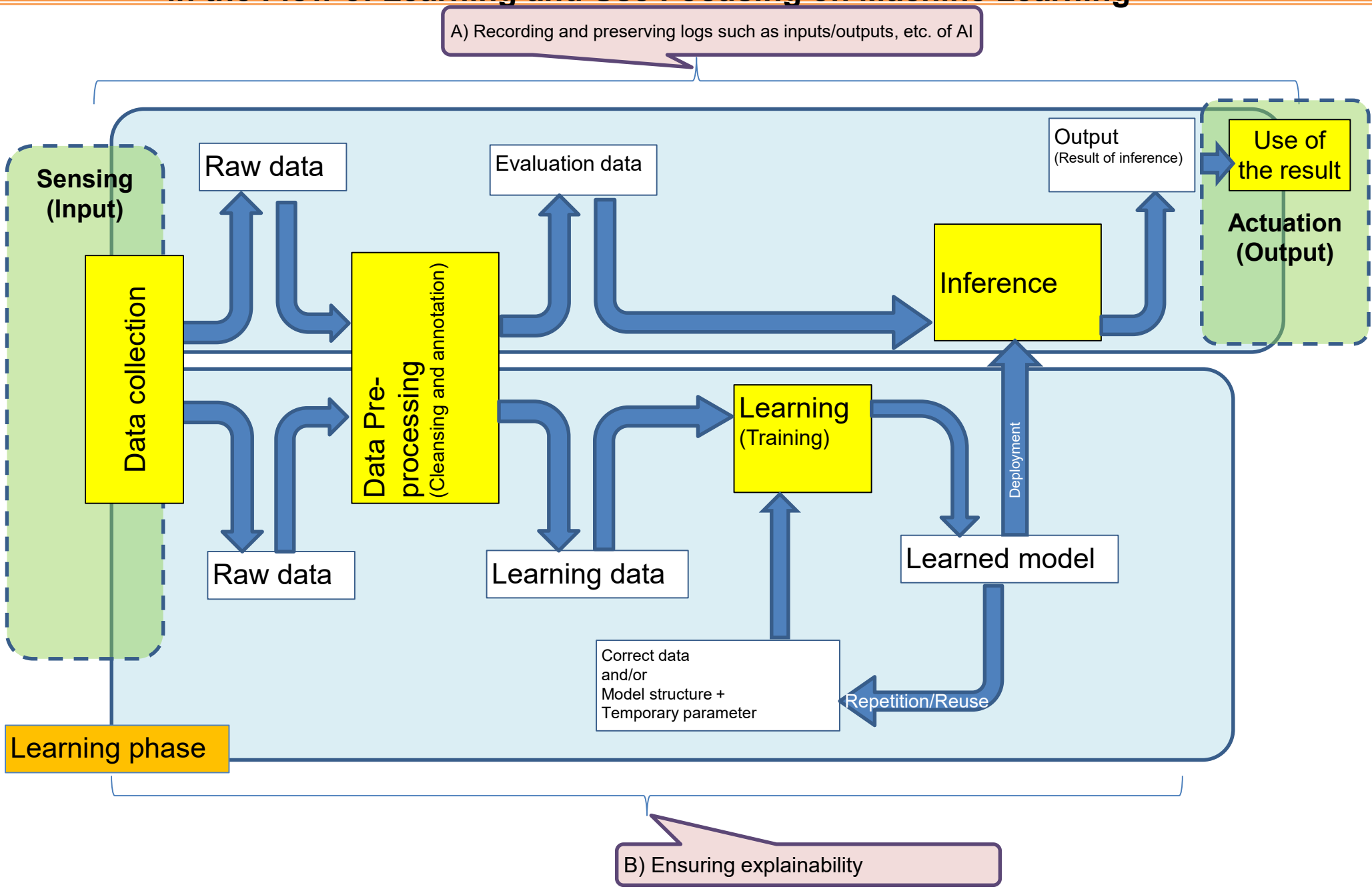
Key issues

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

Each point concerning 9) Principle of Transparency in the Flow of Learning and Use Focusing on Machine Learning



9-A) Recording and preserving logs such as inputs/outputs, etc. of AI

- AI service providers and business users are expected to record and preserve logs, including those on inputs/outputs, to ensure the verifiability of inputs/outputs from an AI. When recording and preserving logs, they are expected to consider the following items with consideration for the characteristics of technology to be used and its usage.

