Rev. 3.1.1.0077-e05 (2023/1/20)

Machine Learning Quality Management Guideline

3rd Edition

January 20, 2023
(Japanese: August 2, 2022)

Technical Report DigiARC-TR-2023-01
Digital Architecture Research Center

Technical Report CPSEC-TR-2023001
Cyber Physical Security Research Center

Technical Report
Artificial Intelligence Research Center

National Institute of Advanced Industrial Science and Technology (AIST)

Foreword and Disclaimer

This English version of the Machine Learning Quality Management Guideline is a translation of 「機械学習品質マネジメントガイドライン 第 3 版」(Machine Learning Quality Management Guideline 3rd edition), published by AIST on August 2, 2022 in Japanese.

The Guideline is assembled by the *Committee for Machine Learning Quality Management* in the National Institute of Advanced Industrial Science and Technology (AIST).

The Guideline is non-binding in relation to laws and regulations/official guidelines. Provisions described as normative in the Guideline have normative meaning only if the Guideline is adopted voluntarily. The Guideline is distributed on an as is basis, without warranties of conditions of any kind, either express or implied.

The Guideline is developed under support from the New Energy and Industrial Technology Development Organization (NEDO).

Table of Contents

1. Summary of the Guideline

The content of this Chapter is informative. Contents included in this chapter that constitute the norms of the Guideline are described again in following chapters.

The overall structure of this summary is as follows. The chapters and sections in the list below show the location of the corresponding content in the main part.

- Section 1.1 explains the background and purpose of the Guideline.
- Section 1.2 presents expected ways of using the Guideline (Chapter 2).
- Section 1.3 analyzes the reasons why quality management of AI is difficult.
- Section 1.4 mentions an overall concept of quality management process, the concept which the Guideline is based on.
- Section 1.5 presents five viewpoints of *external quality* (quality viewpoints that do not depend on implementation methods and can be evaluated only through use), which the Guideline proposes to set as goals (Chapter 3).
- Section 1.6 complements the previous section and explains why some of components generally discussed as “qualities of AI” were not adopted in the previous section and views in the Guideline on how they should be addressed.
- Section 1.7 presents nine aspects of *internal quality* (quality aspects that depend on implementation method and can be managed by measurement or processes), which the Guideline proposes to consider as *means of the quality management* (Chapters 6 and 7).
- Section 1.8 presents an overall image of the development process on which the Guideline depends (Chapters 4 and 5).
- Section 1.9 clarifies the relationship between the Guideline and various external normative documents (Chapter 11).
- Section 1.10 explains the structure of the remaining parts of the Guideline.

1.1. Purpose and Background

The effectiveness of artificial intelligence (AI), especially machine learning technology, has been accepted in broad fields of applications such as manufacturing, automated driving, robots, health care, finance and retail business, and its social implementation seems to start to blossom. On the other hand, it is difficult to identify the cause when any accident occurs or to explain advantages of AI-based products to the amount of investment due to the lack of technologies to measure and demonstrate the quality of AI-based products or services. Consequently, wider acceptance of AI in the society is lagging, causing a big obstacle to the expansion of AI development business.

The Guideline establishes a basis for quality goals for machine learning based products/services and provides procedural guidance for realizing quality through development

process management and system evaluations.

The Guideline aims to enable providers of products and services to evaluate and improve the quality of their systems so as to reduce accidents and/or losses caused by AI malfunctions in the society. Furthermore, it enables stakeholders to express their product quality using provided norms, which can be used for both commercial purposes (e.g., quoting prices of their AI-based products) and social purposes (e.g., to express their responsibility to the society).

1.2. Expected target of the Guideline

The primarily expected users of the Guideline are providers of products and services which are constructed using machine learning (hereinafter referred to as *service developers*[1]) and system developers that actually implement products and services as software. For each product, the service provider and the system developer may be either a single entity that develops products or services and provides them to end users (referred to as *self-development entity*) or two separate entities depending on the sharing of responsibilities based on contract, sub-contracting, or quasi-entrustment. Moreover, a service developer may sell implementation of machine learning components as a separate product. We primarily expect the Guideline to be used as a reference for these entities to share clear goals on required quality in accordance with situations in which products or services are used and to realize said quality throughout the system development process.

Furthermore, as a secondary usage, we expect the quality levels set by the Guideline can be referred to by end users for judging whether a particular system is safe to use. Also, the Guideline is expected to serve as a technical starting point for specifications and evaluation or certification of the quality.

The Guideline is described in a generic way that it can be applied to as broad use cases of machine learning technology as possible. For each of specific application area, one can pick required parts of the Guideline to make a specialized guideline. In addition, we expect each developer may choose to make their own specific version of development guideline to use, based on the document. The research project currently plans to publish some examples of such document as reference[221]. The following list shows some possible cases of use of the Guideline as examples.

