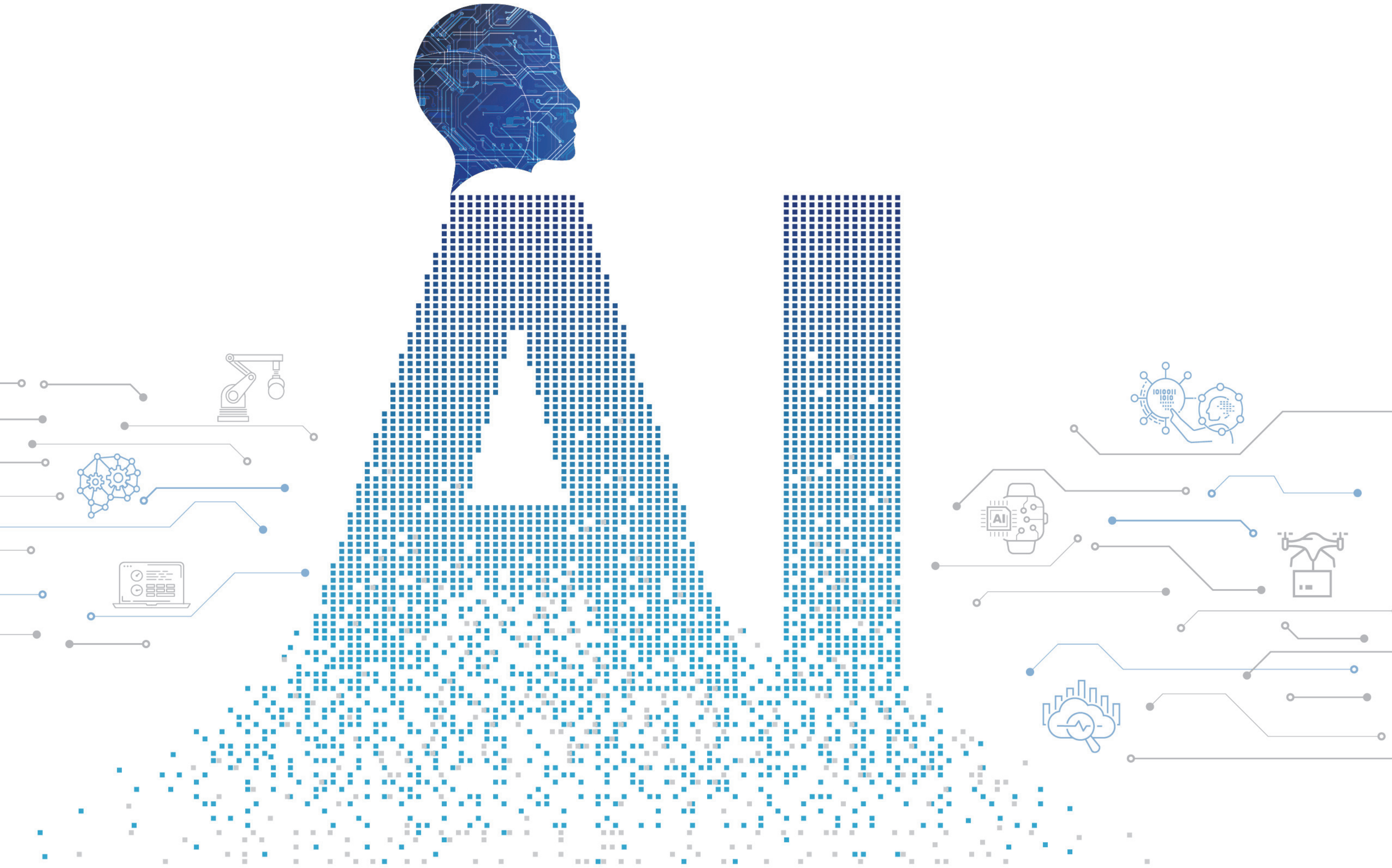
2022

Trustworthy AI Development Guidebook

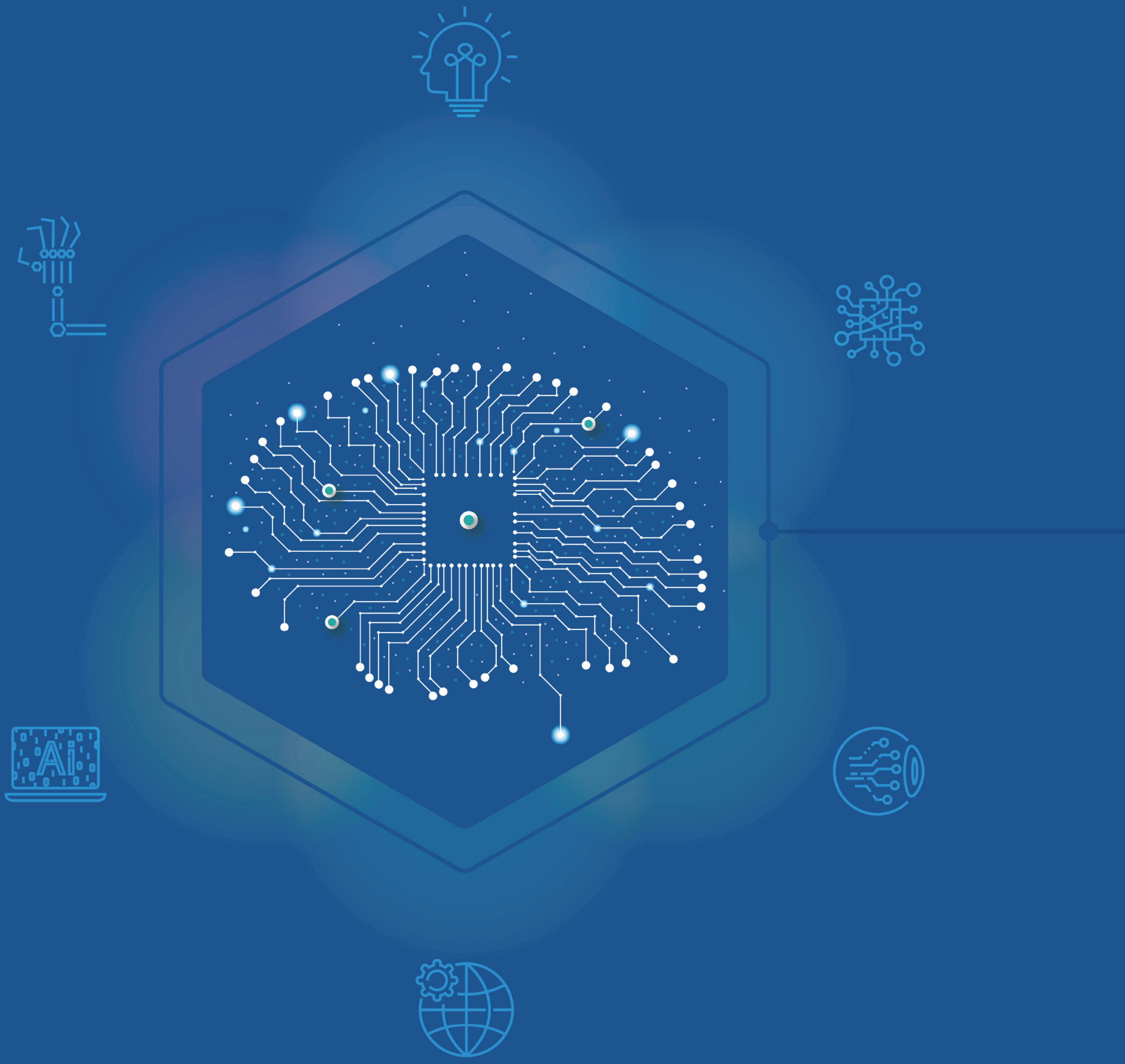


Ministry of Science and ICT

Telecommunications Technology Association

CONTENTS

2022 **Trustworthy AI** Development Guidebook



PART 1

Introduction

1 Introduction

01 Publication background and purpose

Artificial intelligence (AI) technology is used in various fields, including traditional areas such as gaming, simple applications including personalized assistants and voice recognition using speakers, and complex fields such as medical examinations and disease diagnosis, financial services including asset management, and the autonomous driving and operation of automobiles and drones, respectively. As the fields of application expand and have an increasing impact on daily life, securing the trustworthiness of AI has emerged as an important task. Such importance is due to the likelihood of AI creating errors from contaminated data or increased biases occurring from the frequency of AI use and an increase in the number of data, as well as the increased difficulty in understanding the operating principle or mechanism as AI becomes more advanced. In particular, the importance of trustworthiness of AI has grown as its applications are extended to fields directly related to human life and public safety.

As such, various countermeasures are being prepared internationally as the trustworthiness of AI emerges as a global concern. The Organization for Economic Cooperation and Development (OECD) published the “Recommendation of the Council on Artificial Intelligence (May 2019),” which is a recommendation for securing trustworthy AI. In addition, the European Commission (EC) published “The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment (Jul. 2020),” which is a list that allows personnel to verify the trustworthiness of AI on their own. In line with this, Ministry of Science and ICT. of Republic of Korea published the “National AI Ethics Standard (Dec. 2020)” with the goal of realizing human-centered AI.

However, the trustworthy AI principles and standards, which have been released thus far, including those mentioned above, mainly present abstract items from an ethical perspective, and thus, there is a difficulty in applying them in practice. There is a need for a list of verifications that is sufficiently detailed for AI service developers to apply in the field without any additional thought or reasoning process. In particular, R&D personnel of small and medium-sized enterprises with limited human resources and R&D investment capacity have difficulty in directly devising and applying trustworthiness requirements and verification methods.

A trustworthy AI development guidebook was designed to address this issue. The recommendations and guides published by major advanced countries such as the United States (U.S.) and Europe, and by international organizations, were referenced, and 14 requirements and 59 qualitative and quantitative verification items that can be autonomously checked are presented herein.

AI service development personnel such as developers and planners will be able to apply the items presented in this guidebook as they are, securing the minimum trustworthiness, and understand what is important for securing trustworthiness. Further, AI services with a high level of trustworthiness can be developed by devising and utilizing items and verification methods suitable for the field based on the contents of the guidebook. By applying this guidebook, Korean AI-related companies and institutions are expected to contribute to securing more mature AI technology and serve as a solid foundation for global competitiveness.

02 Trends of trustworthy AI

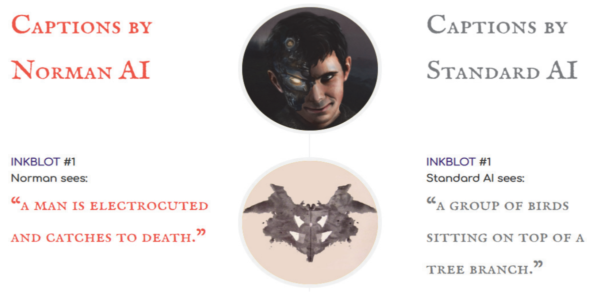
Major national governments, standards organizations, and technical groups around the world are presenting measures tailored to their respective circumstances to secure the trustworthiness of AI. This section examines the problems that arise as AI is widely used, the basic concept of trustworthy AI, and related policies and research trends both at home and abroad.

2.1. Problems with the spread of AI

As new technologies such as the Internet and smartphones continue to change our daily lives and society as a whole, new threats are emerging as AI is being increasingly applied in various fields throughout society and industry. Because many of these problems are either social or ethical based, they cannot be addressed by simply improving or introducing new technologies because the more technology affects people's daily lives, the more people need to redefine the ethical and social issues of which they are normally unaware in accord with such technology. In fact, there are currently many incidents in which AI technology is used for malicious purposes or for creating anti-social and anti-humanistic sentiments.

AI incidents

〈Incident 1: Psychopath AI〉


Captions by Norman AI
Captions by Standard AI
INKBLOT #1
Norman sees:
"A man is electrocuted and catches to death."
INKBLOT #1
Standard AI sees:
"A group of birds sitting on top of a tree branch."

MIT developed "Norman," which was trained by intentionally using an anti-social and anti-humanistic dataset. During a Rorschach test, which is a psychological test using pictures, negative recognition results were derived, including understanding the picture presented as that of a person who had been electrocuted to death (Jun. 2018).

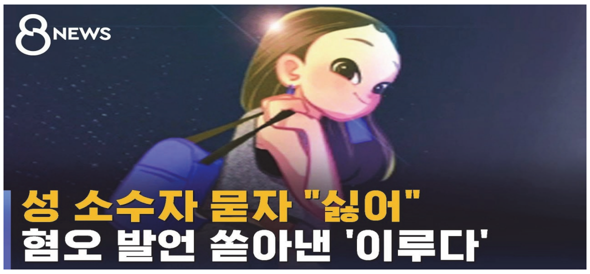
Implication
AI trained based on biased data can form a model that is socially unacceptably biased.

〈Incident 2: Self-driving car fatality〉

A self-driving Uber vehicle struck and killed a pedestrian who was jaywalking. Although detected through the LIDAR sensor, the pedestrian was considered an obstacle on the road that could be ignored and run over owing to the priority given to driving efficiency (May 2018).

Implication
AI worked accurately and efficiently, but resulted in a human casualty resulting from a lack of ethical judgment that human safety comes first.

〈Incident 3: AI chatbot controversy〉



The topic conversational chatbot “Iruda,” developed by SCATTER LAB, was released with a setting that mimics a woman in her 20s. It caused a controversy over remarks regarding the LGBT community, privacy violation, and sexual conversations with some users (Jan. 2021).

Implication
Ethics in securing datasets and bias in learning data caused a social scandal.

〈Incident 4: Deepfake video〉

Using a deep-learning method, the face and voice of former U.S. President Obama were synthesized to create a fake video insulting the current President Trump (Jul. 2018).

Implication
Created fake news and manipulated a video, causing social confusion and public criticism.

2.2. Concept of trustworthy AI

As explored in the aforementioned cases, AI technology projects should be reviewed not only from the technical aspect of “Can it be implemented?” but also from the ethical and social aspects of “Is it okay for this project to exist?” In particular, because AI is used in various fields, it can have an extremely large ripple effect if used without recognizing the ethical flaws in AI systems and learning models. “Trustworthy AI” refers to the value standards that must be adhered to in order to address the risks and limitations inherent in AI technologies, such as data and model bias and “black box” characteristics, as well as to avoid adverse effects in the process of using and disseminating AI. Major organizations are discussing the essential elements for securing trustworthy AI. In general, safety, explainability, transparency, robustness, and fairness have been mentioned as essential factors.

Essential factors of trustworthy AI

Essential factors	Meaning
Safety	When the system operates or functions as a result of an AI judgment and prediction, the possibility of adverse effects on people and the environment is mitigated and eliminated.
Explainability	A state in which the basis for AI judgment and prediction and the process leading to the result are presented in a way that humans can understand, or when a problem occurs, the result that leads to the problem can be traced and drawn.
Transparency	The degree to which the rationale and operation process of essential factors such as safety, explainability, robustness, and fairness of AI conform to universal rationality.
Robustness	The state in which AI maintains the user’s intended level of performance and functionality, even under external interference or an extreme operating environment.
Fairness	Functionality that does not indicate discrimination or bias toward a specific group or draw conclusions containing discrimination or bias in AI data processing.

※ Privacy and sustainability are also being discussed in various ways as essential factors.

References Concept of trustworthy AI under discussion by prominent institutions

- (International Organization for Standardization, ISO) Availability, resiliency, security, privacy, safety, accountability, transparency, and integrity are presented as detailed attributes of trustworthiness (ISO/IEC TR 24028, 2020).
- (Organisation for Economic Co-operation and Development, OECD) AI must have transparency, explainability, robustness, and safety in line with a sustainable society and human-centered values (2019).
- (National Institute of Standards and Technology, NIST) Trustworthiness is a goal that must be satisfied when AI is used for social benefit and economic growth, and is a concept that includes explainability, safety, and security (2020).
- (European Community, EC) The use of AI must be legal, ethical, and technically and socially sound (2019).