[Example Items to be considered in log recording and storage][2]

- The purpose of log recording and preservation (whether the purpose is to identify the causes of accidents or to prevent recurrences in fields that may harm humans' lives, bodies, property, etc.)[3]
- Frequency of log acquisition and recording
- Log accuracy
- Log retention period
- Log protection (Ensuring security, and integrity, etc.)
- Capacity of storage location
- Log time recording
- Scope of log to be disclosed

1) Scenarios that are expected to ensure input and output verifiability assume the case that make sure that end users are not using AI wrongly or with malicious intentions, in addition to the case to clarify the causes of accidents, if any.
2) The considerations described here may be in a trade-off relationship with each other, and it is necessary to consider the balance depending on the social context in using AI and the application of AI. For example, the frequency of log acquisition and retention period are in a trade-off relationship with ensuring the confidentiality and integrity of the logs.
3) In a field that may harm human lives, bodies, and property, it is highly necessary to identify the cause of the accident and then prevent recurrences. Therefore, it is also assumed that it is necessary to increase the frequency of log acquisition and recording, increase the accuracy of the log, and extend the retention period.

9-B) Ensuring explainability

- AI service providers and business users are expected to ensure the explainability of the judgment results by AI for the purpose of ensuring the trust of users and to present evidence of AI behavior with consideration of the social context in case of utilizing AI in a field that has a significant impact on individual's rights and interests. At that time, with consideration of the social context in the AI utilization, they are expected to ensure the explainability of the decision results made by AI by analyzing and understanding what kind of explanation is required and taking comprehensive measures in reference to the following measures.

[Example measures for ensuring explainability]

(Adopt AI software that incorporates an interpretable algorithm)

- Adopt an interpretable model[1] of AI software with high readability in advance.

(Adopt technical methods to explain the decision results by the algorithm)

- Adopt the following technical methods[2] that can explain a black-box model.
  - A global explanation method that replaces the model with an interpretable model, such as a model that makes the AI's prediction and recognition process readable.
  - A local explanation method that presents the basis of prediction for specific input, such as the presentation of key features or the presentation of important learning data and its expression in natural language.

(Manage data provenance)

- Manage when, where, and for what purpose data used for AI learning is collected (data provenance).

(Analyze the overall input and output trends of learning model)

- Analyze AI judgment trends based on combinations of multiple AI input and output (for example, observe changes in output when the input pattern is changed little by little).

(Comprehensive measures)

- With consideration of the needs and opinions of consumer users, clarify parts in which an explanation is lacking, and collaborate with developers to find out what kind of explanation is necessary[3].

1) In general, there is a tradeoff relation between making a learning models interpretable and maintaining the accuracy of AI's judgment.
2) To ensure technical explainability requires its implementation and verification. Therefore, in general, it is a trade-off relationship with the computational cost.
3) It is expected to lead to an essential solution to the issue of explainability by repeating the clarification of necessary explanations with developers' technology development related to the explanations alternately and then sharing of the corresponding technology widely.

9-C) Ensuring transparency when AI is used in administrative bodies

- When administrative bodies use AI, they are expected to ensure the explainability of the decision results made by an AI, according to the social context in the AI utilization with consideration of the reign of law, while ensuring administrative transparency, and keeping within the requirement of proper procedures. For improving its explainability, for example, the following methods are can be considered.

[Example measures to improve explainability]

- Including various social minorities in the development and design process of AI algorithms used by government agencies (co-design).
- Explaining the concept of the construction of learning data (the concept of inclusion and exclusion in learning data), policy judgments made at the stage of algorithm designing, and social impact assessment by introducing AI, and auditing methods for AI.
- Concluding contracts with developers or AI service providers in a form that limits the extent to which developers or AI service providers are able to not disclose a variety of factors that explain the AI's judgment.

10) Principle of Accountability*

Users should make efforts to fulfill their accountability to stakeholders.

Key issues

A) Efforts to fulfill accountability

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

* “Accountability” means the possibility to take appropriate measures, such as to proving the explanation behind the meaning and reason for the judgment, along with providing compensation as needed, in order to gain the understanding of the person who is affected by the result of the judgment.

10-A) Efforts to fulfill accountability

- With consideration for the purpose of the Utilization Principles (1) to (9) described in these Guidelines in order to earn the trust of AI from people and societies, AI service providers and business users are expected to strive to fulfill the corresponding accountability to consumer users and third parties affected by AI utilization. Therefore, based on the nature and purpose of the AI to be used, they are expected to provide and further explain information on the characteristics of the AI system, and communicate with various stakeholders according to their knowledge and capability.

<Reference>

- *Consumer users are recommended to strive to fulfill their accountability according to their knowledge and ability.*

- *If consumer users have any doubts in the decision made by AI, they are recommended to contact developers, AI service providers, and business users as required.*

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

- AI service providers and business users are expected to create, publish and disseminate AI usage policies as described below so that consumer users and others can appropriately recognize the utilization of AI.

i. To create and publish an AI usage policy so that consumer users and third parties are aware of the use of AI when the judgment of an AI could directly affect them, and to provide notifications to them when asked.

ii. Regarding (i), to proactively provide notifications to consumers and third parties in case their rights and interests may have been seriously affected concerning (i)[1]

[Example matters to be described in an AI usage policy]

- Use of AI (if specific functions or technologies can be identified, their names and contents[2])
- Scope and method of AI utilization
- Risks associated with AI utilization
- Consultation counter

- They are expected to publish or notify them not only before use of an AI is started but after its behavior changes or use of it is terminated (especially when assumed risks are changed due to a change in the AI's behavior).