- When an organization for AI-based system development is established for contract, etc.
  - An entity that requests to develop a machine learning based system or a machine learning component which constitutes that system (referred to as *development entruster*[2]) concretizes the Guideline in line with an application and designates it

1 In cases where software developers implement systems in advance and sell these as packaged or customized products, such developers are treated as service providers in the Guideline.

2 The “Guideline for Contract on Use of AI and Data” compiled by the Ministry of Economy, Trade and

for an entity to which development is entrusted (referred to as *development entrustee*) as specifications that serve as contract requirements.

- An entity that is to be a development entrustee refers to the Guideline as a standard for process management to guarantee the quality for a development entruster and use them as a ground or reference to calculate man-hours and the price for order.

– Design and development
  - An entity that designs and develops a machine learning based system or a machine learning component (development entruster or development entrustee) uses the Guideline as a standard for process design, system design and quality management.
– Society use
  - A self-development entity or development entruster bases its self-compatibility declaration on the Guideline with respect to quality requirement as a social norm when a system is provided to the society or used for its own business.
  - The Guideline serves as standards for social consensus on the quality of machine learning based systems in the future.
  - The Guideline serves as standards for the design and operation of a third-party assessment system on the quality of machine learning based system.

The Guideline is currently applicable for machine learning based systems implemented with supervised learning. Although the basic concept can apply to other implementation methods such as unsupervised learning, semi-supervised learning and reinforcement learning, a specific way of treating them will be added in future revisions.

1.3. Challenges behind machine learning quality management

The implementation of artificial intelligence by machine learning is, if simply said, a sort of software. However, a conventional concept of software quality management is not technically feasible to improve and maintain the quality of machine learning based systems. This section sorts out challenges related to *differences between AIs and conventional software* from several viewpoints.

Industry of Japan defines *development entruster* in the Guideline as “user”, and *development entrustee* as “vendor” or “system integrator”, as it is written mainly for business-to-business development agreements. The Guideline, however, uses the term *user* only for end users of products and services, who are finally affected by the quality of the products.

1.3.1. Importance of environmental analysis

Machine learning based products and services are often used under environmental conditions whose complexity is difficult to be understood by humans. Compared to regular programs for which humans have to analyze and identify complexity of all environmental conditions and implement every program codes as rules for judgment, there are high expectations on machine learning technology capable of automatically establishing rules for judgment based on data without having to identify detailed environmental conditions through analytical works by humans as an efficient means for implementing systems that operate under highly-complex environmental conditions.

On the other hand, from the viewpoint of ideal quality management, it is desirable to analyze environmental conditions as precisely as possible and understand risks at the early stage of system design especially from the viewpoint of quality management of systems whose compatibility to rare environmental conditions is called into question. If a machine learning technology is used for these systems, analyses on environmental conditions that used to be carried out at the time of implementing program codes in the past would be omitted. This is the reason why environment analyses at the early stage are important.

Therefore, it is important to consider how deeply the status of use and environmental conditions are analyzed in order to balance between two characteristics (efficiency of implementation and environmental compatibility) at an early stage of system design including differences from the development of conventional software.

1.3.2. Lifetime-long requirement for risk assessment

Generally speaking, a system that operates in real world and constitutes so-called cyber physical systems (hereinafter referred to as *CPS*) has risks originating from external environmental changes in addition to risks that exist in regular software systems. Different from the implementation of regular software capable of theoretically analyzing and simulating unfamiliar situations to a certain degree and taking countermeasures, a machine learning based system established based on real data may not be capable of dealing with significant changes in data trends originating from changes in our surroundings although we have some anticipations on its generalization capability.

Moreover, a machine learning based system may be designed in a way that it can attain practical quality only after additional learning using real data at the operation stage.

From this viewpoint, it is often necessary to introduce a continuous process lifecycle which integrates development and operation so that it can respond to situational changes after the system's initial operation. We need to plan a lifecycle that includes operation-time updates from the early stage of the development planning and to reflect its requirements into both the operation plan and the contracts between stakeholders.

However, ensuring a certain level of quality at the starting of operation is often necessary,

even if quality management will be carried out continuously. It is true particularly for systems with risks that compromise safety. From this viewpoint, the Guideline focuses on quality tests conducted in stages from the implementation to the beginning of operation. Moreover, effects of risks and the required quality level could be changed by varying the form of system operation at the early and full-blown stages, for example, different levels of possibility of risk avoidance due to presence or lack of monitoring or intervention of the system by humans. In this case, it is required to synchronize the time when the form of operation is changed and the time when quality management goals are changed.