2.3. Domestic and foreign trustworthy AI policies and research trends

Major countries and communities such as the EC and the U.S. consider securing trustworthiness in AI as a prerequisite for its social and industrial acceptance and development, and are pursuing policies in this regard. Moreover, studies on securing trustworthiness, centering on the development of related technologies, are active in industry and academia. Specifically, the policies and norms required to secure trustworthy AI are being prepared in earnest at the level of governments of major countries such as the EC and the U.S. In addition, major countries have specified Trustworthy AI and Safe AI as key elements of a national AI strategy. Furthermore, they are striving to create an environment in which the private sector can autonomously check and secure the trustworthiness of AI by preparing appropriate guidelines. In the technology field, academia and global companies in major countries such as the U.S. and Europe are developing technologies necessary to secure trustworthy AI. Republic of Korea also recently announced “AI ethical standards (Dec. 2020)” and “Strategy for trustworthy AI (May 2021)” and is participating in global movements in terms of both policy and R&D.

Trustworthy AI-related policy trends in major countries

Country	Main policy (year)	Characteristics
EC	• Ethics Guidelines for Trustworthy AI (2019) • Declaration on Cooperation on Artificial Intelligence (2018)	Aiming for balanced AI policies such as human-centered values, ethics, and security
U.S	• Guide for Regulation of Artificial Intelligence (2020) • Guidelines for Autonomous Driving (2016-2018) • AI Trilogy Report of Obama Administration (2016)	Focusing on AI technology development support and deregulation policy for AI utilization and promotion by industry sector
China	• New Generation Artificial Intelligence Development Plan (2017)	Business-friendly policies such as government-led large-scale investment, rigorous training of human resources, and data openness and sharing
Japan	• AI Utilization Guidelines (2019) • Social Principles of Human-centric AI (2018) • AI R&D Guidelines (2017)	Comprehensive approach from economic, industrial, social, and ethical perspectives
Republic of Korea	• Human-centered AI Ethical Standards (2020) • National AI Strategy (2019)	Implementing comprehensive policies such as building an AI ecosystem, nurturing talent, expanding industry, and preventing dysfunction, with human-centered AI as a basic value

Trends of trustworthy AI research by major overseas industry, academia, and research institutions

Name of institution	Activities and content
DARPA	Researched the safety, reliability, and explainability of AI system through projects such as Assured Autonomy; in addition, eXplainable AI is in progress.
Stanford University	Researched formal verification techniques to ensure safety of AI, security of learning and control, and methods of securing transparency.
IBM	Under the motto of "Trusted AI," announced the five principles of fairness, alignment of values, robustness, explainability, transparency, and accountability as related measurement and evaluation tools.
Microsoft	Defined six AI development principles including fairness, trustworthiness and safety, personal information protection, and security for responsible AI development and service provisioning.

03 Development process of trustworthy AI development guidebook

This trustworthy AI development guidebook was prepared by carefully analyzing the movements of the international community on recent AI-related incidents and major issues. As stated in the background of the publication, although many domestic and foreign institutions and companies have published ethical principles, standards, and guidelines for securing trustworthy AI, there has yet to be a case in which a detailed methodology was presented from a technical perspective. Accordingly, this guidebook was planned in such a way that data scientists, model developers, system and software developers/verifiers, designers, planners, and service operators who apply practical tasks in the field of AI service development can refer to the methods and standards for securing trustworthiness from a practical perspective and apply them to their particular field. Accordingly, the development process of this guidebook started with identifying the elements necessary to secure trustworthiness from technical and engineering perspectives.

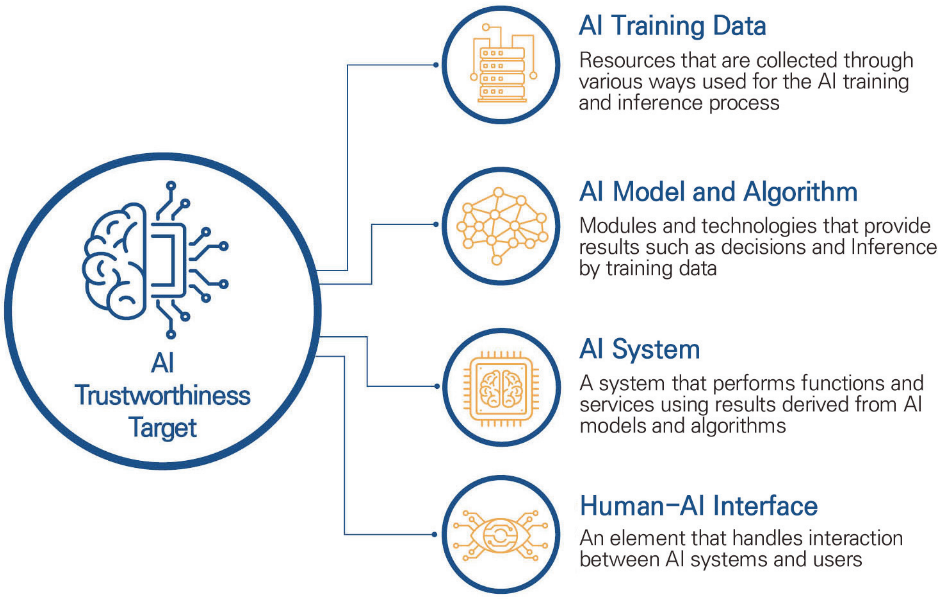
3.1 Development guidebook design elements (AI service composition, lifecycle, and trustworthiness requirements)

An AI service is mainly composed of four elements: AI model and algorithm, data to learn such models and algorithms, a software-based system to be applied to an actual function, and an interface for interacting with users as needed. They are developed, validated, and operated according to the lifecycle of the AI services, either individually or in integration. Thus, verification items that can confirm whether they are properly applied should be provided along with the technical and engineering requirements according to the composition and lifecycle of the AI services for developers.

1 Introduction

First, the methods of securing trustworthiness for the four elements constituting the AI service are as follows.

AI service composition



AI Trustworthiness Target

AI Training Data
Resources that are collected through various ways used for the AI training and inference process

AI Model and Algorithm
Modules and technologies that provide results such as decisions and Inference by training data

AI System
A system that performs functions and services using results derived from AI models and algorithms

Human–AI Interface
An element that handles interaction between AI systems and users

AI service composition	Methods of securing trustworthiness
Training data	verify that bias and fairness are excluded from data used in AI training and the inference process
Model and algorithm	verify whether AI derives safe results according to the model and algorithm and can explain and whether it is robust to malicious attacks
System	verifies whether the entire system to which the AI model and algorithm are applied works according to the AI–derived result, and whether countermeasures exist in the case of errors
Human–AI interface	verifies whether AI system users can understand the operation of the AI system and whether it prevents human error and AI malfunctions

Second, the AI service lifecycle can be defined as follows.

AI service lifecycle

Lifecycle phases	Key actors	Main activities
1. Planning and design	• Business decision-maker • Data scientist • System operator	- Define business models and KPIs - Collect requirements according to the purpose of the AI system and the entire lifecycle - Proof of AI system concept and define required resources
2. Data collection and processing	• Data provider • Data scientist • Domain expert	- Prepare measures to secure data quality and provide information for data users to understand - Establish data access control and de-identification policies such as personal data protection - Data labeling and documentation of dataset characteristics - Prepare dataset for building AI model
3. AI model development	• Data scientist • AI model developer • System engineer	- Implement AI model according to business purpose - Confirm and verify the implemented AI model - Tune AI model, analyze data, and consider additional data needed - Evaluate the performance of the final AI model
4. System implementation	• AI model developer • System engineer	- Verify compatibility with legacy system - Test functional unit, validate system, approve deployment version - Perform pilot test of AI system
5. Operation and monitoring	• Business decision-maker • AI model developer • System operator	- Ensure performance through system monitoring and AI model retraining - Monitor system trustworthiness such as model bias detection, fairness, and explainability - In the case of a fatal problem, make a decision to dispose off the system

The AI service lifecycle refers to the process of implementing and operating the components discussed in the first section. It is similar to the engineering process or lifecycle addressed in existing software systems, but separate data processing and model development steps are required owing to the nature of AI. The definition of the main activity is slightly different in the other phases. In much of the literature, the lifecycle of AI or an AI service is currently defined through a division into six to eight phases. A representative includes a lifecycle presented by the OECD and ISO/IEC. This guidebook simplifies this into five phases without distorting the nature and activities of each lifecycle phase such that the personnel can easily apply it by referring to the lifecycle presented by the two organizations as representative examples.

The phases of an AI service lifecycle are repetitive and cyclical and may not be sequential. This guidebook explains steps 1–5 sequentially to help with the understanding; however, the sequence may in fact have no meaning in the process of collecting and processing actual data or developing and operating a model.

Third, to define the requirements for trustworthy AI, by applying the core requirements of the "Ethical standards for AI," the following four requirements for the verification items were derived along with the requirements from a technical viewpoint.

1. Respect for diversity   2. Accountability   3. Safety   4. Transparency

International organizations for standardization such as EC, OECD, IEEE, and ISO/IEC have presented trustworthy AI by subdividing its sub-attributes. Moreover, in academia and industry, the sub-attributes constituting trustworthiness form a separate discipline. In particular, ISO/IEC 24028:2020 provides keywords in the form of considerations needed to ensure trustworthiness. These include transparency, controllability, robustness, resilience, fairness, safety, privacy, and security, whereas the relationship between keywords or the relationship with trustworthiness was not defined. As such, terms that are similar but slightly different depending on the viewpoint are being defined in various documents and there is no agreed definition yet. The attributes and keywords presented by various organizations such as EC, OECD, IEEE, and ISO/IEC were comprehensively analyzed, and a consensus was sought by collecting the opinions of experts from the domestic academic circles, research circles, and industry. After deriving trustworthy AI attributes through such a broad process of sharing opinions, they were matched with the 10 requirements of the National AI Ethical Standards, and requirements that can be addressed in the technical aspect were then selected.

3.2 Derivation of requirements and verification items for securing trustworthy AI

As the next step, specific requirements and verification items were derived. First, the technical requirements to be complied with based on the policies, recommendations, and standards announced by standards organizations, technical groups, international organizations, and major national governments for securing trustworthy AI were deduced and clearly presented. In addition, the cases announced for the purpose of securing trustworthy AI in Republic of Korea, such as the AI Personal Information Protection Autonomous Checklist (May 2021) and the Financial AI Guidelines (Jul. 2021) were examined. Through this process, important contents were reflected in the guidebook and duplicate contents were removed or reduced. The relevant references are as follows.

Key references of trustworthy AI

Name of institution	Date of publication	Recommendations and standards
Republic of Korea	Nov. 2020	National AI Ethics Standards (Draft)
EC	Jul. 2020	The Assessment List for Trustworthy Artificial Intelligence
ISO/IEC	Mar. 2021	ISO/IEC TR 24029-1:2021, Artificial Intelligence (AI) - Assessment of the robustness of neural networks - Part 1: Overview
	Jan. 2021	ISO/IEC 23894, Information Technology - Artificial Intelligence Risk Management
	Nov. 2020	ISO/IEC 24027, Information technology - Artificial Intelligence (AI) - Bias in AI systems and AI aided decision making
	May 2020	ISO/IEC TR 24029:2020, - AI - Overview of Trustworthiness in artificial intelligence
Google	May 2019	People + AI guidebook
European Telecommunication Standards Institute (ETSI)	Mar. 2021	Securing Artificial Intelligence (SAI 005) - Mitigation Strategy Report
OECD	May 2019	Recommendation of the Council on Artificial Intelligence
World Economic Forum (WEF)	Jan. 2020	Companion to the Model AI Governance Framework

The requirements derived through this are illustrated in the table below, and the results of responding to the core requirements of AI ethics are shown.

Results of matching technical requirements and ethical factors for securing trustworthy AI

Requirements		Respect for diversity	Accountability	Safety	Transparency
Requirement 01	Planning and implementation of risk management for AI system		✓		✓
Requirement 02	Providing detailed information for data utilization		✓		✓
Requirement 03	Removal of abnormal data to ensure data robustness			✓	
Requirement 04	Removal of bias in collected and processed training data	✓	✓		✓
Requirement 05	Ensuring security and compatibility of open-source library		✓	✓	
Requirement 06	Removal of bias in AI model	✓			
Requirement 07	Establishment of response countermeasures against AI model attacks			✓	
Requirement 08	Providing AI model specifications and explanations of outputs		✓		✓
Requirement 09	Providing confidence value for AI model output				✓
Requirement 10	Removal of bias that may occur when implementing AI system	✓			
Requirement 11	Implementation of safe mode of AI system		✓	✓	✓
Requirement 12	Fostering user's understanding of AI system instructions				✓
Requirement 13	Securing traceability of AI system			✓	✓
Requirement 14	Providing the service coverage and the target of interaction		✓		✓

3.3 Collecting opinions from industry, university, and research personnel

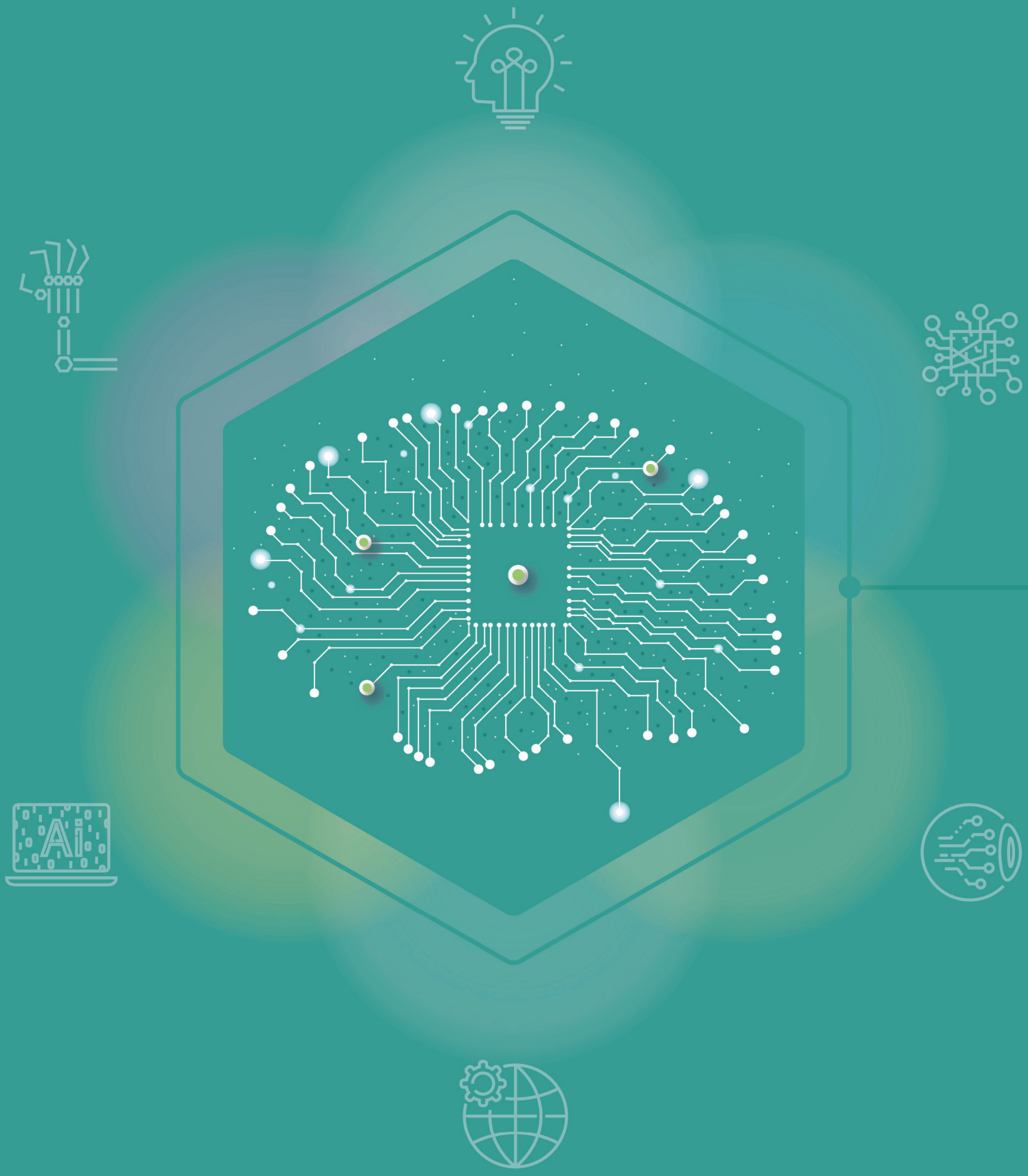
After selecting the requirements for securing trustworthiness, each item was examined from a technical viewpoint, and the guidebook was upgraded to fit the perspective of the personnel in the field. This review included aspects of technical feasibility, utility, and comprehensiveness. Whether each detailed check item meets the requirements (validity), whether it is practically usable at the development site (utility), and whether the content for verification covers a wide range of research content from the past to the present (inclusiveness) were validated. To this end, a number of AI experts directly reviewed the content through participation and were reviewed, and their various opinions were collected and considered. Regarding the AI experts, a variety of researchers from industry and academia, including corporate planners, development project leaders, professors, lead researchers of a national research institute, and related national policy officers were recruited.