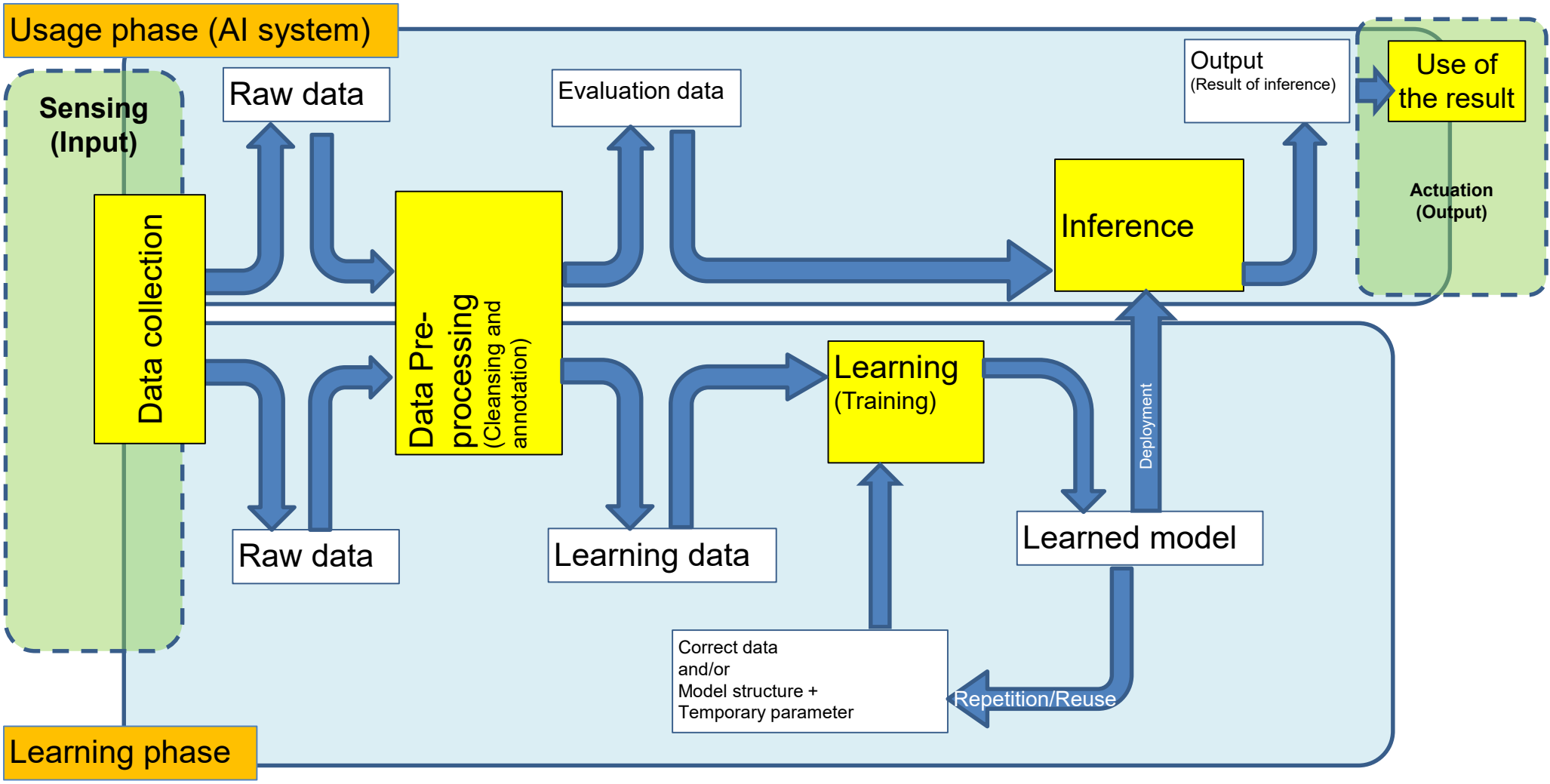
1) It is considered that AI service providers and business users are required to publish usage policies related to AI if the judgment of AI to be used directly affects consumer users and third parties. In other words, if AI is only used as an analytical tool for human thinking, or if AI is making a draft, but it is practically guaranteed that humans will ultimately judge, it is not always required to announce the usage policy regarding AI. (However, even in such a case, it is recommended that the announcement should be voluntarily published.)
2) It is expected to announce the proper utilization methods and the risks associated with utilization along with the functions and characteristics of the AI. For example, "The service XX is judged by the deep-learning model and cannot guarantee 100% of the expected behavior (therefore please be careful)".