Furthermore, if the system implementation is updated at the time of operation, there is a risk that sometimes the quality deteriorates, called regression in software engineering, even if that update aims to improve the quality. Therefore, we need a certain form of quality monitoring and countermeasures at the time of operation, but they vary depending on forms of development and operation. Section 4.3 sorts out some possibility for such operation models.

1.3.3. Quality assurance depending on data

A request for *quality of data* used for building machine learning is often mentioned in data-oriented development. The Guideline assumes that the quality of data itself is not the ultimate goal of quality management but rather the cause of quality deterioration or means for ensuring quality. Of course, it is almost indispensable to ensure a certain level of data quality to actually ensure the machine learning quality through the lifecycle process management such as the Guideline. In addition, when a machine learning based system is developed by sharing works, learning data may be sold, bought or shared for development. In such a case, there is a room for discussion by treating *data quality* as an independent character. Moreover, there has been pointed out at recent academic meetings that there are security risks such as contamination of learning results due to intentional injection of improper data. Effects of such improper data cannot be detected by simple numerical evaluations but require checks on data sources from broader perspectives.

The Guideline is supportive of ensuring the quality by means of numerical evaluations to the extent as possible, but at least for now, we need to secure the quality through qualitative data quality management as well.

1.4. Basic concept of quality management

(note) The Guideline defines *quality in use* as quality to be provided to end users in the whole system. On the other hand, *external quality* and *internal quality* are examined at component level[3] which will be combined into a single system.

External quality of each component is quality required as a part of system from an objective perspective. It includes, for example, security, reliability, and consistency. This external quality includes both qualitative and quantitative qualities and cannot always be expressed in single measurable indicator.

On the other hand, *internal quality* of each component refers to quality as a unique characteristic which is specifically measured or evaluated through acts of development including design when the component is created.

Based on this concept, Figure 1 explains hierarchical quality model in which external quality of each component is realized through its improved internal quality and is required for achieving internal quality of the element in one layer outside. *External quality* of the whole system is realized and provided by a provider of products/services for the sake of *quality in use* viewed from their end users.

[3] The term *internal quality* roughly coincides with the "internal measure of the quality" in ISO 25000 series. The concept behind the *external quality* is not directly associated with "external measure of the quality", because the Guideline assumes that external quality is not generally satisfactorily measurable as a numeric quality indicator. Instead, the Guideline ensures satisfaction of an external quality *level* through assessments of internal qualities.