04 Target of trustworthy AI development guidebook

The trustworthy AI development guidebook targets stakeholders, including all organizations and individuals who are directly or indirectly involved in or influence the process of implementing AI services. In particular, the main subjects are planners, data collectors and processors, AI model developers, system and software developers, and testers who need to manage trustworthiness from a technical viewpoint. The key requirements they should review to ensure the trustworthiness of AI at each phase of the AI lifecycle are as follows.

Requirements for securing trustworthiness at each phase of the AI lifecycle

Lifecycle phases	Key actors	Requirements
1. Planning and design	• Business decision–maker • Data engineer • System operator	– Review the requirements for securing trustworthiness throughout the entire lifecycle of the AI system and establish an application plan
2. Data collection and processing	• Data provider • Data engineer • Domain expert	– Secure management measures for data errors and biases that may occur in the process of acquiring training data
3. AI model development	• Data engineer • AI model developer • System engineer	– Establish countermeasures against biased output of learning model or attacks – Provide an interpretation method for the output of the learning model
4. System implementation	• AI model developer • System engineer	– Prepare countermeasures for possible biases or errors in AI system development – Provide user–friendly explanation for the results of AI services
5. Operation and monitoring	• Business decision–maker • AI model developer • System operator	– Establish a countermeasure by tracing the cause in the event of an AI system problem



PART 2

Requirements

Table of Contents for Requirements

1 Planning and Design

2 Data Collection and Processing

3 AI Model Development

Table of Contents for Requirements

1 Planning and Design

Requirement 01 Planning and implementation of risk management for AI system

Accountability Transparency

▶ It is necessary to prepare response measures against risk factors that may arise during the implementation and operation of AI systems (e.g., model misidentification, functional misoperation, and security and personal information issues) by identifying those factors beforehand and analyzing their scale (gravity and impact).

01-1 Have you analyzed the risk factors that may occur during the lifecycle of the AI system?

[Yes I No I N/A]
□ □ □

- There are four categories of risk management: risk identification, risk analysis, risk evaluation, and risk treatment. To secure the reliability, the four activities are maintained and repeated during all phase of the AI lifecycle for risk elimination and prevention. The concept, definition, and overall flow of risk management are introduced in ISO 31000:2010.
- However, methods of identifying, analyzing, and evaluating risks that may hinder the process of securing AI reliability potentially differ from existing software- and hardware-based systems. Categories of risk factors that should be examined from the perspective of AI reliability are presented in ISO/IEC 24028:2020 and ISO/IEC 23894.2.
- After identification of risk by factor types, the causes, circumstances, and conditions where risk factors may occur are analyzed. Then, the size of the impact that the risk factors may have on the AI system, people, and surrounding environment is analyzed.

01-2 Did you take measures to eliminate, prevent, or minimize the effects of the risk factors?

[Yes I No I N/A]
□ □ □

- Response measures should be prepared according to the risk factors analyzed in 01-1. Eliminating the cause of risk factors can not only prevent human casualties and accidents, but also minimize the spreading effects or negative impacts from such accidents.
- Response measures refer to all technical methods that may be applied in processes such as execution and operation means, software and hardware functions, and model learning methods and strategies. Response measure classifications are presented in ISO/IEC 24028:2020.
- All parties interested in implementing AI should take the aforementioned into consideration to prepare response measures against risk factors and verify whether the risks have been eliminated or minimized.

2 Data Collection and Processing

Requirement 02 Providing detailed information for data utilization

Accountability Transparency

▶ A dataset for training AI may collect new data during the development process, which may be used as training data for other similar systems. However, if there is insufficient information on collected data, such as the source and characteristics, it may be difficult to reuse the data or identify the cause behind a problem brought on by the data. Detailed data information should therefore be provided so that the collected data may be properly used, and the cause behind a problem may be clearly tracked.

02-1 Have you provided detailed information that supports a clear understanding and data usage?

[Yes I No I N/A]
□ □ □

- Metadata may be defined data that provide an explanation. By recording the characteristics of raw data using the metadata, it is possible to deliver data information when those data are reused in the future or when additional data in the same format are collected.
- Information on data, such as metadata, and a detailed manual, should be secured so that not only developers but also concerned parties using the AI system may comprehend and apply the collected data.
- Sources, formats, methods of data collection/filtering/processing, data license, and protective attributes that may cause bias are some examples of information on collected data that should be delivered to the concerned parties.

02-2 Have you recorded and managed the sources of the data?

[Yes I No I N/A]
□ □ □

- The quality of the training data is one of the important factors that greatly influences the performance of an AI model. It is possible to use various open-source datasets to employ quality training data.
- In the case of open-source datasets, errors may be identified in the process of the data being used by numerous users, which may lead to version changes in the data owing to dataset modification and reconstruction.
- It is vital to carefully manage information on training data such as the source, construction time, and open-source dataset version to respond to AI model problems that may be caused by the dataset itself.

2 Data Collection and Processing

Requirement 03 Removal of abnormal data to ensure data robustness

Safety

▶ Data used to train AI models should not be influenced by outliers, poisoning, and evasion. Data robustness should be ensured by applying verification and protection methods against the aforementioned influences.

03-1 Have you identified data outliers and checked for normality and errors?

[Yes I No I N/A]
□ □ □

- Outliers and data errors in training data refer to all errors that may occur in the process of collecting and processing the dataset for the training data. These errors may present themselves in various forms of noise in the data, bias within the training data, erroneous labeling, and labeling omissions. It would be impossible to fully ensure the performance and robustness of the AI model if the aforementioned errors were not verified and treated.

03-2 Have you made an effort to protect the data against attacks?

[Yes I No I N/A]
□ □ □

- Because the process of developing or operating an AI service is exposed to attacks intended to alter training data or minimally modify input data to generate results that are different from the expected output, it is necessary to review and apply response measures to such attacks.

Type of attack technique	Content of attack technique	Main defense methods
Data poisoning attack	AI services, in general, are retrained with newly collected data after a model has been put in place so that they can adjust to the distribution of input data (e.g., intrusion detection system). During this procedure, an attacker may contaminate the training data by carefully implanting perturbed data that would damage the normal functions of AI services.	• Adversarial training • Gradient Masking (Distillation) • Feature Squeezing
Evasion attack	To prevent training models from correctly identifying input data, the attacker subtly varies the forms of noise in the input data that would perturb them. Although such changes are not apparent to humans, they would have a significant effect on the output of deep–learning models.	

Requirement 04 Removal of bias in collected and processed training data

Respect for diversity | Accountability | Transparency

▶ Measures should be established to recognize and eliminate biases that may occur during the collection and processing of data that would be used for training. Primacy should be given to checking biases that may occur during data collection. In addition, because biases may occur when selecting characteristics for training or when labeling and sampling data, it is necessary to prepare measures to eliminate them.

04-1 Have you taken measures to mitigate biases stemming from human and physical factors during data collection?

[Yes I No I N/A]
□ □ □

- Human factor–based biases result from people's conscious or unconscious biases toward certain information.
  - ✓ Human biases: automation bias, group attribution bias, implicit bias, and in–group bias
- Human biases may be prevented by establishing clear data collection and revision standards for collecting data such that the data characteristics are not biased for each collector. Another method is to obtain a sufficiently diverse pool of data reviewers to correct biases during the revision process.
- Data biases may also occur from physical factors employed in data collection tools or methods. Factors such as image–taking tools or saving equipment may physically limit the collection of data regarding the color, brightness, and resolution of images.
- The aforementioned may not only make it difficult to identify the age or race of subjects in the images, but also cause AI training to be based on data collected through certain methods. Hence, the diversity should be increased by either eliminating physical factors that may cause biases or using a variety of collection tools.

04-2 Did you analyze the features used in training and prepare the selection criteria?

[Yes I No I N/A]
□ □ □

- It is important to sort out discriminatory components included in the data in advance to eliminate biases, which calls for the need to analyze features used in training and establish the selection criteria.
- Discriminatory components refer to matters that may arouse social criticism and biases based on training results owing to information on gender, race, and socially vulnerable classes included in the data. ISO/IEC 24027 defines them as protected attributes and marks them as features that should not be reflected in data training to prevent biases.

Protected attributes that may evoke social criticism

ISO/IEC 24027	IBM Watson OpenScale	Google
age, gender, race, income, family relations, level of education, height, weight, and disability status	age, gender, race, marital status, and address	race, gender, disability status, and religion

04-3 Did you check for and prevent biases that may possibly occur during the data labeling?

[Yes I No I N/A]
□ □ □

- AI models of a supervised learning series require labeling in the training data. During such labeling processes, however, biases may arise from the reflection of certain intentions of the labeler, omission of characteristic information owing to mistakes, and unconscious judgments.
- The aforementioned may be caused by a lack of expertise on the part of the labeler and an absence of consistency in task and judgment standards. Therefore, biases should be prevented by identifying potential sources of biases that may be brought on by the labeler, evaluating labeling outcomes, and providing education on the task standards. Furthermore, various labelers should be recruited to minimize biases created by them, and a sufficient number of reviewers should be secured to carry out bias prevention tasks.

04-4 Did you carry out data sampling through a data distribution verification to prevent biases?

[Yes I No I N/A]
□ □ □

- Sampling is a method used to verify the distribution of the overall data by extracting certain data from a dataset based on consistent criteria. The extracted sample data should represent the distribution of all collected data to achieve significance. If this is not the case, there may be unintentional biases during the sampling process.
- Stratification sampling is one of the main methods of sampling. It classifies a population by considering its characteristics so that there are no overlaps between the samples. It then extracts a sample from each class to build a sample set that reflects the characteristics of the population to prevent biases.

3 AI Model Development

Requirement 05 Ensuring security and compatibility of open-source library

Safety Accountability

▶ In the design and development stage of the AI model, various open sources can be used to shorten the development period and apply the latest technology trends both quickly and flexibly. When using open sources, the list and version of the source should be frequently checked to identify operational and security risks.

05-1 Have you verified the security and compatibility of the open-source library?

[Yes I No I N/A]
□ □ □

- As the version of the open-source library changes, legal and technical issues may arise. Therefore, if an open-source library is being used for model development, any changes following the release of a new version or any issues discovered in the current version should be tracked.
- A major legal issue is the license, whereas a major technical issue is compatibility and vulnerability. Identifying these issues is vital to checking the operational and security risk factors.

Requirement 06 Removal of bias in AI model

Respect for Diversity

▶ Because the type of AI model or the goal of the system may lead to bias, a technique for removing it should be considered during the process of building the model.

06-1 Did you apply methods to eliminate model biases?

[Yes I No I N/A]
□ □ □

- AI models learn potential biases in the data, and even amplify them further. Therefore, not only should the method of removing latent bias from the data be applied to the data purification step, techniques for removing or mitigating model bias should also be applied to the model development procedure.
- Bias mitigation techniques are divided into three variations according to how they are applied. These are bias mitigation techniques applied before model learning (preprocessing), techniques applied during model learning (in-processing), and techniques applied after model learning (post-processing). An appropriate technique should be selected from among them and applied on the basis of the AI model and target mission to be implemented.

Requirement 07 Establishment of response countermeasures against AI model attacks

Safety

▶ Because AI models are vulnerable to the theft of learning data and functions by malicious users or to various abuses, appropriate preventive of mitigating measures must be taken.

07-1 Did you introduce response measures against model extraction attacks?

[Yes I No I N/A]
□ □ □

- Model extraction attacks can be executed in two ways: 1) constructing an alternative model with a performance similar to a learning model in service by analyzing predictions for different inputs of the learned model and extracting classification criteria, and 2) extracting the input data, hyperparameter information, and the hierarchical structure of the model. Several methods can be applied to mitigate these AI attacks, such as limiting the number of queries and obfuscating the prediction results.

Case Study Results of model extraction attacks on cloud-based machine learning

Service	Model Type	Queries	Time(s)
Amazon	Logistic Regression	650	70
BigML	Decision Tree	1,150	631

Stealing Machine Learning Models via Predictions APIs, Usenix, 2016

- The results of a study that created a model similar to the logistic regression provided by Amazon Cloud with 650 queries over a 70-s period using a model extraction attack method.
- Similar AI models can be created by acquiring information on the regression coefficient of the linear classifier and the path of the decision tree.

Requirement 08 Providing AI model specifications and explanations of outputs

Accountability Transparency

▶ It is difficult to know which factors contributed to the predicted results based solely on the output of the AI model. In addition, a number of such AI models may be used to obtain the final result of the system. Through this process, to ensure user trust in the prediction result of the AI model, it is necessary to provide information about the model, as well as an explanation about the process used to derive the result.

08-1 Did you give thought to the application of techniques to explain the expected results of the AI model?

[Yes I No I N/A]
□ □ □

- For users to feel confident about the prediction results of the AI model and its operation, the system should assist users in understanding the judgment or derivation process by which the AI model provides predictions as well as provide explanations and grounds for its conclusions.
- The review and application of XAI (eXplainable AI), which can provide the basis for model judgment in a way that humans can understand, should be considered. XAI technologies such as surrogate models, attention mechanisms, and an internal analysis can be introduced depending on the elements that require explanation and the characteristics of the AI models.

08-2 Did you provide details of the AI model transparently through a fact sheet?

[Yes I No I N/A]
□ □ □

- One of the ways to secure transparency in AI systems is to secure fact sheets, which are bundles of various results generated during the development, testing, and distribution of AI models or services. When a fact sheet is secured, results such as the purpose of the model, input/output information, performance, bias, and reliability can be transparently disclosed upon the user's requests for information related to the AI model.
- IBM's AI Fact Sheet 360 Project proposes ways to ensure transparency in artificial intelligence systems through these fact sheets and aims to explain the main information and components of AI models through fact sheets as needed without disclosing algorithms for the system developed.

Requirement 09 Providing confidence value for AI model output

Transparency

▶ The trustworthiness of an AI model can be used to explain the results more specifically and accurately. Essentially, trustworthiness refers to the reliability of the algorithm of the derived results by showing the performance indicators of AI models, such as accuracy, precision, and recall, together with uncertainty. If trustworthiness is to be used, it is necessary to review in advance whether displaying it will assist users in making decisions. Moreover, measures need to be taken in the event of low trustworthiness.