<Reference>

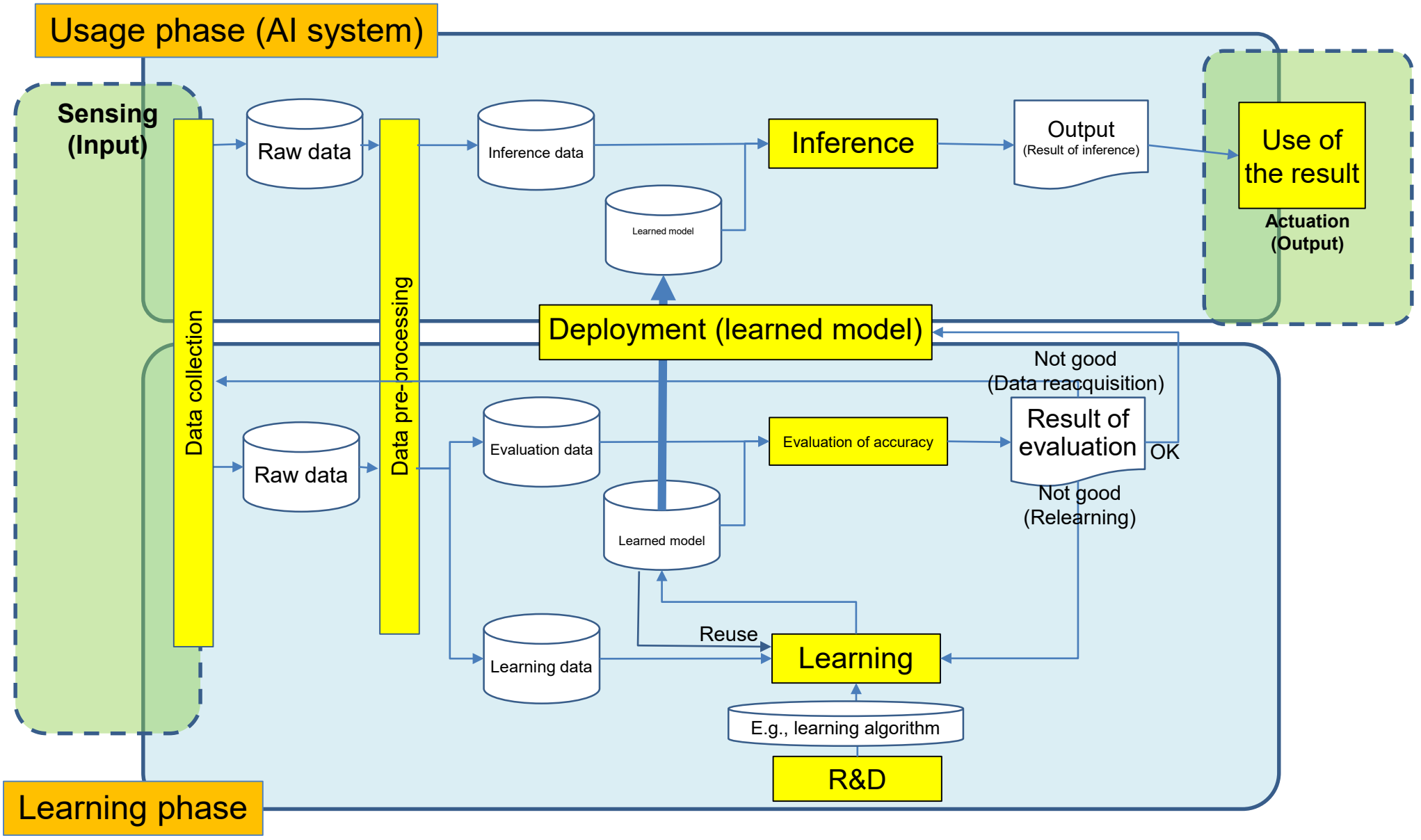
- *If consumer users have any doubts in the decision made by AI, they are recommended to contact developers, AI service providers, and business users as required.*

(Reference 1) Outline of AI System

The Flow of Learning and Use Focusing on Machine Learning (Simple Version)