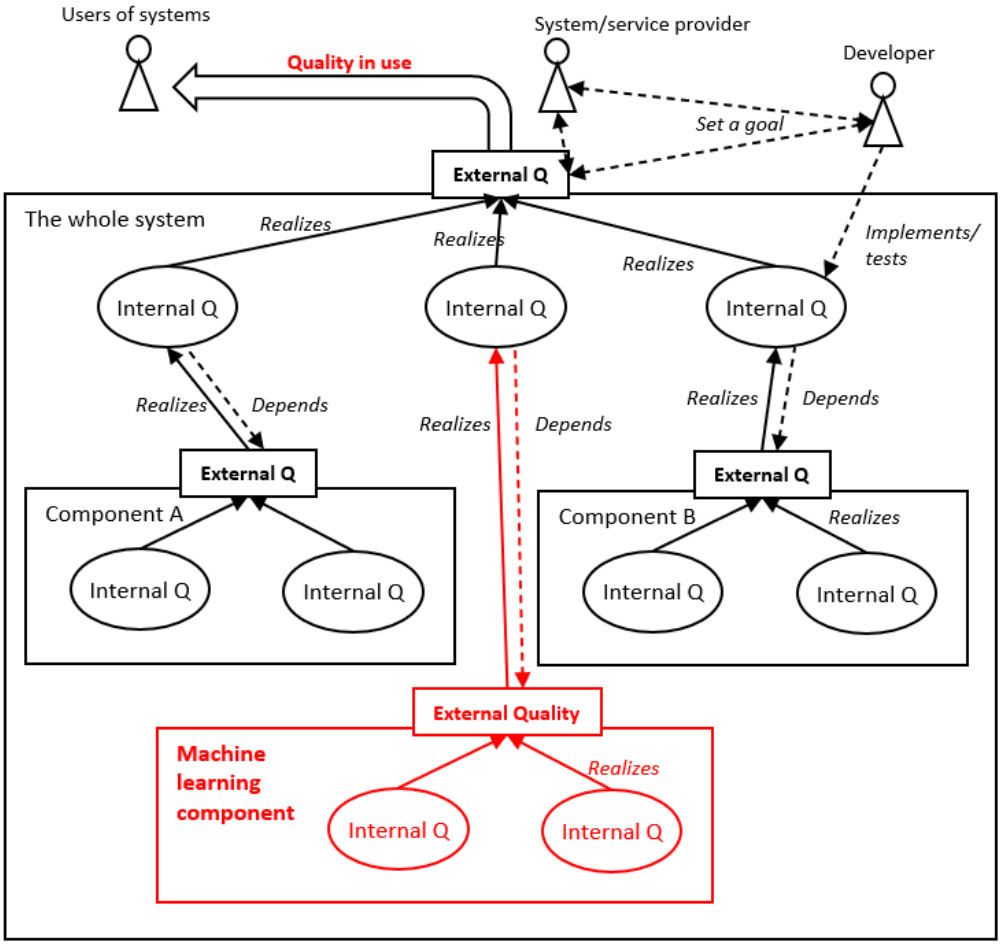

Figure 1: Hierarchical quality model

In the Guideline, *quality* of a machine learning based system itself is understood by 1) *quality in use* expected to be satisfied in the whole system in use, 2) *external quality* expected to be satisfied in components of the system built by means of machine learning, and 3) *internal quality* which constituent components built by means of machine learning uniquely have. *Quality* is understood as what satisfies a required level of *external quality* through improving *internal quality* of machine learning components and realizes *quality in use* of a final system (Figure 2).

Four viewpoints listed in Section 1.5 and an overarching viewpoint, AI security, were set as external qualities of machine learning components for target of quality management. We focused viewpoints unique to machine learning components as internal qualities to achieve said external qualities and extracted nine viewpoints listed in Section 1.7 at this stage. The quality goals were categorized by level in relation to the first four viewpoints for external qualities, and the performance goals were set for each of the nine viewpoints for internal qualities in accordance with said categorized levels in order to achieve those goals through various technologies and development process management. This is the basic concept of quality management in the Guideline.

Since qualities in use of the whole system to be realized ultimately have different focuses

depending on its application, it is not specified specifically in the Guideline (See the following supplement).