09-1 Did you provide a confidence value for the AI model output results that require one?

[Yes I No I N/A]
□ □ □

Step 1: Determining whether the results require proof of trustworthiness

- Providing trustworthiness to users, that is, demonstrating uncertainty with performance indicators of the AI model, allows users to receive objective information about the trustworthiness of the results as well as information on how accurate the AI model is. Consequently, it can be useful for assisting the decision-making of the user, but it may also cause problems such as confusion. Therefore, the necessity of trustworthiness must be evaluated in light of the situation and context.

Step 2: Calculating trustworthiness

- Indicators such as the precision, reproducibility, and the mean average precision (mAP) can be used to estimate the accuracy. Methods of estimating uncertainty include ensemble and dropout techniques.

[Yes I No I N/A]

09-2 Did you lay out a solution plan in case the trustworthiness is low? □ □ □

Step 1: Define the meanings according to the confidence interval

- The confidence intervals for trustworthiness should be determined once trustworthiness is determined, and its meaning should then be defined. When trustworthiness is indicated in the lesion diagnosis function, the confidence interval of the diagnosis result (e.g., 50% or less, 51%–60%, 61%–80%, 81%–90% or higher) and the definition of trustworthiness for each section should be determined.

Step 2: Prepare a solution in the event that trustworthiness does not meet the expectations

- The results of the AI model can be derived without trustworthiness meeting expectations owing to restrictions such as the input or response time. There should be several solutions to this situation, for example, preparing and providing follow-up information to users and implementing a function to warn service personnel about the low trustworthiness of the model.
  - ✓ For example, if the AI fails to recognize the user’s input value, that is, his or her voice, and consequently has an inaccurate input value, this will result in low performance and uncertainty. AI speakers deliver users with avoidant responses such as "I don't know what you mean," which can also be seen as a measure against trustworthiness falling short of expectations.

4 System Implementation

Requirement 10 Removal of bias that may occur when implementing AI system

Respect for Diversity

▶ If bias is not considered in the implementation stage of the AI system, the AI system may become biased with the background knowledge or prejudice of the system designer or developer. The design should therefore aim to identify and remove any biases.

10-1 Did you make an effort to eliminate biases from the source code and user interface?

[Yes I No I N/A]
□ □ □

- In addition to bias caused by the data and model, bias may occur through a source code created by a specific person and a user interface that implicitly induces a specific selection.
- To prevent bias in the implementation stage of the AI system, it is necessary to periodically review the written code and determine whether specific class access was omitted during the code implementation process or whether the bias of the developer was reflected in the code.
- In terms of the user interface and interaction, the system should be designed in such a way that prevents the possibility of presentation bias or ranking bias, among other bias types, by checking in advance whether such bias exists.

4 System Implementation

Requirement 11 Implementation of safe mode of AI system

Safety | Accountability | Transparency

▶ AI can produce results or make decisions that can negatively affect individuals or society as a whole. As a result, safe modes should be implemented and reporting should be allowed.

11-1 Did you apply safe mode in the case of an attack, a deterioration in the performance, or social issue?

[Yes I No I N/A]
□ □ □

- Fail–Safe is a general concept used throughout the industry and refers to a mechanism and function that can maintain a safe state even if a problem occurs owing to a failure or error. This can also be applied to AI systems. Whenever external attacks, human errors, deterioration of artificial intelligence models, social criticisms, or accidents from the existence of bias are expected in AI systems, either their causes should be identified and resolved, or the user should be provided with a way to regain normal function. As the coping mechanism operates, it is said to be in safe mode.
- Methods and examples of implementing safe mode are as follows.
  - ✓ When a problem occurs in the system, stop the function and switch to the screen providing feedback
  - ✓ When a problem occurs in the system, the service is restored to the initial screen or state
  - ✓ When the trustworthiness of the results as judged by the AI is low or the possibility of a problem is high, avoid making decisions or giving guidance to the user regarding the situation
  - ✓ Identify the malicious intentions of the user and refuse to input
  - ✓ In the event of a system problem during an automatic and autonomous operation, induce human intervention
  - ✓ Provide guidance and response to expected user errors

11-2 Did you generate a report when a problem occurred with the AI system?

[Yes I No I N/A]
□ □ □

- AI systems can exhibit bias or a performance degradation owing to external attacks and a misuse of users during service. To enable the system operator to comprehend this, the system must either have its own diagnostic function or a feature that allows the user to notify the operator regarding the relevant matter.
- The system's own diagnostic function should be capable of responding to a poor service performance, or inspecting external attacks to the greatest extent possible and reporting such facts to the system operator.
- Using the user report function is a way for the user to communicate certain issues to the system operator when a problem arises, such as a temporary error in the system or bias in the derived result. This function should enable reporting by humans in addition to its own diagnostic function.

Requirement

12 Fostering user's understanding of AI system instructions

Transparency

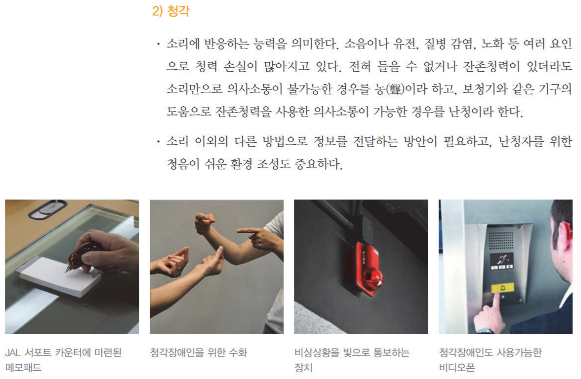
▶ Even if a technique that provides explanations for the prediction results of a model is applied, it is often difficult for the user to immediately understand and interpret such results. Therefore, the operator or service provider of the AI system should evaluate whether the results provided to the user are understandable, interpretable, and explainable.

12-1 Have you analyzed the characteristics and limitations of the users of the AI system?

[Yes I No I N/A]
□ □ □

- To evaluate whether the results of the AI system are appropriate, users who read the results must first be considered. Because the level, depth, and context of the result (description) depend on the user, a detailed analysis of the user should be conducted.

Reference Seoul Universal Design Integration Guidelines



2) 청각

· 소리에 반응하는 능력을 의미한다. 소음이나 유전, 질병 감염, 노화 등 여러 요인으로 청력 손실이 많아지고 있다. 전혀 들을 수 없거나 잔존청력이 있더라도 소리만으로 의사소통이 불가능한 경우를 농(聾)이라 하고, 보청기와 같은 기구의 도움으로 잔존청력을 사용한 의사소통이 가능한 경우를 난청이라 한다.

· 소리 이외의 다른 방법으로 정보를 전달하는 방안이 필요하고, 난청자를 위한 청음이 쉬운 환경 조성도 중요하다.

JAL 서포트 카운터에 마련된 메모패드

청각장애인을 위한 수화

비상상황을 빛으로 통보하는 장치

청각장애인도 사용가능한 비디오폰

In the 'Seoul Universal Design Integration Guidelines', various characteristics (gender, age, nationality, body size, disease, and cognitive ability) of users who can use public facilities were defined and analyzed in advance.

12-2 Did you provide sufficient instructions based on the user characteristics?

[Yes I No I N/A]
□ □ □

- Users who use the service are diverse, and the results of the AI system may be interpreted and misunderstood from different perspectives. Thus, in view of the user characteristics analyzed in 12-1, reference items aiding the evaluation of the explanation are collected. For an explanatory evaluation, the clarity, specificity, and accuracy can be considered.

5 Operation and Monitoring

Requirement 13 Securing traceability of AI system

Accountability Transparency

▶ Technical countermeasures, such as system logs, data monitoring, and the tracking of contributions to decision-making between AI models and humans, should be put in place to track the causes of problems at the operational stages of the AI system.

13-1 Have you established measures to track and respond to the decision-making of the AI system?

[Yes I No I N/A]
□ □ □

- AI systems can be programmed to make decisions unassisted, or through the intervention of an operator or user. Further, if the AI system is designed and developed to enable learning to take place during operation, then continuous monitoring of learned data and models is required.
- In the case of AI systems, unlike traditional software, the lifecycle process is repeated. Thus, even in the service operation stage, it is important to secure a tracking plan that considers the entire lifecycle.
- To track the impact of the AI system output results that may be produced by functional factors, such as the construction of artificial intelligence models, datasets, and systems themselves, as well as human factors including artificial intelligence system operators and users, log collection target information must be defined and monitored at each stage of the system.

13-2 Do you regularly manage the records of changes to the training data?

[Yes I No I N/A]
□ □ □

- In AI models, the learning model varies depending on the data used. As a result, the design of the model or the change in the main parameters may be made together. Consequently, when the training data are changed during the model development process, it is necessary to allow the management of the training data versions and causes of the changes to be tracked.
- The use of open-source tools for managing the training data versions, and the establishment of an autonomous system, may help accomplish this. In addition, information should be provided on the cause of the changes in the training data, the structure of the revised training data, and the expected output result of the training model such that stakeholders who use or operate the training data can assess the impact of such changes.

13-3 Do you periodically update the history of the training data being managed?

[Yes I No I N/A]
□ □ □

- Learning–based AI models that are heavily influenced by data may show a poor performance when dealing with new data, and additional learning that includes new data may be necessary. When evaluating the impact of the AI model because of new data, it is necessary to record and manage how the model performance changes according to the ratio of new data included in the training data.
- Further, when evaluating the performance impact of existing, learned AI systems based on new data, through performance comparison analyses using representative artificial intelligence algorithms in the domain, it is necessary to check whether procedures such as a redesign and retraining of the AI models are needed.

Requirement 14 Providing the service coverage and the target of interaction

Accountability Transparency

▶ To ensure that the service provided by the AI system is used properly, and is not misused or abused, the purposes, nature, limitations, and interacting subject should be described to the user in advance.

14-1 Have you provided explanations encouraging the proper uses of the AI service?

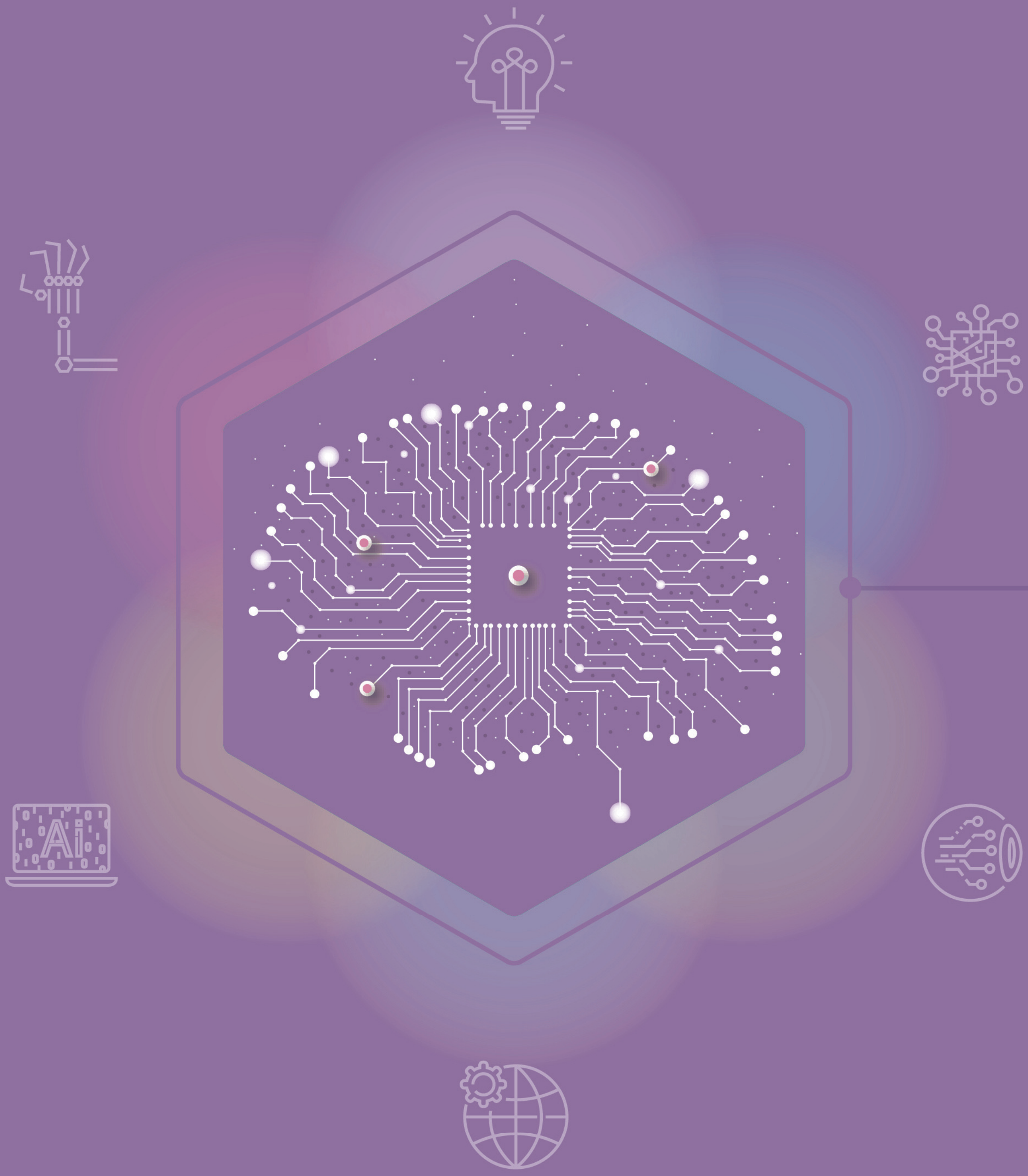
[Yes I No I N/A]
□ □ □

- With the expansion of AI applications, there are cases in which users falsely believe that the service will function in a much broader way than it will in reality. Hence, it is vital to prevent users from misusing or abusing AI technology and to adjust their expectations of services by explaining the purposes, scope, and limitations of the services.
- In particular, if a service personifies a non-human object, there is a high likelihood that the initial expectations of the user will be incorrect. Consequently, the service provider must set the expectations of the user for the service by explaining the purpose and intention of the service, to what extent the service is provided, and any limitations related to the service.

14-2 Did you clearly explain the target of interaction?

[Yes I No I N/A]
□ □ □

- AI systems have recently become increasingly personified to enhance user intimacy and usability. However, as AI technology advances, it is becoming more difficult to distinguish it from human intelligence. This could lead to users being confused about whether they are interacting with a human being or a system. The service provider should therefore inform the user of the subject of the interaction to reduce any confusion the user may experience.


Ai

PART 3

Verification and Validation Items

Table of Contents: Verification and Validation Items

1 Planning and Design

2 Data Collection and Processing

Table of Contents: Verification and Validation Items

Table of Contents: Verification and Validation Items

1 Planning and Design

Requirement 01-1 Have you analyzed the risk factors that may occur during the lifecycle of the AI system?

E-01 Did you derive the risk factors of AI systems and identify their ripple effects?

[Yes I No I N/A]
□ □ □

- The risk factors of AI systems are different from those that can occur in software and hardware-based systems. In contrast to software defects and errors and hardware aging and wear, AI systems should consider bias, a lack of explanations, and attacks perpetrated against the models, which may appear as characteristics of a data-based analysis. The classification and outline of these factors are presented in ISO/IEC 23894.2 and ISO/IEC 24028.
- For each risk factor derived, the causes and possible consequences should be analyzed. Possible consequences refer to phenomena and accidents that can have a negative social impact, including accidents that harm the human body, and discriminatory phenomena that give rise to social problems.
- The consequences of the risk factors can be measured based on the severity and frequency of their occurrence. This is a ripple effect of such factors. Countermeasures should be put in place based on the evaluation of the risk factors, beginning with those that have high ripple effects.
- In calculating and evaluating the ripple effect, a scale suitable for the situation can also be introduced and combined with the severity and frequency of occurrence.