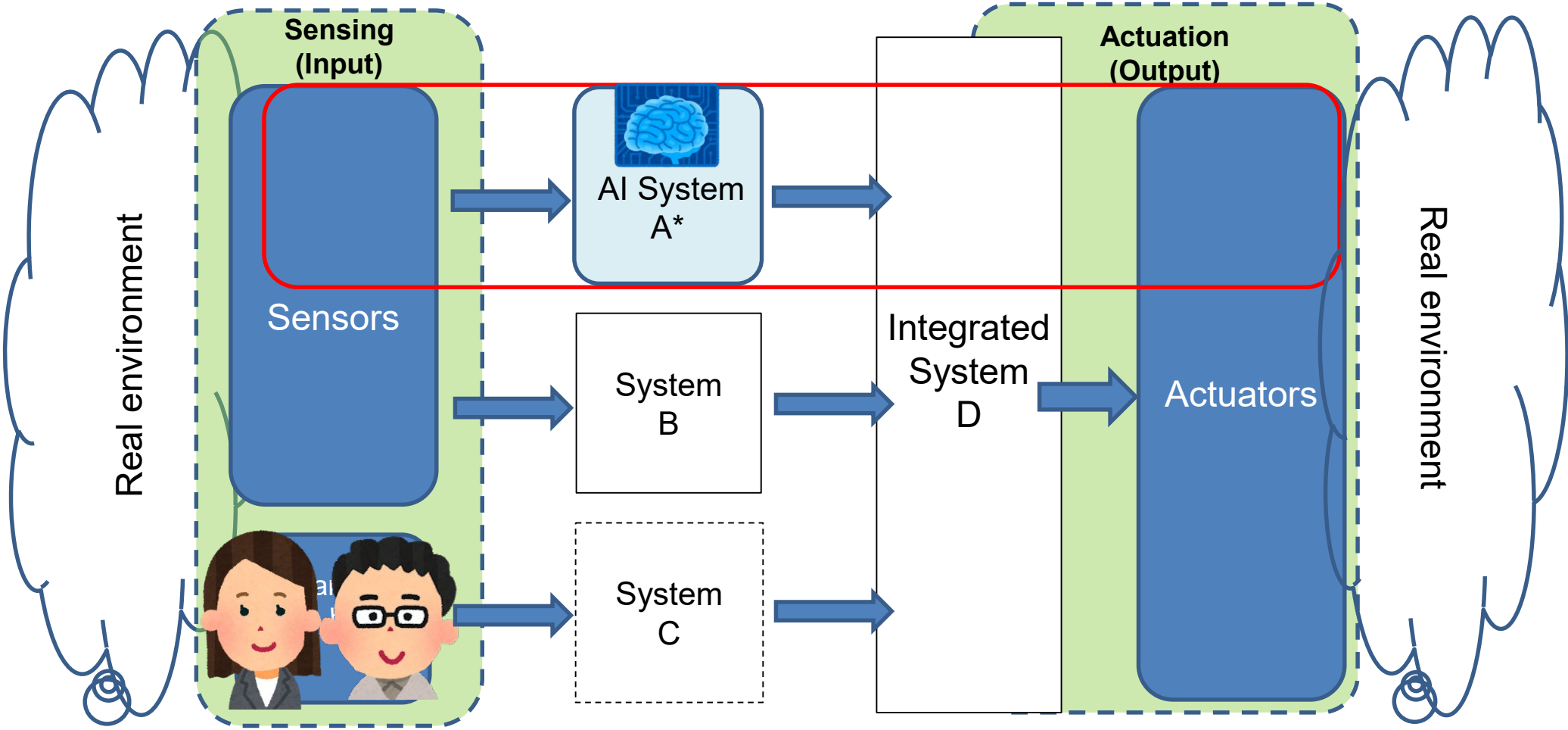
The Flow of Learning and Use Focusing on Machine Learning



An Example of AI System (Integrated with Other Systems)