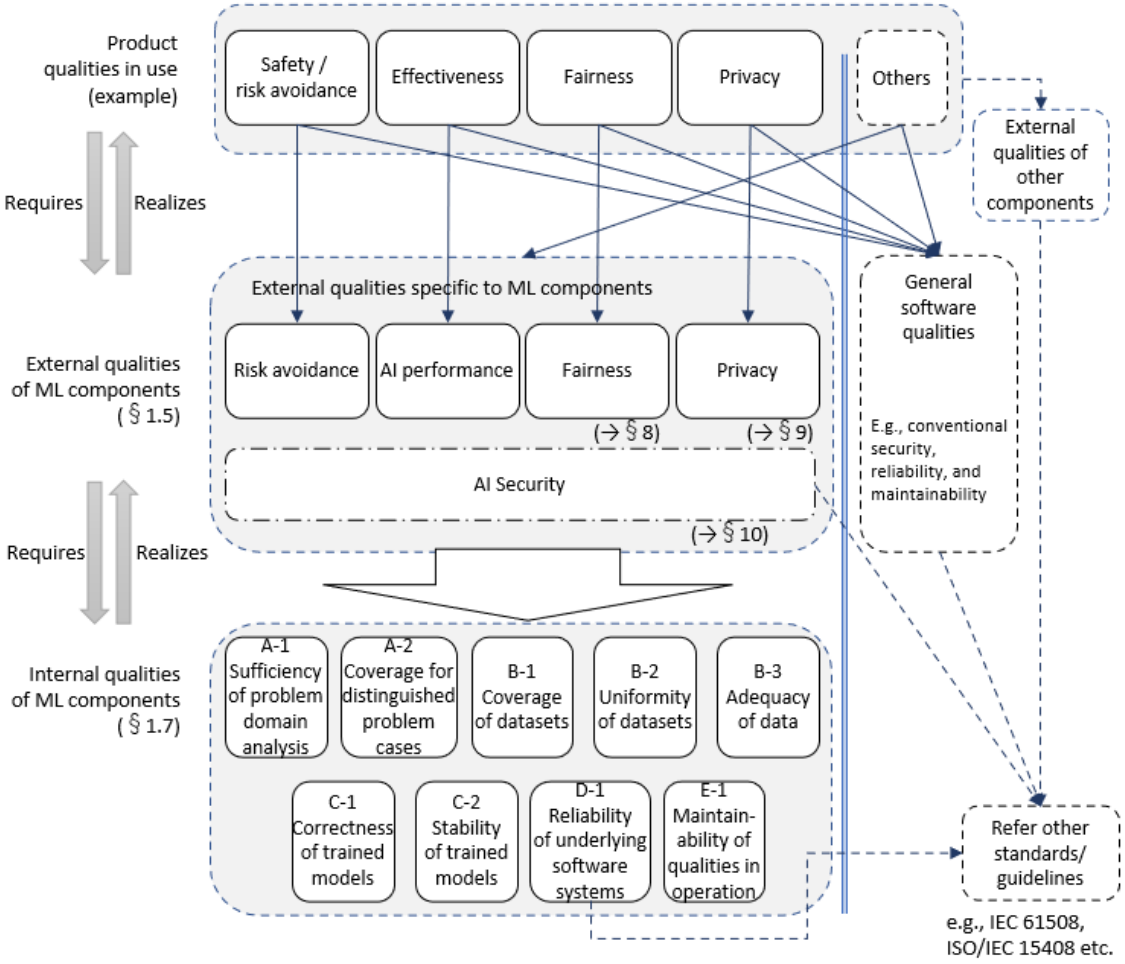

Figure 2: Overall structure of realization of product quality

Example 1) A module to recognize objects based on front images installed on a self-driving car is envisioned. One of qualities in use of automobiles is *safety for avoiding collisions with obstacles under all environmental conditions under which automobiles can be driven*. In order to realize this, the object recognition module must have a characteristic that *it can correctly recognize objects in all possible climate conditions, times etc.* as an external quality. Internal qualities required to realize this include completeness of learning conditions. This is realized by the process of, e.g., preparing machine learning training data.

Example 2) AI to estimate stock prices included in automated stock trading services is envisioned. One of qualities in use of the whole service is *maximization of profits*. A module to estimate stock prices must have a characteristic that *minimizes errors in stock price predictions and maximizes the total of envisioned trading results* as an external quality. Internal qualities

required to realize this include, for example, the precision of reasoning of machine learning. This is realized by optimizing machine learning training parameters.

(Supplement) When a product or service is developed using strict processes to evaluate safety and reliability of the whole industrial systems such as IEC 61508 [12] or IEC 62278 [15], a requirement for external qualities is of course identified in relation to machine learning components in the design process of whole conventional systems.

On the other hand, if products or services are used for IT services that have strong request for machine learning technologies, an expected level of risk is not as high as that of industrial systems in many cases. Therefore, it may not be necessary to apply a strict risk management process to the whole system development.

Moreover, if we expect machine learning components to "make a wise decision" in any way as an overall use of AI technology, the viewpoints listed in this section may often require somewhat direct response to external qualities and qualities in use of the whole system. The above two cases illustrate specific cases.

Based on such observations, Figure 2 is described in a way that qualities in use of the whole system and external qualities of machine learning components are paralleled. Moreover, five qualities in use, *safety*, *effectiveness*, *fairness*, *privacy*, and *AI security*, are exemplified.

1.5. External quality characteristics to be achieved

First, the Guideline lists the following four characteristics, *risk avoidance*, *AI performance*, *fairness*, and *privacy*, as axes of the external quality of machine learning components and specifies for each of them quality levels. The characteristics are explained in Sections from 1.5.1 to 1.5.4.

Those characteristics are, particularly in machine learning models, sometimes difficult to achieve together at the same time retaining the best possible result for each. For example, a model trained with a lot of data representing risk-prone cases to strengthen its risk avoidance capability sometimes performs poorly for the other cases, specifically in terms of accuracy. It is also known that measures to improve a model's privacy or fairness sometimes reduces its accuracy. During the development of machine learning based systems, it is often necessary to consider priorities among those characteristics and find compromises to keep up more than one of them at sufficiently high levels.

Next, the Guideline names *AI security* as a characteristic that influence all four others, a concern about malicious operations influencing from outside the system the level of achievement in the four characteristics. While measures to address AI security may similarly reduce the general performance in the other four characteristics, they can be regarded as means to maintain the other important characteristics even in a hostile environment. The scope of the problems to be treated as within a category of AI security is explained again in Section 1.5.5.

1.5.1. Safety or Risk Avoidance

The *safety*, on the machine learning component level, is a property to reduce possibilities of generating undesirable, probably harmful outputs from a machine learning component. The following are some specific examples related to the safety property:

- Oversight in object recognition for automated driving and false recognition of type thereof;
- Oversight of foreign objects in the detection of contaminants in a food factory's production line; and
- An order exceeding the acceptable level in automated trading of securities due to illegal input such as spoofing

We have set seven levels of safety, varying from the one that affects many human lives and continuity of businesses to the one that only causes minor losses of opportunities for profit (Section 3.1). Since the characteristics of safety are closely related to the properties handled in safety specifications for conventional systems, we have set four levels (Levels 4–1) corresponding to the existing specifications (e.g. SIL of IEC 61508-1[12]) in response to strong industrial requests taking the combination and affinity with those specifications into consideration. On the other hand, typically for the application of machine learning to IT services and smart devices, we added three levels from the practical point of view, because only minor damages that are not considered as risks in conventional industrial products are envisioned in many cases so that these requests for quality are categorized into the same level *Not Applicable* in the existing safety specifications.