Reference Risk factors of AI systems

- Regarding the risks to be recognized by AI systems, Appendix A of ISO/IEC 23894.2 presents the following keywords. The risk factors should be specified and subdivided based on these keywords.
  - Fairness, Safety, Robustness
  - Security, Privacy
  - Transparency, Explainability
  - Environmental Impact
  - Accountability, Maintainability, Availability, Data Quality
  - Expertise
- In addition, ISO/IEC 24028:2020 summarizes the vulnerabilities and risks that can be considered when implementing an AI system (Chapter 8: Vulnerabilities, Threats, and Challenges), which can be referenced.
  - AI specific security threats
  - AI specific privacy threats
  - Bias
  - Unpredictability
  - Opaqueness
  - Challenges related to the specification of AI systems
  - Challenges related to the implementation of AI systems
  - Challenges related to the use of AI systems
  - System and Hardware faults

Requirement 01-2 Did you take measures to eliminate, prevent, or minimize the effects of the risk factors?

E-02 Did you derive a risk factor removal plan and check the resulting effect?

[Yes I No I N/A]
□ □ □

- Technical methodologies, such as implementation and operation methods that may generate risk factors, software and hardware functions, and model learning techniques and strategies, should be derived. The classification and outline of these methodologies are presented in ISO/IEC 24028.
- Previously, the ripple effect of the risk factors was evaluated during their analysis. Countermeasures should be targeted primarily at those risk factors that have the greatest ripple effect.
- When countermeasures are applied, it is important to reevaluate the ripple effect to ensure that the risk factors are eliminated or prevented, or that their impact is reduced.

Reference Risk factors of artificial intelligence systems and countermeasures.

- ISO/IEC 24028:2020 summarizes the vulnerabilities and risks that can be considered when implementing an AI system (Chapter 8: Vulnerabilities, Threats, and Challenges). For additional reference, relevant countermeasures are presented in a comprehensive manner (Chapter 9: Mitigation Measures).

Chapter 8 Vulnerabilities, threats and challenges	Chapter 9 Mitigation measures
8.1 General	9.1 General
8.2 AI specific security threats	9.2 Transparency
8.3 AI specific privacy threats	9.3 Explainability
8.4 Bias	9.4 Controllability
8.5 Unpredictability	9.5 Strategies for reducing bias
8.6 Opaqueness	9.6 Privacy
8.7 Challenges related to the specification of AI systems	9.7 Reliability, resilience and robustness
8.8 Challenges related to the implementation of AI systems	9.8 Mitigating system hardware faults
8.9 Challenges related to the use of AI systems	9.9 Functional safety
8.10 System and Hardware faults	9.10 Testing and evaluation
	9.11 Use and applicability

2 Data Collection and Processing

Requirement 02-1 Have you provided detailed information that supports a clear understanding and data usage?

E-03 Did you describe the data characteristics both before and after data cleaning?

[Yes | No | N/A]
□ □ □

- Data cleansing is a step of data screening and processing for the building of training data prior to data labeling, and users who use only data that have been cleaned cannot accurately identify the features of the raw data. Therefore, considering the possibility of collecting additional data in the future, the relevant information for data cleaning and the data features before and after cleaning should be described.
- Data cleansing can basically be applied by removing or correcting some of the data according to a set rule using an open-source tool or through a visual inspection, and the cleaned data can be visualized for an analysis of the data features.
- If raw data are directly collected, it is necessary to present the criteria for data cleansing, such as the purpose of the data building, data type, and domain characteristics and information on the cleansing tool. The following is an example of the criteria used for data cleansing with respect to different data types:
  - ✓ Image data: image size, aspect ratio, image quality, equipment used for image capturing, personal information policy, copyright, etc.
  - ✓ Text data: text volume, syntactic accuracy, relevance of the content of the text, relevance of the topic, etc.
  - ✓ Voice data: volume, pronunciation accuracy, noise, inaudibility (according to the acceptable range), personal information, copyright, etc.

E-04 Did you distinguish between the training data and the metadata? Did you secure the specification data for each type of data?

[Yes | No | N/A]
□ □ □

- For the use of a training dataset for AI model training, information regarding the dataset is required, which is referred to as metadata. Metadata can be provided in JSON or XML formats, and the following information can be included depending on the dataset type.
  - ✓ Image metadata: date of image capturing, location of image capturing, exposure, etc.
  - ✓ Text metadata: title, text length, date of creation, etc.
  - ✓ Voice metadata: date and time of recording, length, person responsible for recording, speaker information, number of speakers, etc.
- As described above, metadata and training data should be differentiated, and specification data for the respective data should be created to facilitate the use of data for AI model training from the developer's point of view.

E-05 Did you explain the reason for selecting and reflecting the protective attributes?

[Yes I No I N/A]
□ □ □

- During the process of training an AI model using a large dataset, the model may learn various types of biases such as within the dataset itself or latent bias. This may lead to a degradation of the performance of the AI model and the service of the AI system owing to ethical issues such as sexism and racism.
- Data bias can be mitigated by analyzing variables in the data to identify specific variables that have a significant impact on causing biased results, designating these variables as protective attributes, and ensuring that these attributes are not reflected in the process of the AI model training.
  - ✓ Representative open-source analysis tools for detecting and examining data bias include the Google What-if Tool and IBM Fairness 360.
- Therefore, in consideration of future users of the collected/constructed data, explanations should be provided regarding the purpose of the AI system, the reason for selecting the protective attributes of the dataset, the process, and the status of reflecting the selected attributes.

E-06 Did you conduct training and create work guidelines for the labelers?

[Yes I No I N/A]
□ □ □

- Data labeling involves the annotation (correct answer) of raw data for training an AI model and is applied by multiple labelers. Labeling can directly affect the model performance and ensure the quality of the dataset, and thus it is important to conduct training for the labelers and prepare detailed work guidelines.
- For labeling tasks, the target, scope, detailed procedure, and labeling tool may vary depending on the type of data. The general procedure of the data labeling is as follows, according to which, training and guideline documents for the labelers should be secured.
  - ✓ Data acquisition and cleansing: Conduct raw data acquisition and data cleansing
  - ✓ Organizing the target and scope of the data labeling: Define the target and scope to determine items to be labeled in the raw data. In particular, detailed criteria need to be established according to the type of data (labeling some of the data, de-identification of personal information, class definition and management, etc.)
  - ✓ Establishment of labeling methods and procedures: Decide on methods of labeling such as automatic, semi-automatic, or manual methods according to the information to be labeled; Prepare detailed work standards such as allocation of the labeling work and criteria for labeling by data type
  - ✓ Conduct data labeling: Upon completion of labeler training based on detailed work guidelines, data labeling is applied (depending on the previously determined labeling work approach, in the case of an automatic or semi-automatic method, an appropriate labeling tool selection and training should be conducted)

Requirement 02-2 Have you recorded and managed the sources of the data?

E-07 If using an open-source dataset, did you specify the source?

[Yes I No I N/A]
□ □ □

- When an open-source dataset is used to train an AI model, errors or biased output that were not detected at the time of training may occur. In addition, the biased output may be associated with ethical issues following changes in social perception, and thus there is a possibility of data bias that was not recognized at the time of building the open-source dataset.
- Therefore, when building a learning-based AI model using an open-source dataset, the clear source of the acquired data and the related information must be recorded and managed to identify the cause of the data bias that may occur in the past, present, and future.

Requirement 03-1 Have you identified data outliers and checked for normality and errors?

E-08 Did you identify possible errors by visualizing the overall distribution of the learning data?

[Yes | No | N/A]
□ □ □

- After the data cleansing step, which is one of the steps of data preprocessing, additional input errors can be identified by visualizing the overall data distribution. In particular, such visualization of the data distribution is highly useful in the analysis and understanding of data used for training an AI model.
- There are various techniques for visualizing a data distribution depending on the characteristics of the data. First, there is a distribution plot that visualizes the data distribution by utilizing the mean, variance, and deviation of all data, as well as a categorical plot for the presentation of categorical data, and a matrix plot for visualization of 2D matrix data.

Types of visualization methods	Description
Histogram	Visualize the data in the form of a histogram for variables
Kernel Density Estimation plot	Visualize the data in the form of a density estimation plot for one or two variables
Empirical Cumulative Distribution Function plot	Visualize the cumulative distribution of all data
Rug plot	This is a plot for visualization of the data distribution, displayed as marks along the x- and y-axes, and is mainly used to supplement other diagrams

E-09 Did you apply the outlier identification technique for the training data?

[Yes | No | N/A]
□ □ □

- One of the key activities in data preprocessing is to detect and remove data outliers. Unlike in the case of data omission, for data outliers, the value of the data is already determined; however, it is a value outside of the normal range when considering the values of the entire dataset, and thus it is difficult to detect outliers through a simple search process.
- The methods of outlier detection are mainly based on finding data points that differ from others when considering the entire dataset using statistical techniques. Representative techniques of outlier detection include the Z-score and interquartile range.

Outlier detection methods	Description
Z-score	As the simplest statistical method, the Z-score quantifies how far an observed data point is from the total data distribution using the distribution mean and standard deviation of a given dataset.
Interquartile range	Divide all data into two parts based on the median (Q2), and then divide the data into the left median (Q1) and right median (Q3) to produce a total of four ranges. Find the quartile range (Q3-Q1), and if the data are outside the range, they are determined as outliers.

Requirement 03-2 Have you made an effort to protect the data against attacks?

E-10 Did you provide defense measures against attacks such as data poisoning and evasion?

[Yes I No I N/A]
□ □ □

- Various defense techniques are available to defend against adversarial attacks and to enhance the robustness of AI services. In particular, representative techniques for preventing evasion and poisoning attacks during the data design and model training stages include adversarial training, gradient masking, and feature squeezing.

Defense techniques	Description
Adversarial Training	This is the most intuitive algorithm used for understanding. When training an AI model, the number of cases that can be used as adversarial are considered in advance and are included in the training dataset. However, the performance of adversarial training can only be ensured as long as there is a process in place for generating adversarial data with a sufficient number and diversity.
Gradient Masking /Distillation	Considering the fact that most adversarial attacks are made by observing the gradient through the model inference process, the method prevents the exposure of the gradient of the learning model as an output (gradient masking) or in the architecture of the model, and to avoid providing clues regarding the direction of an adversarial attack, the gradient is not emphasized as a type of regularization method.
Feature Squeezing	As another method to prevent adversarial attacks, apart from the existing AI model, another AI model can be added to determine if the given input is an adversarial case. In addition, if a system is composed of an ensemble of multiple models, it is possible to avoid white-box attacks on a specific model. Considering that it is difficult to develop adversarial attacks that can be universally applied to multiple AI models, studies have reported that this technique is useful as a defense measure.

Requirement 04-1 Have you taken measures to mitigate biases stemming from human and physical factors during data collection?

E-11 Did you apply procedural and technical means to eliminate human bias?

[Yes I No I N/A]
□ □ □

- Human bias in the process of collecting data originates from the bias of the person in charge of the data collection. In this case, it is necessary to prepare guidelines for a data collection operation to reduce the individual deviation of the data collection operators, recruit operators from various backgrounds and thus rule out a bias formation based on specific backgrounds and tendencies, and ensure a pool of sufficient data-labeling reviewers for the results of the data collection.

E-12 Did you use a heterogeneous collection device to secure data diversity?

[Yes I No I N/A]
□ □ □

- When data are acquired using specific hardware and equipment, it may be difficult to obtain a large number of consistent data owing to the environment of the data acquisition and constraints. In this case, because such limitations adversely affect the ability to ensure diversity in the acquired data, it is necessary to secure the quantity and diversity of the data by using a large number of equipment and heterogeneous devices.
- However, in this case, because the routes of data acquisition and environment (e.g., camera filming, web crawling) are different among the different devices, data cleansing and inspection must be properly applied to ensure consistency of the data for their utilization after data acquisition.

E-13 Did you check the data for bias that might have been caused by the hardware?

[Yes I No I N/A]
□ □ □

- For data acquisition and generation, hardware or equipment such as cameras and microphones are used. In this case, owing to physical factors such as system specifications and the data acquisition environment, bias may occur, such as data collection for limited situations and scenarios.
- It is therefore necessary to establish a plan to check and handle such factors during data acquisition. The table below presents an example of collecting video data using video filming equipment.

Steps of bias prevention measures	Description
Writing a scenario for filming	• Writing a scenario for filming according to the method and purpose of the video • Filming plans such as a written schedule for each scenario, hiring a professional filming company, etc.
Review of the filming scenario	• Inspection of the filming scenario and details according to the scenario and with the person in charge of filming
Establish a data collection environment suitable for labeling purposes	• Selection of filming location (e.g., confirmation of data acquisition environment and no-entry zones) • Checking preparation process for filming (e.g., camera, lighting, and subject) and setting the filming environment according to the scenario (e.g., angle of view, composition, image quality, and items required for filming)
Acquisition of data and determining additional information to be acquired	• Proceed with filming according to the pre-determined filming technique (e.g., filming lens, focal length, frame, balance, and resolution) and filming method (e.g., filming angle, distance, ratio, quality, and quantity) • Identification of additional information to be acquired for each purpose (e.g., data size, length, and output format)
Review of appropriateness of the raw data acquired	• Check and confirm the adequacy of the raw data acquired for a pre-planned purpose (filtering of data that do not meet the purpose of data acquisition or low-quality data) • Checking for possible legal issues with the acquired data

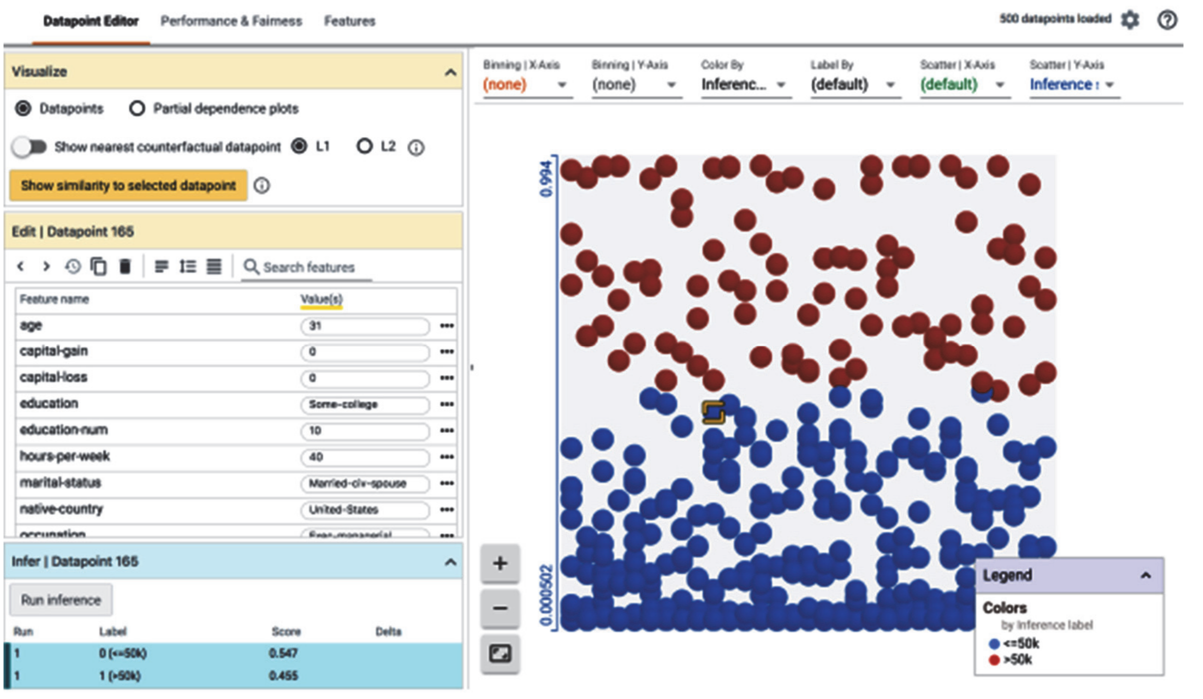
Requirement 04-2 Did you analyze the features used in training and prepare the selection criteria?