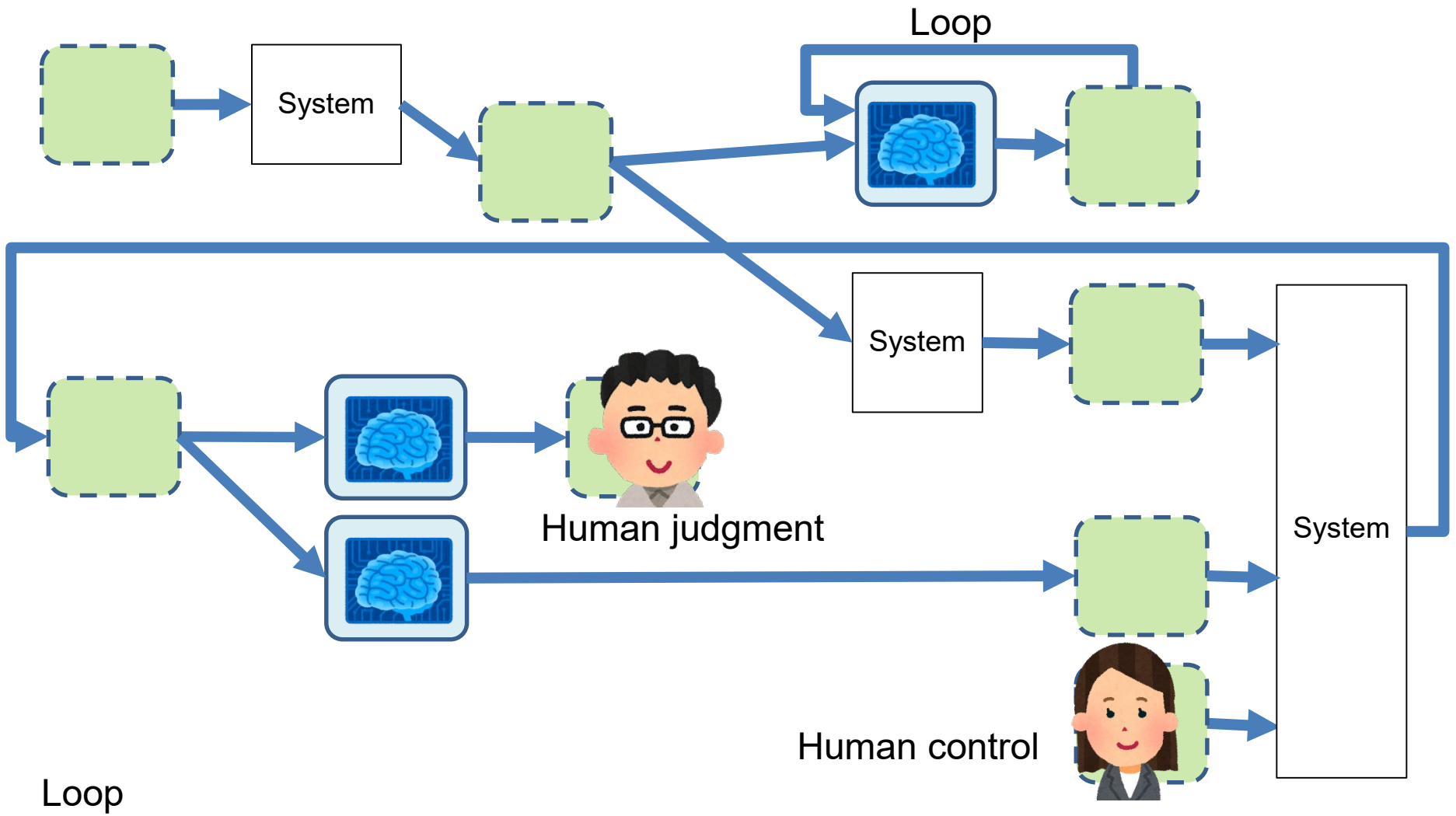
Results obtained from an actual environment through sensors, and then the results processed by AI systems A and B, and those processed by human beings received by system C, are integrated into system D, and these results are reflected in the control of actuators that affect the real environment.

Example: Automatic driving is usually based on sensed peripheral image information. AI System A, which makes inferences about running and stopping the automobile, instructs integrated system D. On the basis of the instruction, actuators related to the operation of autonomous driving is controlled. A kill switch (a switch to stop the automobile safely) is provided in System C in order to respond to abnormal situations, including situations when the driver senses danger. In response to forced stop input from humans, integrated system D instructs the actuator to stop the autonomous vehicle safely.