Generally, in machine learning systems, it is not practical to strictly guarantee a characteristic that *they always operate safely at all times* and machine learning components themselves may not achieve qualities in use required for the whole system. When a system is actually built, its realization depends in many cases on *safety assurance valves* by implementing peripheral software not on machine learning components. The Guideline recognize requests for qualities in use of the whole systems and requests for external qualities of machine learning components separately in the same way as conventional specifications for safety and set levels required for external qualities of machine learning components based on risk assessments on the whole system and its configuration (Section 5.1.1.5).

1.5.2. AI performance

As the second axis of characteristics, we focus on applications to fields in which the usefulness of machine learning functions is emphasized and categorize it as the axis of characteristics of *AI performance*. A typical application in which priority is given to AI performance rather than to safety includes a scenario in which higher average performance is required than negative effects caused by individual misjudgments such as request forecasting by

retail stores and forecasts of investment decisions.

In some cases, both *AI performance* and *safety* are required. For example, on a price forecasting application, an excessive purchase could cause more indirect negative effects on management environment than expected, resulting in big economic losses that cannot be assessed based only on the average value of profits. In this case, it is necessary to clarify an optimal middle ground of two axes of characteristic, *AI performance* and *risk avoidance*.

Since specific targets to be achieved of AI performance (Key Performance Indicator (KPI)) differ from one application to another, three levels were set from the viewpoint of how much the achievement of those goals is required.

1.5.3. Fairness

A kind of social norms or ethics is often required for AI. From the engineering not humanities viewpoint, many of these social norms boil down to various existing characteristics of qualities in use and the process of their setting. Under these circumstances, we can extract *fairness* as a quality viewpoint specific to AI, which has not been taken into consideration in the past, by using a recent statistical technique.

Although conventional software must make a *fair judgment* in many cases, its realization completes mainly when examinations are made in the design stage. Therefore, other quality characteristics such as *reliability* (software operates correctly as designed) were emphasized in the implementation stage. However, in machine learning, probabilistic and statistical behaviors are included in the implementation process and learning results so that a prior examination is not enough to ensure fairness, and machine learning components as software elements must directly realize *fairness* as a quality.

From this perspective, the Guideline picks up *fairness* as the third external quality characteristic. In the Guideline, *fairness* is defined as the absence of different treatment of inputs according to factors other than conditions that the machine learning component should use as a basis for judgment, with respect to functional requirements specified at the level of the system or machine learning component. Examples of such fairness would be cases in which it is explicitly or implicitly required that there is no bias in the treatment of racial differences or gender in, for example, health insurance estimates or personnel hiring scoring. More precisely:

- *Fairness* is defined as the fact that multiple possible inputs or people who produce inputs are not *treated differently, due to differences in situations other than those defined in requirements*, under some defined criteria, compared to the requirements for situations that should be used as a basis for judgment, which are defined as functions required for judgment, classification, etc. performed by a system or component.
  - Depending on the functional definition of the system, *situations other than those defined in requirements* may be specified in a reverse way, as *differences in specific situations that should be treated equally*. The *situation* can be an attribute that is explicitly included in the value of the input to the machine learning component, or

it can be a hidden feature that is not included.

- The criteria for *different treatment* can be variously specified depending on the strength of the fairness requirements of the system, for example, *no intentionally different treatment, no statistical difference*, or *algorithmically guaranteed same treatment*.
- How to elaborate this definition of *fairness* for individual cases is discussed in more detail in Section 8.5 of the Guideline.

– In contrast, the term *bias* refers to the statistical correlation of the output of an implemented machine learning component with various characteristics of the input. By this definition, unless the output is a constant or random function, it will have some bias, but some of the unintentional biases may cause damage to fairness.

Four elements *FAST (fairness, accountability, sustainability, and transparency)* are generally pointed out as social requests for machine learning based systems. The Guideline focus firstly on fairness which can be directly analyzed as a statistical characteristic.

More detailed discussions on Fairness property, including its background, are given in Chapter 8.

The Guideline does not consider the property that certain levels of safety and AI performance are required for all attributes as fairness property, but instead as a property of safety and AI performance. For example, a request for quality concerning the effect of differences in gender or physical constitutions in protective devices for safety is sorted out as a *request for safety so that they are safe for all genders and physical constitutions*, not as a *request for no differences in safety due to gender difference*. This is because a technical way of realizing fairness can be sufficiently balanced out with other performance indicators and it is not desirable to solve the problem by improving fairness for some attribute values by reducing safety for the other values.

Furthermore, as regards this external quality characteristic, we examined *which aspects of fairness are measurable and can be subject to quality management* and *how can those aspects be realized in machine learning based components* in the Guideline. *What kind of fairness realizes social justice in a specific system* should be examined in advance outside the scope of the Guideline. These social norms and ethics will be discussed further in Section 1.6.3.