E-14 When selecting a protective attribute, did you conduct a sufficient analysis?

[Yes I No I N/A]
□ □ □

- If an analysis is not properly applied when selecting protective attributes, the performance of the model may be degraded. Therefore, if there is a protective attribute that affects the model output, it will be necessary to observe and analyze how the model output changes while changing a portion of the data from the given dataset.
- The Google What-if Tool visualizes and shows the trend of change in the inference results according to changes in the data for a machine-learning-based regression and classification model, how much the set protective attribute affects the unfairness of the results, and how the performance results change.

Reference Example of Google What-if Tool Screen Capture



● Datapoints ○ Partial dependence plots

Show nearest counterfactual datapoint ● L1 ○ L2

Binning | X-Axis (none)
Binning | Y-Axis (none)
Color By Inferenc...
Label By (default)
Scatter | X-Axis (default)
Scatter | Y-Axis Inference :

Edit | Datapoint 165

Feature name	Value(s)
age	31
capital-gain	0
capital-loss	0
education	Some-college
education-num	10
hours-per-week	40
marital-status	Married-civ-spouse
native-country	United-States
occupation	Exec-managerial

Infer | Datapoint 165

Run	Label	Score	Delta
1	0 (<=50k)	0.547	
1	1 (>50k)	0.455	

E-15 Did you exclude features that can cause bias?

[Yes I No I N/A]
□ □ □

- By using the selected data features when training an AI model, efficient learning and a reduction in computer resources and costs can be achieved, and during the analysis of the relationships between multiple features, latent bias may be detected through a more in-depth understanding of the data.
- During the data acquisition and processing, minimizing bias can be considered through the utilization of feature selection techniques. Examples of feature selection techniques include the filter method, wrapper method, and embedded method. These methods are based on the use of features that have high correlation coefficients through analyzing the statistical correlation of features in the data, or the use of subsets that show a good performance for some of the features.

2 Data Collection and Processing

E-16 In preprocessing the data, did you unnecessarily remove any features?

[Yes I No I N/A]
□ □ □

- The feature selection technique allows mitigation of latent bias and an improvement of the model performance; however, if the technique is applied in excess, it may cause overfitting problems or even bias.
- In particular, when feature selection is applied for all data, the same features are used in a cross-validation, which may cause bias. It is therefore necessary to check to prevent an excessive feature selection and the exclusion of features.

Checklist items	Actions
Is domain knowledge available?	If so, it is best to build tentative features based on the available domain knowledge.
Are the features related to each other?	If not, it is recommended to apply normalization to fit the scale.
Are the features inter-dependent?	If so, it is recommended to extend the feature sets by combining related features.
Do input variables need to be removed for dealing with cost, speed, etc.?	If not, it is better to separate or construct a weighted sum function of the features.
Should features be evaluated individually for the filtering or understanding of the features for the model?	If so, it is recommended to use the variable ranking method.
Is a predictor required?	If not, there is no need to apply the feature selection.
Are the data noisy?	If so, it is recommended to remove outliers using the top-ranking variable.
Is it clear what to do first?	If not, use a linear predictor, as well as a forward selection or 0-norm embedded technique.
Are new ideas, time, computing resources, and sufficient data available?	If so, it is recommended to try different methods.
Is a reliable solution required?	If so, it is recommended to run the approach multiple times and use a bootstrap method.

Requirement 04-3 Did you check for and prevent biases that may possibly occur during the data labeling?

E-17 Did you clearly define and communicate work standards to the data labelers?

[Yes I No I N/A]
□ □ □

- Data-labeling techniques include automatic, semi-automatic, and manual methods depending on whether a labeling tool is used. In this case, the labeler is involved in the labeling process, and thus the labeler's latent bias may be reflected in the labeling.
- These latent biases tend to depend on individual judgment because no clear guidelines are established for many different labeling tasks. Therefore, detailed labeling guidelines should be prepared to detect and prevent biases. In addition, sufficient training should be provided to labelers based on the guidelines to minimize the potential for biases among the labelers.

E-18 Did you recruit different data labelers?

[Yes I No I N/A]
□ □ □

- To reduce human bias in the data-labeling process, it is necessary to secure a large number of data labelers and ensure that the demographic characteristics and background knowledge of the labelers have sufficient diversity without any biased tendencies.
- To verify the diversity of the data labelers, two main points need to be checked. First, it is necessary to check whether a method such as crowdsourcing has been introduced. Second, by examining and analyzing the demographic characteristics and background knowledge of the data labelers, it needs to be confirmed whether the labelers actually show a uniform distribution with diverse backgrounds.
  - ✓ Crowdsourcing: This refers to placing an external contract such that the general public having undergone labeling-related training can participate in the data-labeling process. In this way, data labelers with more diverse backgrounds can be recruited than the existing group of data labelers.

E-19 Did you secure various data-labeling inspectors?

[Yes I No I N/A]
□ □ □

- Human bias may still arise despite having recruited data labelers from various backgrounds. It is therefore necessary to have a pool of data-labeling reviewers who check whether the labeling results are consistent with the purpose of the data acquisition and specifications and ask for corrections if required.
- Data-labeling reviewers, as in the case of data labelers, should also show an unbiased distribution with various backgrounds. It is therefore necessary to check whether methods such as crowdsourcing have been adopted, and whether the distribution is diverse and uniform through a survey and analysis of the data-labeling reviewers.

Requirement 04-4 Did you carry out data sampling through a data distribution verification to prevent biases?

E-20 Did you apply sampling techniques for bias prevention?

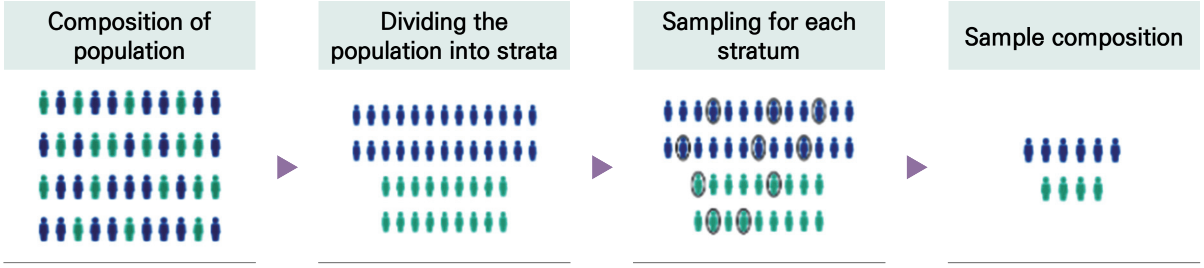
[Yes I No I N/A]
□ □ □

- In the case of sampling demographic data that may cause social prejudice and discrimination, a sampling technique that can prevent bias owing to such source data is applied, and it is necessary to check whether the necessary activities and information are generated during the process of applying the technique.

Reference Example of sampling techniques – Stratified sampling

- In the stratified sampling method, to prevent bias against demographic factors, the population is divided by stratum or class constituting the population before data sampling, and the sampling is then conducted through a divided stratum.

Process of stratified sampling



Composition of population ▶ Dividing the population into strata ▶ Sampling for each stratum ▶ Sample composition

- While applying the stratified sampling method, it is necessary to check ① how the population composition is divided into strata, ② the classification criteria, ③ whether the sampling ratio is the same for each stratum, and ④ the result of the total sample composition.
- The classification criteria based on the stratum and detailed figures such as sampling the ratio may be set differently depending on the service/technology to be implemented using AI, and the information included in the dataset to be handled. The person in charge of sampling shall provide an appropriate basis for the criteria applied.

3 AI Model Development

Requirement 05-1 Have you verified the security and compatibility of the open-source library?

E-21 Did you confirm the license, security vulnerability, and compatibility of the open-source library being used?

[Yes I No I N/A]
□ □ □

- Legal and technical issues may arise from the use of an open-source library depending on changes to the version, and thus it is crucial to manage the version in use and consider the following issues.
- Legal aspect: Check license
  - ✓ An open source can be used free of charge. However, because licenses exist independently, it is the user's responsibility to examine the license notice and license type of the open source before using it to make sure they are aware of their obligations.
- Technical aspect: Check compatibility and security vulnerability
  - ✓ Changing the library version may cause compatibility problems with the development environment, language, tools, and other library versions. Therefore, it is necessary to select the type and version of the open-source library in consideration of compatibility, such as identifying the dependency among libraries.
  - ✓ Security vulnerabilities may be found in the open-source library in use. To minimize the impact of security vulnerabilities, it is necessary to continuously check security vulnerabilities and release notes from version changes to detect and respond quickly.

Requirement 06-1 Did you apply methods to eliminate model biases?

E-22 Did you select the bias removal technique according to the model to be developed?

[Yes I No I N/A]
□ □ □

- Techniques for mitigating bias by AI models are divided into three types. Such techniques can be applied before model learning ("preprocessing"), during the learning process ("in-processing"), and after the learning process ("post-processing").
- Appropriate techniques should be selected and applied according to the characteristics of each method, the AI model to be implemented, and the target mission.

Technique and Indicators	Type of technique			Description
	Pre	In	Post	
Weight redesignation	✓	✓		A method of assigning weights to a sample of learning datasets
Labeling redesignation	✓			A method of modifying the label of the data sample for learning
Variable blinding	✓			A method of preventing the classifier from responding to sensitive variables
Variations	✓		✓	A method of converting data and model prediction distribution during numerical data-based learning
Sampling	✓			A method of removing bias through sampling in learning data
Regularization		✓		A method of correcting the class distribution, which has a significant influence on bias during classification
Optimizing restrictions		✓	✓	A method of assigning a correction value to the loss function of the classifier
Threshold			✓	A method of removing bias when the resultant inference is close to the decision boundary
Correction			✓	A method of setting the positive prediction ratio for equal distribution with the positive data instance ratio

E-23 Did you select and manage quantitative indicators for bias evaluation and monitoring?

[Yes I No I N/A]
□ □ □

- Indicators that quantitatively measure bias can be divided into five categories as shown in the table below. They should be selected based on the model and the mission goals to be developed, and whether the mitigation of bias should be continuously measured and managed.

Classification	Indicator
Parity-based indicator	Statistical/demographic equity indicators, disparity effect indicators
Confusion matrix-based indicator	Equalized opportunity, equalized odds, overall accuracy equity, conditional use accuracy equity, response equity, non-compensation equalization
Score-based indicator	Positive and negative class balance indicators
Counterfactual-based indicator	Post-assumption fairness
Individual fairness indicators	Generalized entropy index, shale index

Requirement 07-1 Did you introduce response measures against model extraction attacks?

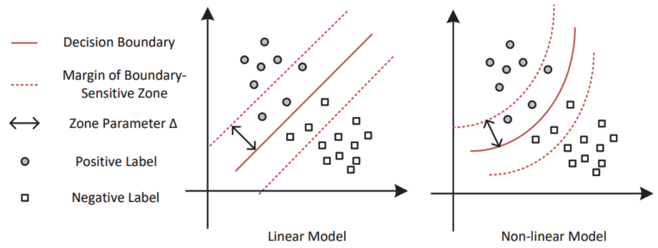
E-24 Did you apply a defense technique to prepare for model extraction attacks?

[Yes I No I N/A]
□ □ □

- The major methods for mitigating attacks on the AI model involve performing several response processes, such as limiting the number of queries for AI services per specific time interval, detection and warning of suspicious queries, and obfuscating prediction results.

Classification of defense technique	Defense technique details
Limiting the number of queries	A technique for defending repetitive queries against model attacks by limiting the number of queries that can be entered within a specific period of time.
Learning-based monitoring	A technique that actively defends against model attacks by using machine learning, such as pre-detection and warning notifications, and executing equivalent defense techniques.
Obfuscating the prediction results	A technique that arbitrarily lowers the accuracy of the prediction result when the prediction result is close to the decision boundary and hinders the extraction of detailed attributes of the model.

Reference A Boundary Differentially Private Layer



The BDPL technique designates the criteria for determining the classification and the area around it as the Boundary Sensitive Zone. With this area being protected, this technique makes it more difficult to extract the model by mixing noise with the result when the input from the outside is closer to the sensitive area.

Requirement 08-1 Did you give thought to the application of techniques to explain the expected results of the AI model?

E-25 Did you provide explanations for each stage of the model output?

[Yes I No I N/A]
□ □ □

- AI services can use multiple AI models internally to simultaneously perform complex tasks, or users and systems can interact through multiple stages to obtain final output results.
- AI systems use pre-learned models, and input/output structures have a great influence on performance depending on the model design. Hence, appropriate explanations must be provided to users for each step of AI services to facilitate their use of the services.

E-26 Did you provide the basis for the output result in such a way that the user can accept it?

[Yes I No I N/A]
□ □ □

- In the case of AI systems using deep-learning technology, they exhibit excellent performance while having low explainability. Low explainability can undermine confidence in model predictions and the trustworthiness of the entire system, and thus it is important to establish the basis for user-acceptable output results.
- As a way to secure the basis for artificial intelligence model output results, the introduction of XAI (“eXplainable AI”) technology through model input factors, model interior, explanatory variables, and proxy model analysis can be considered. Thus, the basis for model inference and output results can be visualized and presented to the user.
- XAI technology has been continuously researched to secure the explainability of AI models, and it is important to select techniques suitable for explanation through analyzing the advantages and disadvantages of AI models and XAI techniques before introducing the technology. An overview of the representative methodology of XAI technology is presented below, as well as its advantages and disadvantages.

Classification	Methodology	Advantage	Disadvantage
Model-agonistic explanation (Based on inductive reasoning about the black box)	LIME	Regardless of the knowledge-based model for parameters and architecture, it can be checked only with input and output (advantage shared with the proxy model -based methodology); it is suitable for practical application, as an explanation of a specific sample is easy.	The unit of explanation changes from time to time; the essence of the model cannot be explained because it only indirectly explains the input and output while not explaining the AI model.
Model-specific explanation (Based on knowledge of parameters and architecture)	LRP	It is intuitive and allows for checking of the contribution inside the hidden layer to allow for identification of what the hidden layer may have detected.	The abstract concept learned by the neural network model cannot be known merely by expressing the contribution in the heatmap.
	Explorative sampling considering the generative boundaries of DGNN	The characteristics of the grid used in the complex generation model can be estimated through samples between each grid.	In the process of viewing and judging multiple samples, it is difficult to express ambiguous changes and the boundaries that divide them using language; or, even if expressed, the bias of the analyst may be involved given the nature of the example-based explanation.
	Rule extraction	The Decision Tree sequence, which helps understand the neural network, is transformed into a format.	In the process of abbreviating the neural network model, information on the model is inevitably omitted.

E-27 If it is difficult to apply XAI technology (eXplainable AI), did you apply an alternative?

[Yes I No I N/A]
□ □ □

- If it is difficult to apply XAI technology according to the service of the AI system, the developer should consider the second-best way to increase the explainability and reliability of the AI system. In this case, the trustworthiness of the system may be secured by verifying the validity of the actual system and explaining the analysis result of the verification.
- This can be achieved by applying the various test techniques employed in existing software development to AI systems. The validity of a system can be verified by using the following validation methods in combination.
  - ✓ Perform repeatability tests in the production environment
  - ✓ Conduct the counterfactual fairness test
  - ✓ Identify exceptions and perform tests of handling exceptions
  - ✓ Conduct the appropriateness test of the AI model according to the data
  - ✓ Perform repeatability tests under single or diverse complex conditions
  - ✓ Conduct an AI model error rate test on different subgroups of the target population

Requirement 08-2 Did you provide details of the AI model transparently through a fact sheet?