*Equivalent to the use phase (AI system) in the previous slide

The AI System Cooperates with Other (AI) Systems



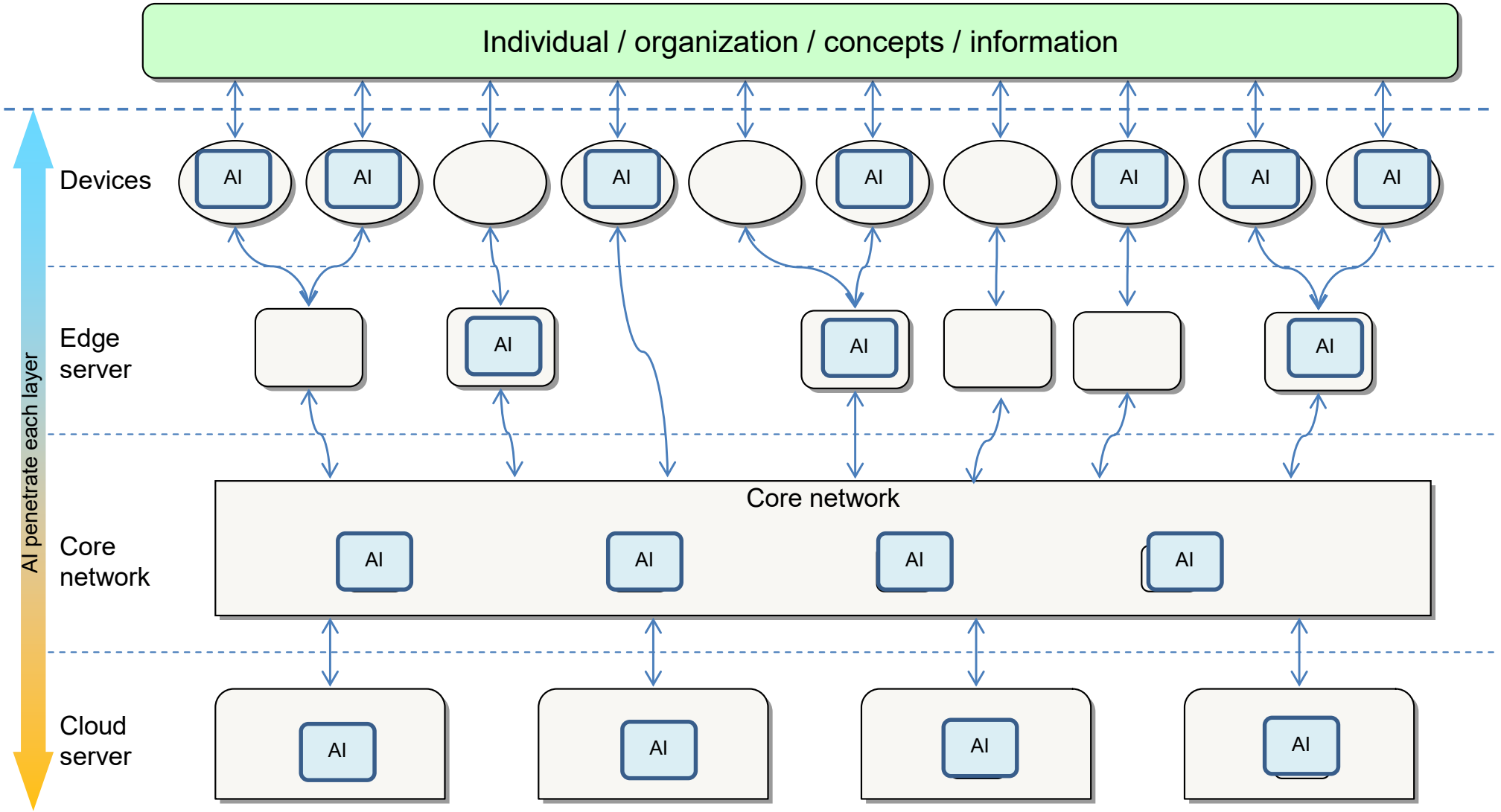
System General system

AI System

Input: Sensor information, human control, and judgment

Output: Judgment information, equivalent to actuation based on the judgment information

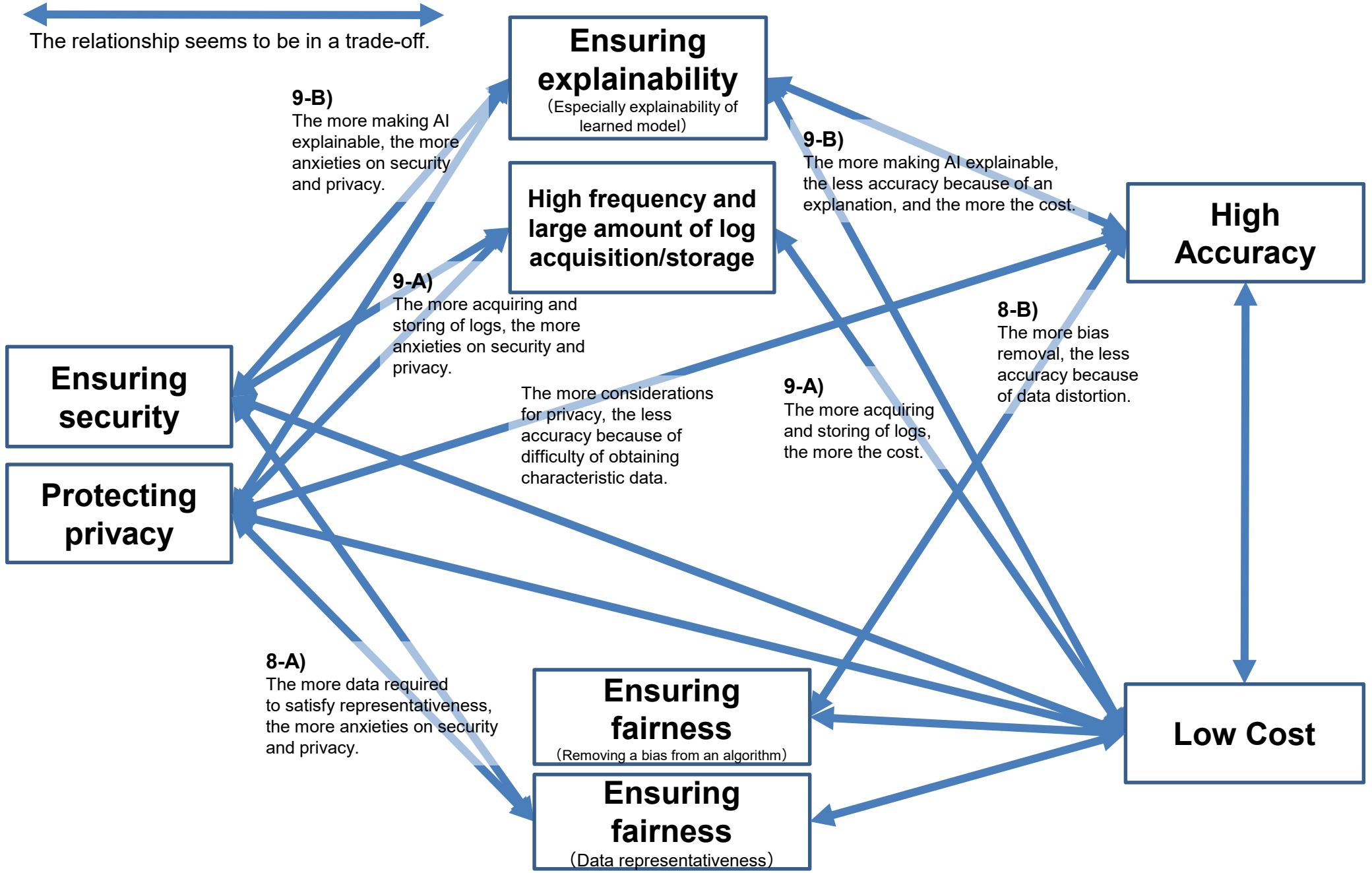
AI (systems) Penetrate into Each Layer of Information and Communications Network Systems to Achieve Cooperation and Coordination with Each Other