1.5.4. Privacy

Information system applications in the real world may handle confidential information that should not be widely disclosed to the public in the process of implementation. In particular, the protection of personal information, such as various attributes and preferences of natural persons, is strictly required to respect the right to privacy as a fundamental human right, and the handling of such information is often strictly regulated by laws in each country or region. Among information handled by the system, personal information stored apparently as such, e.g., gender, race, and facial images, are easy to handle in the sense that it is straightforward to carefully consider its handling at the stage of system design and operation. Meanwhile,

statistically processed information such as average values should be handled often taking into account the fact that there are situations where information is revealed even though it would be normally assumed that individuals cannot be identified (e.g., if the input value is for two persons, each person can know the specific value of the other from the average), in other words, situations where reidentification is possible despite attempts to avoid identification of individuals.

In a system using machine learning, a machine learning model that uses personal information as training data should be treated as data including the personal information of each person in some cases (e.g., when used as an authentication system to identify each person) or not in the other cases (e.g., when used to estimate the age range of a person other than the person in question), depending on the purpose of use of the system. In the former cases, it is often sufficient to treat the machine learning model with conventional privacy protection mechanisms. However, in the latter cases, it is technically unclear whether making the model publicly available would unexpectedly leak personal information that may lead to privacy infringement, comparably to or even more so than conventional statistical processing data, and it has been pointed out that information leakage does occur in some cases. In such cases, a different countermeasure against personal information leakage targeting machine learning models than the conventional countermeasure is necessary.

Assuming the relevance of the case of such unexpected information leakage to this characteristic, we decided to deal with the need to take measures to protect privacy through the development of machine learning systems as a quality perspective on *privacy*. This characteristic is treated in depth in Chapter 9. Note that although *privacy* is used in this section as a term that particularly corresponds to the strength of the aspect of social demand along with fairness, it may be technically applicable from a similar perspective to the handling of confidential information such as trade secrets and other information of companies rather than individuals.

1.5.5. AI security

Like other IT systems, AI-based systems should also be assumed to be used in a variety of malicious ways. Even when the above four external qualities are achieved as expected in normal, natural use, if a malicious person modifies the environment or input data, he or she can intentionally change the behavior of the machine learning model to prevent the achievement of these external qualities, which is generally pointed out as a weakness of machine learning. Particularly with systems that are intended to be used by general users or in a city environment, it is difficult to detect or prevent such modification of input data, and a special handling is sometimes needed if the systems are machine learning based. *AI security* is chosen as an external quality to address these cases.

As a precaution, the external quality *AI security* in the Guideline does not cover attacks that can or should be handled in terms of conventional IT security or system security. For example, falsification of training data considering particular machine learning algorithms classified as data poisoning attack is considered as a target of the quality, but attacks that intrude into the

development environment of AI models to falsify training data are not. Attacks that tamper with image data to cause software malfunctions such as buffer overruns are also outside the scope of the quality. These should be managed by the security management process as before and, if necessary, by conventional standards, such as the Evaluation Assurance Level (EAL) of ISO/IEC 15408 [3].

On the other hand, it has been reported that unlike security measures for conventional IT software, which is expected to work according to specifications when implemented accordingly, machine learning models trained with data can often malfunction with certain input data, even if the training data are within the assumptions of the specifications (e.g., image data from a real environment that have not been altered after their capture by the camera) and that such input data can be intentionally created. It has also been pointed out that when publicly available data or machine learning models, such as open-source software, are used to build systems, malicious modifications that are difficult to detect can be included in those public data. Against these problems, conventional approaches such as pre-checking of software security do not work, and they should be addressed concurrently with other quality perspectives during the development of machine learning systems.

The quality perspective of *AI security* to be addressed during the development of machine learning models is discussed in detail in Chapter 10, including risks, examples, and possible measures.

1.6. Possible other aspects for AI quality

The Guideline set forth that the quality is managed with a focus on five viewpoints (safety, AI performance, fairness, privacy, and AI security) as described above. Generally, various viewpoints are under discussion concerning *the AI quality*. In this section, we summarize our thoughts on the relationship with the Guideline for some perspectives other than the five we have organized so far.

1.6.1. Explainability of AI

The *explainability* of AI systems is sometimes emphasized as one of the components of trustworthiness. In fact, there are several aspects to considering explainability as a quality. First of all, from the perspective of *explainability of being able to use the product with confidence*, or in other words, explainability of quality, the Guideline considers this as a property required for quality management rather than quality itself. Throughout the Guideline, quality management is viewed as a series of activities that enable quality management activities to be explained and accepted afterwards with conviction, and explainability from this perspective is the very purpose of the Guideline.