E-28 Did you prepare detailed information on the system development process and how the model works?

[Yes I No I N/A]
□ □ □

- Increasing the transparency of an AI system and providing system users with information that allows identification of AI-based program components are key factors in improving system reliability. To this end, if a fact sheet is prepared when developing an AI model, information that allows identification of the AI system components can be provided to the user.
- When preparing the fact sheet, stakeholders involved in the AI lifecycle need to be considered and relevant information should be included so that each stakeholder can identify and select necessary information. The following is an example of the information that must be present in the fact sheet with respect to different stakeholders.

Stakeholders	Fact sheet information
Business Decision-makers	General purpose and direction of the AI system, name of the services in the system, and intended purpose of each service, etc.
Data Scientists and System Developers	Dataset specification used for training and preprocessing techniques, architecture of the training model, input/output specifications, model training parameters, etc.
Model Tester	Information about the test dataset composition and evaluation results, such as key test performance, bias, confidence value, etc.
Model Operator	Performance indicators with respect to model operation and monitoring results, environmental factors affecting performance degradation, environmental conditions for achieving optimal results, etc.

Requirement 09-1 Did you provide a confidence value for the AI model output results that require one?

[Yes I No I N/A]

E-29 Did you review the necessity of providing a confidence value? □ □ □

- Expressing the confidence value of the results derived by the AI system may benefit people’s decision-making processes when using AI. However, it can sometimes serve as an obstacle to decision-making. Therefore, the confidence values must be checked and whether they must be provided should be confirmed rather than providing confidence values for all results immediately.
- Examples of two cases where it is better not to provide confidence values are given as follows:
  - ✓ First, when it is assessed that the provision of the confidence value will not have a significant impact on the user's decision-making. If the effect of providing the confidence value is not clear, one may think that providing it with more subdivided categories will help users in decision-making. However, this could cause confusion, contrary to the expectation. For example, if there are two results derived by the AI system, with confidence values of 85.8% and 87.0%, respectively, the user may be confused about which result to use.
  - ✓ Second, it is better not to provide a confidence value when it is too high or too low. If the user is informed that the confidence value is 100% for the output of the system, the user may blindly accept the system output.

E-30 Did you calculate the confidence value and define the confidence level of the model based on the calculation results?

[Yes I No I N/A]
□ □ □

- To define the confidence value, uncertainty should be calculated along with certain indicators, such as precision, recall, and mAP (mean average precision). Uncertainty is the size of variance of a random variable and is an indicator of how confident the AI model is in the derived results. Techniques of uncertainty estimation include Bayesian neural networks, ensembles, and dropouts.
  - ✓ Dropout is a technique that randomly selects and drops nodes in a neural network and connections between each node.
  - ✓ When leveraging the feature that different neural networks are generated for each time of training both the dropout-applied neural network and Bayesian neural network, the same input value is provided to multiple neural networks that are generated as a result of the training. For the multiple output values that are derived, the average and variance can be calculated. Subsequently, uncertainty is the calculated value of variance.
- By combining the calculated results of the output performance (e.g., precision, recall) and uncertainty, the confidence level can be defined. For example, when there is a model that predicts true or false values, the basis for the prediction “Because the model's prediction probability is high (98%) and the uncertainty about the prediction is low (1%), we can be confident in the "true" result derived through the AI model.” can be presented to the user.

E-31 Did you derive the threshold of the model performance, as well as the confidence value provided if it is less than or equal to the threshold?

[Yes I No I N/A]
□ □ □

- A threshold of the output values can be categorized by thresholds for performance indicators (e.g., precision, recall) of an AI model and a threshold for uncertainty in its performance. To derive the threshold of the output values, first, it is necessary to define the problematic situations that may arise when using the AI model, and identify important variables that govern the problematic instances. In this case, a problematic situation includes both a situation that poses a threat to the life or property of the user, as well as a situation where the model performance is lower than the expected level or the level of quality that needs to be maintained.
- The output threshold value can be derived via various AI techniques, from the conventional techniques of linear discriminant analysis, support vector machine, convolutional neural network, and long-short term memory, to the relatively new techniques such as graph extrapolation network, simple framework for contrastive learning of visual representations.

Requirement 09-2 Did you lay out a solution plan in case the trustworthiness is low?

E-32 Did you provide further explanation to the user when the confidence level of the model output is less than or equal to the threshold?

[Yes I No I N/A]
□ □ □

- If the results of the AI system do not exhibit the expected confidence level owing to certain constraints, the user should be provided with additional information on follow-up procedures. Here, follow-up procedure information specifically guides the user's next actions when the AI system confidence value is low.
- There are various methods for providing user follow-up procedure information, such as the N-best alternatives, numerical indicators, graph-based visualization method, and categorical visualization method. Because each method has different advantages and drawbacks, considering the application field of the AI system and characteristics of main users, the optimal method should be selected and information should be presented to facilitate an easy and intuitive user understanding.

E-33 Did you develop a function to warn the stakeholders when the model performance is below the allowable threshold?

[Yes I No I N/A]
□ □ □

- The performance of the AI model considers the time needed for object detection or recognition and how effectively the explanation is conveyed to the user, as well as the accuracy, precision, recall, and F1-score metrics. Performance degradation refers to a situation where the model performance is below the acceptable threshold. To prevent such degradation in performance, continuous monitoring and related action plans are required.
  - ✓ An acceptable threshold for model performance is defined as follows: For example, when an AI model requires a performance level with 95% accuracy and 1% uncertainty, the threshold values for accuracy and uncertainty are 95% and 1%, respectively. If the model performance is above or below these specific values, problem situations may be induced.

4 System Implementation

Requirement 10-1 Did you make an effort to eliminate biases from the source code and user interface?

E-34 Did you confirm a possibility of bias caused by the source code: for getting access to data; and so on?

[Yes | No | N/A]
□ □ □

- In an AI system, when implementing the data access method for the data to be used in the model source code, various bias types may occur, such as omitting access to a specific class.
- Particularly, in a rule-based system, when hard-coded rules are used based on the expert knowledge in a specific field, the output may be biased to a specific class, potentially causing cognitive bias. Therefore, to mitigate the occurrence of system bias, it is helpful to select a broad range of experts covering various backgrounds.
- To check for bias that may occur during the design and development phases of an AI system, open-source tools (e.g., FairML, Google What-If Tool) can be utilized. These tools perform periodic statistical analysis of output data to detect unknown bias or notify the presence of a risk regarding a function according to pre-set metrics of fairness evaluation. By applying these tools, bias can be quickly detected and alleviated in the implementation process.

E-35 Did you confirm the bias caused by the user interface and the interaction method?

[Yes | No | N/A]
□ □ □

- In AI systems, user interaction bias may occur either through implicitly being induced by the user interface or intentional misuse by the user.
- To prevent the user interaction bias, factors that may cause such bias (e.g., presentation bias, ranking bias) should be detected and removed in advance during the design and implementation phases of the user interface.
  - ✓ Presentation bias: This bias type is caused by depending on the presented information method. For example, a user can only click on the content that is visible when using a product, hence, only the displayed content will be clicked, while others will not. Owing to such characteristics of the user interface, only clicks on specific content may be induced.
  - ✓ Ranking bias: This bias occurs according to the ranking of the presented information. The dominant trend is that the users tend to assume that the result appearing at the top is the most relevant and important one. Therefore, results presented at the top may be more frequently selected by users than those presented at the bottom.

Requirement 11-1 Did you apply safe mode in the case of an attack, a deterioration in the performance, or social issue?

E-36 Do you have a policy to deal with exceptions to a problematic situation?

[Yes I No I N/A]
□ □ □

- In the event of a system problem, whether exceptions, such as blackout, screen transition, reset to the initial state of service provision, input rejection, and decision avoidance, are handled properly should be checked.
- Whether the user of an AI system is provided with an explanation regarding why the system operation is out of order and the responsive measures in place should also be checked when the exception handling is made as described above.

E-37 Have you applied a security mechanism to strengthen the security of the AI system?

[Yes I No I N/A]
□ □ □

- When developing an AI system, through the application of an AI security architecture and deployment solution that utilizes security mechanisms such as isolation and detection, it is possible to ensure not only the security of AI data and model but also the general security of the AI system.

Security mechanism	Description and examples
Isolation	The security of the AI system is ensured by dividing the main functions used for decision –making into modular units and setting access control mechanisms between modules.
Detection	Continuous monitoring of attacks on AI systems enables comprehensive analysis of the network security status and measurement of the current risk level.

E-38 When a problem arises, do you consider human intervention?

[Yes I No I N/A]
□ □ □

- When an AI system leverages the assessment results given by the AI model to control system operation or provide information that can affect human safety and environment, human intervention is sometimes required. This applies to cases where there would be significant consequences for improper operation and function of the AI system because the uncertainty in the assessment result derived by the AI model is high.
- This trend is markedly exhibited in systems of automated operations based on AI models. In addition to the exception handling policy and security mechanisms, various methods of resolving the uncertainty of AI models through direct or partial human intervention should be considered.
- For example, when the detection result of an AI model that performs steering using obstructive object detection in front of an autonomous vehicle is unclear or has high uncertainly, a function that induces driver intervention and transfers the vehicle to the human driver control may be considered.

E-39 Do you provide guidance and responses regarding expected user errors?

[Yes I No I N/A]
□ □ □

- The external causes of user errors mostly originate from the user of the service end–product, while the internal causes generally originate from an internal system operator that generates the service product. Therefore, service personnel should understand the following user error types and define and analyze the errors that may occur related to different user error types in advance
  - ✓ Omission error: Errors caused by omitting a task to be performed
  - ✓ Commission error: Errors caused by incorrectly performing a task
  - ✓ Sequence error: Errors caused by performing a task out of sequence
  - ✓ Time error: Errors caused by not completing a within a set time
  - ✓ Extraneous error: task Errors caused by performing a task that should not have been performed
- Examples of preemptive measures for user errors are given as follows:
  - ✓ Constraint setting: To prevent incorrect user input, this measure restricts the user's selection to some extent, or defines the acceptable options for display, to prevent processing an incorrect user input. For example, in the case of an AI–based counseling chatbot, a list of questions that are frequently asked in practice is provided first; then, the user selects from the list rather than freely asking random questions.
  - ✓ System Suggestion/Correction: Types of frequent user errors are cataloged. If a similar user error occurs during actual service, the system automatically corrects the error or suggests a correct input. For example, if there is a typo in the search, correct spelling can be recommended.
  - ✓ Setting default values: It is possible to reduce user errors by providing values that are mandatory in the system and frequently used as default values first or by providing relevant examples.
  - ✓ Rechecking, result provision, revoke: The input entered by the user is rechecked and the expected output is delivered in advance. The user error can also be prevented by including certain functions, such as revoking the execution for incorrect output.

Requirement 11-2 Did you generate a report when a problem occurred with the AI system?

E-40 Have you established a reporting procedure for ethical issues, such as prejudice or discrimination?

[Yes I No I N/A]
□ □ □

- The possibility of ethical issues, such as prejudice or discrimination, in the AI system is checked. Moreover, the status of the established reporting function or procedure for these problems is checked.
- Regarding the procedure for reporting ethical problems, first, in–system standards and checklists shall be prepared to evaluate the reliability of the provided AI system. Examples of major items in the checklist are given as follows:
  - ✓ Human rights protection, privacy protection, respect for diversity, non–infringement, publicity, solidarity, data management, accountability, safety, transparency
- In addition to the automatic in–system reporting framework, a function for manually reporting to the system operator should be developed if the user identifies an ethical problem during system operation.

E-41 Have you set up indicators and procedures to assess the performance degradation of the system?

[Yes I No I N/A]
□ □ □

- For AI systems, performance degradation may occur owing to reasons such as continuous data accumulation, service function expansion, and environmental changes in the service deployment and operation phase; unlike the performance degradation causes of a general software.
- Because it is difficult to immediately determine the cause of an AI system's sudden performance degradation during service operation, whether the performance indicators and procedures are set for continuous evaluation and management of system performance degradation should be checked.
- Performance indicators may include F1–score, intersection over union, and mean average precision (mAP). If performance degradation is confirmed through the evaluation, it should be reported to the system operator. Then, the operator should prepare procedures, such as identifying the cause of the degradation and proceeding with the process of improvement.

Requirement 12-1 Have you analyzed the characteristics and limitations of the users of the AI system?

E-42 Did you analyze the detailed considerations according to the user characteristics?

[Yes I No I N/A]
□ □ □

- The main focus in the service planning stage is the preference and needs of the users, and for an evaluation of the explanations, the characteristics of different users should be considered. As example considerations of the user characteristics when children are among the service users, whether the user may have limited understanding of the graphs, terms, and vocabularies should be considered.
- Example factors to consider for a user characteristic analysis are as follows.

Category	Subcategory	Considerations
Age	Children, adults, older adults	In the case of children, the user may have a limited understanding of vocabulary and words compared to adult users.
Disability status	Disabled, Non-disabled	Limitations that may arise from physical handicaps should be considered. Examples include body size, physical abilities, and cognitive abilities.
Knowledge	Novice, Professional, etc.	The differences in knowledge level owing to the discrepancies in prior background knowledge and experience of related services should be considered.

Requirement 12-2 Did you provide sufficient instructions based on the user characteristics?

E-43 Did you establish the criteria for an explanatory evaluation according to the user characteristics?

[Yes I No I N/A]
□ □ □

- Because there are various types of users of a service, it is necessary to determine the characteristics and details for a comprehensive evaluation of an explanation. The evaluation criteria for the explanation may include categories such as the specificity, clarity, and adequacy. For subcategories, the content to be considered in each category may vary depending on the data type or modality. The following shows example evaluations of an explanation.

Category	Evaluation items
Clarity	• Are there any expressions, terms, or vocabulary that may mislead or cause a misunderstanding from user perspectives? • Is there any unnecessary explanation? • Does the explanation contain all information the user expects and aims to obtain?
Specificity	• Do the explanations use clear subjects, objects, and verbs to elicit specific actions of the user?
Adequacy	• Does the given explanation not require a specific level of knowledge of the user? • Does the user need background knowledge or prior experience? • Do you provide explanations of the terminology and abbreviations considering the reader? • Is the timing for providing the explanation appropriate?
Accuracy	• Do all of the illustrations and descriptions in the material/data provided with the explanation match? • Does the explanation of the expected results provided in advance agree with the actual results? • Does the explanation exactly match the internal algorithm?

E-44 Did you avoid the use of technical terms that are difficult for users to understand?

[Yes I No I N/A]
□ □ □

- When explanations are provided through texts, it is recommendable to avoid technical terms as much as possible in consideration of readers with different backgrounds, and to write additional explanations for terms if necessary. For example, among natural language processing technologies, one that converts a specific word in a sentence into an appropriate word tailored to the user level may be applied to the interface.

4 System Implementation

E-45 Did you use clear expressions to elicit specific actions and the comprehension of the user?

[Yes I No I N/A]
□ □ □

- A good explanation should elicit specific actions and understanding from users. It is therefore important to write the explanation in a concise and clear manner to avoid unclarity in the interpretation.
- In terms of a visual presentation, keeping the colors consistent with the outputs such as success, failure, warning, and risk can help users understand the system outputs intuitively at a glance. In addition, for an explanation provided by text or voice, an example of a clear expression is to clearly specify the object without using a demonstrative pronoun. In addition, when similar pronunciations appear consecutively, it is recommended to substitute one word into another word.

E-46 Is the location and timing of the explanation appropriate?