Cited from: Report of Conference on the Networking among AIs—Wisdom Network Society (WINS) Produced by the Networking among AIs —Toward a society that goes beyond the 4th Industrial Revolution—

(Reference 2) Examples of Trade-offs for Each Item

Examples of Trade-offs

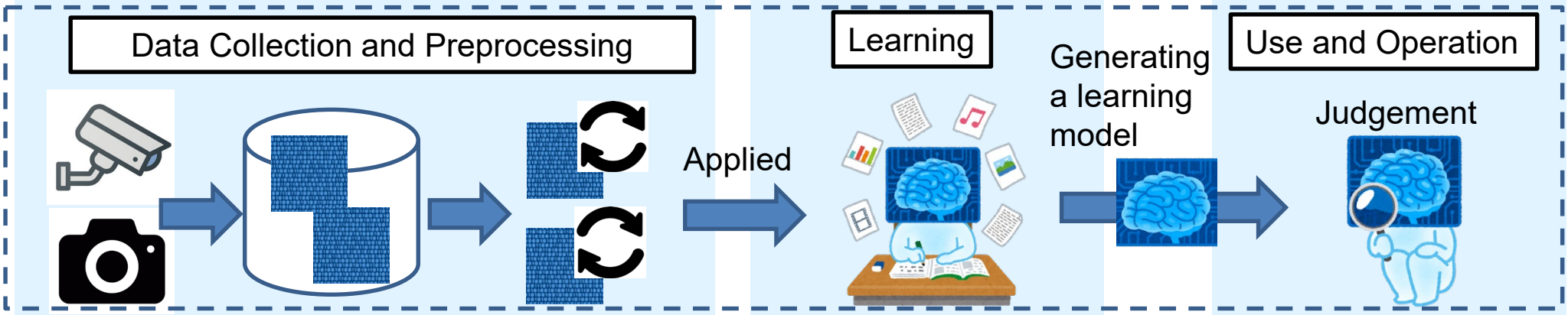


Reference 3: Application Example of the Principles

AI Utilization Guidelines (Example)

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

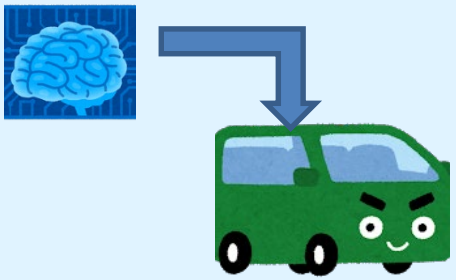
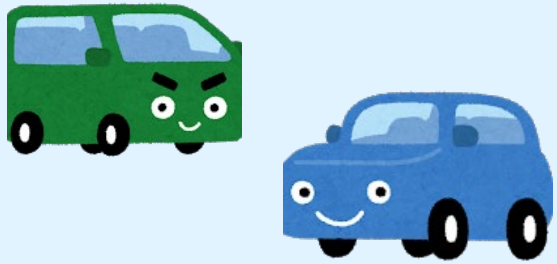
[Fairness]
Attention to unfair discrimination by algorithm:
Not to discriminate in the computing process

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

AI Utilization Guidelines (Example)

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

Ex. Autonomous driving

Build system with AI	Deployment	Use and Operation
		
[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