[Yes I No I N/A]
□ □ □

- Another important point of consideration is to place the well-written explanation at an appropriate position at a suitable timing. To this end, consideration should be given to whether the explanation should be provided once or repeated multiple times, and where should the explanation be placed to maximize the readability of the users.
- In addition, user research techniques such as a web log analysis of E-47 and A/B testing can be utilized to examine whether the position and timing of the written explanation are appropriate.

E-47 Did you use a variety of user survey techniques to evaluate the user experiences?

[Yes I No I N/A]
□ □ □

- User experience (UX) refers to all aspects of what an individual thinks and feels while using a specific product, system, or service. In addition, it includes system characteristics such as utility, ease of use, and efficiency perceived by the individual. User research techniques can be used to evaluate the explanation.
- User research methods can be largely categorized with respect to methods of approach and data acquisition methods. First, it is divided into quantitative (indirect) and qualitative (direct) user research depending on the method of approach applied by the user research technique and can also be divided into behavioral user research and attitudinal user research depending on the method of data acquisition. In consideration of the approach and data acquisition methods applied, appropriate user research techniques should be selected to evaluate the UX.
  - ✓ Classification based on method of approach and details of such methods
    - – Quantitative (indirect) user research: A method of indirectly collecting data on user behavior or attitude through tools, etc. (e.g., web log analysis, A/B testing, surveys, customer support data analysis)
    - – Qualitative (direct) user research: A method of directly observing user behavior or attitude (interviews, focus group interviews, prototype testing)
  - ✓ Classification by method of data acquisition and details of the methods
    - – Behavioral user research: A method of investigating what users do (user behaviors, e.g., web log analysis, A/B testing, eye/tracking, web log analysis)
    - – Attitudinal user research: A method of investigating what users say (user words, e.g., card sorting, in-depth interview, survey of user needs)

5 Operation and Monitoring

Requirement 13-1 Have you established measures to track and respond to the decision-making of the AI system?

E-48 Have you developed a monitoring plan for changes in the data source?

[Yes | No | N/A]
□ □ □

- Methods such as web crawling can be utilized to obtain training data for AI models. Although web crawling has the advantage of quickly obtaining a large amount of data through related open sources (e.g., Apache, Nutch, and Scrapy), the data source of the web page to be crawled may be changed in real time, or if there is a failure that makes the access to the target page impossible, the distribution of the collected data may be compromised, such as a lack of data for a particular class.
- In particular, changing the data source of an AI system that continuously performs real-time learning of crawled data may have a direct impact on its performance. Therefore, it is necessary to be able to respond to problems such as data source anomalies or duplicate collection through the monitoring of the data collection process.

E-49 Have you developed a plan to track the contribution of AI systems to the decision-making?

[Yes | No | N/A]
□ □ □

- To understand the contribution of the model to the decision-making of the AI system, information such as the output information of the previous model and the status of human intervention (e.g., system operator and user) in the final decision need to be traced.
- To this end, detailed criteria for a contribution to the decision-making of the AI system should be internally established, such as classifying into cases in which the decision is made entirely by the AI model, the decision is made by a human upon reviewing the results derived by the model, and the decision is primarily made by a human whereas the output of the model is utilized as a secondary measure, such as for only specific events. In addition, the plans and methods for tracing the contribution in the process of the system operation (e.g., log collection) should be established.

E-50 Did you implement a log collection function for tracking the decision-making of AI systems?

[Yes I No I N/A]
□ □ □

- To ensure the traceability when considering the entire lifecycle of an AI system, the continuous collection of information is necessary, such as the learning process of the model, the results of the decision-making during an operation, and the user input data. To this end, information for collection of the process-specific log of the system needs to be selected, the level of importance of the information should be defined, and the log record format needs to be determined for log collection.
- In particular, to trace the cause of errors in the AI system operation process, it is necessary to analyze the causes of the errors, including aspects of the model building method and the dataset, and thus the logs must be collected when considering these two aspects.

Types of error	Example causes of error
Errors in terms of model building method	Generation of various types of error data in terms of model building process, architecture, and learning model owing to insufficient control of the target selection of the model/data, data collection, data cleansing, data labeling, etc.
Errors in terms of dataset	Degradation of training data quality owing to inadequate dataset design, syntactic accuracy violation, duplication in data construction, etc.

E-51 Have you collected and managed user logs for the continuous monitoring of user experiences?

[Yes I No I N/A]
□ □ □

- An analysis of service usage logs can serve as the most basic method for examining not only the service operation status but also the problems experienced by users. Service logs are continuously collected during the service operation and can be accumulated in various forms according to the status of the service improvement/advancement.
- Using the log of the server infrastructure, monitoring of the service operation status can be achieved, and the user interaction log can be used to analyze which services users use the most and in which services users experience errors. To this end, log analysis software can be used in terms of infrastructure, and for users, the company may have an internal system for log collection according to the interface or call of interaction, or they may utilize a log analysis tool.

5 Operation and Monitoring

Requirement 13-2 Do you regularly manage the records of changes to the training data?

E-52 When data are changed, do you manage the different versions?

[Yes I No I N/A]
□ □ □

- In the course of AI model development, when a change in data occurs, such as updating the training data or re-applying the labeling owing to errors, the model, which is the result of training, is also changed. Therefore, when there is a change in the training data, it is necessary to manage not only the version of the training data used, but also the AI model trained based on the corresponding data version.
- To this end, the adoption of an open-source-based data version control tool (e.g., data version control) for Machine Learning projects needs to be considered, or the system of the training data version control needs to be developed in-house to apply the version control for both the training data and the model.

E-53 In preparation for data changes, have you established procedures to deliver information to stakeholders?

[Yes I No I N/A]
□ □ □

- In the AI system development process involving the participation of a large number of stakeholders, to facilitate an understanding of various actions such as the design of an AI model, tuning of the hyperparameters and retraining owing to changes in data, explanations should be provided considering the roles of different stakeholders.
- The information that needs to be provided for each type of stakeholder according to a change in data is as follows:

Stakeholders	Information to be provided
Business Decision-maker	Rather than explaining details of changes in the model according to changes in data, it is necessary to focus on changes in the purpose of the existing system, service intentions, etc., or the direction of the system as a whole.
Data Scientist	Explaining the difference in the characteristics, format, and scale of the existing data, and the changed data are required.
System Developer	With reference to the explanation of changed data, an explanation is required regarding the compatibility with the existing model, model structure redesign, detailed strategy of the model retraining (e.g., objective function, learning time, learning algorithm), and expected changes in the output.
Model Tester	An explanation is required regarding the changes in the test dataset composition, major performance evaluation results for the redesigned and retrained model, and results of the performance comparison with the existing model.
Model Operator	An explanation is required by collecting and analyzing the results of the operation and user monitoring for the changed model that was completed through the validation process.

E-54 Have you implemented measures to track the data flow and lineage?

[Yes I No I N/A]
□ □ □

- In the case of AI systems, owing to changes in data, system changes such as a model extension or redesign may follow. It is therefore necessary to continue tracing the flow and lineage of data inducing a change in the system.
- A data flow can be traced by dividing the cases into the reverse direction, forward direction, and end-to-end architecture for a data change, and the considerations for the traceability are as follows:
  - ✓ Is it necessary to record and maintain metadata for tracing the data flow and lineage?
  - ✓ Is it useful to create a data inventory, data dictionary, data change process, and document control mechanism?
  - ✓ Can data be traced back to their source?
  - ✓ Is it useful to have a "feature repository" with the Application Programming Interface (API), database, and files for tracing a data flow and lineage?
  - ✓ Should developers document the data narrative and data diary, and provide a clear explanation of the type of data used, as well as the methods and reason for the data collection?
  - ✓ Is it useful to have a data policy team to manage the data flow and lineage tracing?

Requirement 13-3 Do you periodically update the history of the training data being managed?

E-55 Do you record and manage the rate of new data among the training data?

[Yes I No I N/A]
□ □ □

- The performance of the learning-based AI model visibly degrades when a performance test is applied using new data. In particular, if the characteristics of the dataset used for training are completely different from those of the previously used dataset, or if the entire dataset is replaced, there may be a significant degradation in the model performance, and additional training may be required.
- Therefore, when new data are added, to trace the change in the performance of the AI model, the proportion of new data used for training or testing should be recorded, and the consequent change in performance of the model should be traced.

E-56 When securing new data, do you reconduct a performance evaluation of the AI model?

[Yes I No I N/A]
□ □ □

- After acquiring new data, to use the new data in the AI system, it is necessary to compare the performance with the existing AI model. Even if the new data are similar to the existing training data according to human judgment, the trained AI model may experience a difference from the characteristics of the data used in the existing training data.
- Therefore, it is necessary to carry out performance evaluation and analysis using the representative AI algorithm of the domain for new data. For performance evaluation following the acquisition of new data, refer to the following process.
  - ✓ Apply existing learning models and related representative AI models for a performance evaluation and comparative analysis
  - ✓ Select performance indicators appropriate for the target AI field and model
  - ✓ Conduct an experimental design for a performance evaluation (selection of quantitative or qualitative tests, parameter setting of test models, detailed experimental plan, etc.)
  - ✓ Proceed with the experiment and analyze the results (new data are evaluated according to the results, or if necessary, a decision is made on the redesign, extension, or retraining of the model.)

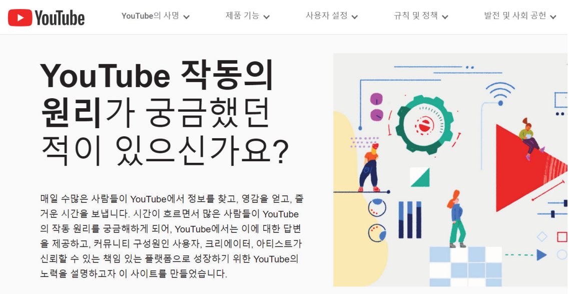
Requirement 14-1 Have you provided explanations encouraging the proper uses of the AI service?

E-57 Did you provide information on the purpose and goal of the service?

[Yes I No I N/A]
□ □ □

- A service goal includes the purpose offered by the service provider to the AI system. Objectives refer to the benefits of using the system, in particular the manner by which the user can achieve them. By explaining the service goals and objectives, the service providers allow users to select functions that are appropriate to the context of use.

Reference Service goals and objectives of YouTube



YouTube 작동의 원리가 궁금했던 적이 있으신가요?

매일 수많은 사람들이 YouTube에서 정보를 찾고, 영감을 얻고, 즐거운 시간을 보냅니다. 시간이 흐르면서 많은 사람들이 YouTube의 작동 원리를 궁금해하게 되어, YouTube에서는 이에 대한 답변을 제공하고, 커뮤니티 구성원인 사용자, 크리에이터, 아티스트가 신뢰할 수 있는 책임 있는 플랫폼으로 성장하기 위한 YouTube의 노력을 설명하고자 이 사이트를 만들었습니다.

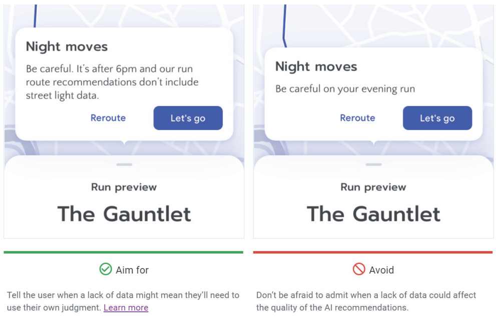
YouTube, the largest video streaming platform worldwide, states its service goals and the principles based on which it achieves users' objectives in a separate website.

E-58 Did you provide information on the limitations and scope of the service?

[Yes I No I N/A]
□ □ □

- Service providers can control user expectation by providing information regarding the limitations and scope of the service. The quality of a service outcome may be affected by various factors such as user group characteristics, the environment in which the service is used, and the data used. Accordingly, it is important to provide users with information regarding the limitations and scope of the service.

Reference Google AI+ design guidelines



Night moves

Be careful. It's after 6pm and our run route recommendations don't include street light data.

Run preview

The Gauntlet

Night moves

Be careful on your evening run

Run preview

The Gauntlet

✓ Aim for

Tell the user when a lack of data might mean they'll need to use their own judgment. Learn more

⊘ Avoid

Don't be afraid to admit when a lack of data could affect the quality of the AI recommendations.

The Google AI+ design guidelines recommend that the factors affecting the quality of a service outcome be explained to users. In this context, Google recommends that users explicitly notify the company if they experience a limitation while using the service.

Requirement 14-2 Did you clearly explain the target of interaction?

E-59 Did you provide guidance to users to help them recognize that they are interacting with AI?

[Yes I No I N/A]
□ □ □

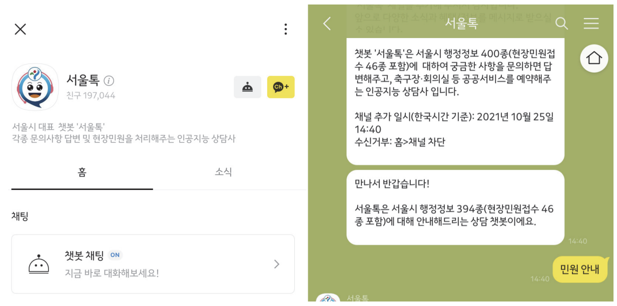
Step 1: Verify the utilization of anthropomorphism

- Anthropomorphism refers to the attribution of humanoid features to a non-human object to transform them into human-like targets of interaction to increase usability. Accordingly, if users interact with an anthropomorphized target, they should be informed that they are not interacting with a human to prevent confusion as well as to allow them to adjust their expectations.

Step 2: Explicitly state the target of interaction for each aspect of the service

- In particular, if anthropomorphism is used only partially in the system, then the scope of AI application must be stated explicitly.

Reference Example of interaction: Seoul Talk



서울톡
친구 197,044

서울시 대표 챗봇 '서울톡'
각종 문의사항 답변 및 현장민원을 처리해주는 인공지능 상담사

챗봇 채팅 ON
지금 바로 대화해보세요!

챗봇 '서울톡'은 서울시 행정정보 400종(현장민원접수 46종 포함)에 대하여 궁금한 사항을 문의하면 답변해주고, 축구장·회의실 등 공공서비스를 예약해주는 인공지능 상담사 입니다.

채널 추가 일시(한국시간 기준): 2021년 10월 25일 14:40
수신거부: 홈>채널 차단

만나서 반갑습니다!

서울톡은 서울시 행정정보 394종(현장민원접수 46종 포함)에 대해 안내해드리는 상담 챗봇이에요.

14:40

민원 안내

14:40

The Seoul Metropolitan Government operates an AI chatbot named "Seoul Talk" to simplify the processes for administration and civil petition. To avoid confusion, users are informed that they are interacting with an AI chatbot, not a human service representative, when adding Seoul Talk as a friend in a messenger service. Additionally, the Seoul Talk chat window is used.

List of abbreviations

AI	Artificial Intelligence
ALTAI	Assessment List for Trustworthy Artificial Intelligence
API	Application Programming interface
BDPL	Boundary Differentially Private Layer
CNN	Convolutional Neural Network
EC	European Commission
ETSI	European Telecommunications Standards Institute
EU	European Union
GEN	Graph Extrapolation Network
IEC	International Electrotechnical Commission
IoU	Intersection over Union
ISO	International Organization for Standardization
LDA	Linear Discriminant Analysis
LIME	Local Interpretable Model-agnostic Explanation
LRP	Layer-wise Relevance Propagation
LSTM	Long-Short Term Memory
mAP	mean Average Precision
MLOps	Machine Learning model Operationalization management
NIST	National Institute of Standards and Technology
OECD	Organization for Economic Cooperation and Development
SimCLR	Simple framework for Contrastive Learning of visual Representations
SVM	Support Vector Machine
TAI	Trustworthy AI
TR	Technical Reports
WEF	World Economic Forum
WIT	What-If Tool
XAI	eXplainable AI