Next, from the standpoint of the explainability of the behavior, explainability in this sense

can be considered as one of the means to achieve quality explainability, which is the objective of the Guideline. Indeed, if the behavior of a machine learning system is fully explained, it is highly likely that its behavior can be logically analyzed and its quality explained in the same way as an ordinary program. And even if the behavior cannot be perfectly explained, machine learning elements with a simple structure may be expected to be easier to explain the content of the behavior, and at the same time have lower quality concerns that may lead to different behavior than expected. On the other hand, however, when the environment is complex as described in Section 1.3.1, even ordinary software programs that would logically be perfectly explainable often do not necessarily lead to reliability, and the explainability of behavior and the explainability of quality are not always considered to be the same thing.

1.6.1.1. Explainable AI

The idea of *explainable AI* is often referred to as a form of achieving explanatory behavior. Once the technology for explainable AI is well established, where the behavior of AI can be logically re-described and organized at a level that is sufficiently understandable, it is expected that the achievement of various qualities can be logically explained from the standpoint of quality management. However, at least at present, such technology is still under research and has not reached the stage where it can be adopted stably and universally as an all-purpose means to achieve quality. Of course, if the technology can be applied, it can be a powerful means of achieving quality in relation to these guidelines. In particular, it could be a strong means to obtain explanations of internal quality specific to fairness (Chapter 8) and stability of machine learning models (Section 6.7, Internal Quality C-2). It should also be noted, as mentioned above, that even if the algorithmic computational processing of machine learning elements can be perfectly explained, it does not always explain their quality in the real world.

1.6.1.2. Transparency

For the same reason, we consider transparency to be one of the means of achievement for quality management. At least in the field of machine learning, transparency of behavior is a rather similar property to explainability of behavior, and may be useful in the analysis of fairness and stability. Also, in applications where the resulting machine learning model is not used as it is, but is rather designed logically based on the parameters inside the obtained model and implemented as a rule base or as an ordinary program not considered as a machine learning model in the Guideline, transparency and explainability as an ordinary program may be considered.

1.6.2. Social aspects such as ethicalness

As qualities required for machine learning based systems, ethicalness and social validity of results and judgments made by machine learning components may be included. For example, in fields closely related to personal information (e.g., prior scorings for hiring, crime forecast, etc.), it may be required to guarantee that differences in gender or races do not affect judgments legally or socially. Moreover, in fields that are likely to cause human and economic damages, there may be a case where the validity of possible judgments with different risks is called into question (e.g., a judgment made by a self-driving system under a situation where some human damage is inevitable).

The Guideline do not directly examine *what kind of judgment* machine learning components should make to have social validity in those cases, since it should be sorted out by humans in advance as a part of required definitions at the beginning of system development. Then, a machine learning based system draws out processing results with a high probability to the extent possible toward *correct* output sorted out by humans, and this is considered as a request for qualities in use.

As such, *fairness and privacy*, which have not been treated as a direct external quality in conventional software engineering in particular, is listed in Sections 1.5.3 and 1.5.4 and Chapters 8 and 9 as one of the external quality axes. As for ethical aspects, we summarize some international efforts in Section 11.2 for reference.

1.6.3. Limit of response to complex external environment

As a general problem of AI, discussions on the limit of complexity of environments where AI can be utilized can draw attention. Machine learning based systems are often incorporated into a part of so-called cyber-physical systems in open environments such as streets and public spaces. Therefore, it is impossible for AI to make expected judgments under all environmental conditions if ultimate conditions are included. This issue is inherently common not only to machine learning but also to devices and software that operate in open environments. Conventional specifications for reliability engineering such as IEC 62998 [18] seem to intend to tackle this issue somehow.

The Guideline follow the overall concept of past specifications. The adequacy of risk analysis and system design required for realizing the quality in complex external environments becomes subject to examination of quality management methods with regard only to differences in characteristics from conventional software and points to remember for unique analysis and design originating from those differences.

1.7. Internal quality characteristics subject to quality management

In the Guideline, the following nine characteristics, in five categories, were extracted as characteristic axes of quality management, for the achievement of five external qualities described above. Section 12.1 describes the outline of analyses leading to this extraction.

- A: Designing quality structures and datasets
  - A-1: Sufficiency of problem domain analysis
  - A-2: Sufficiency of data design
- B: Dataset quality
  - B-1: Coverage of datasets
  - B-2: Uniformity of datasets
  - B-3: Adequacy of data
- C: Quality of machine learning models
  - C-1: Correctness of trained models
  - C-2: Stability of trained models
- D: Quality of software implementation
  - D-1: Reliability of underlying software systems
- E: Operational quality
  - E-1: Maintainability of qualities in operation

In addition to this section and Chapter 6, we will reorganize our thinking on *fairness* and *privacy* in Chapters 8 and 9, with respect to perspectives specific to each.

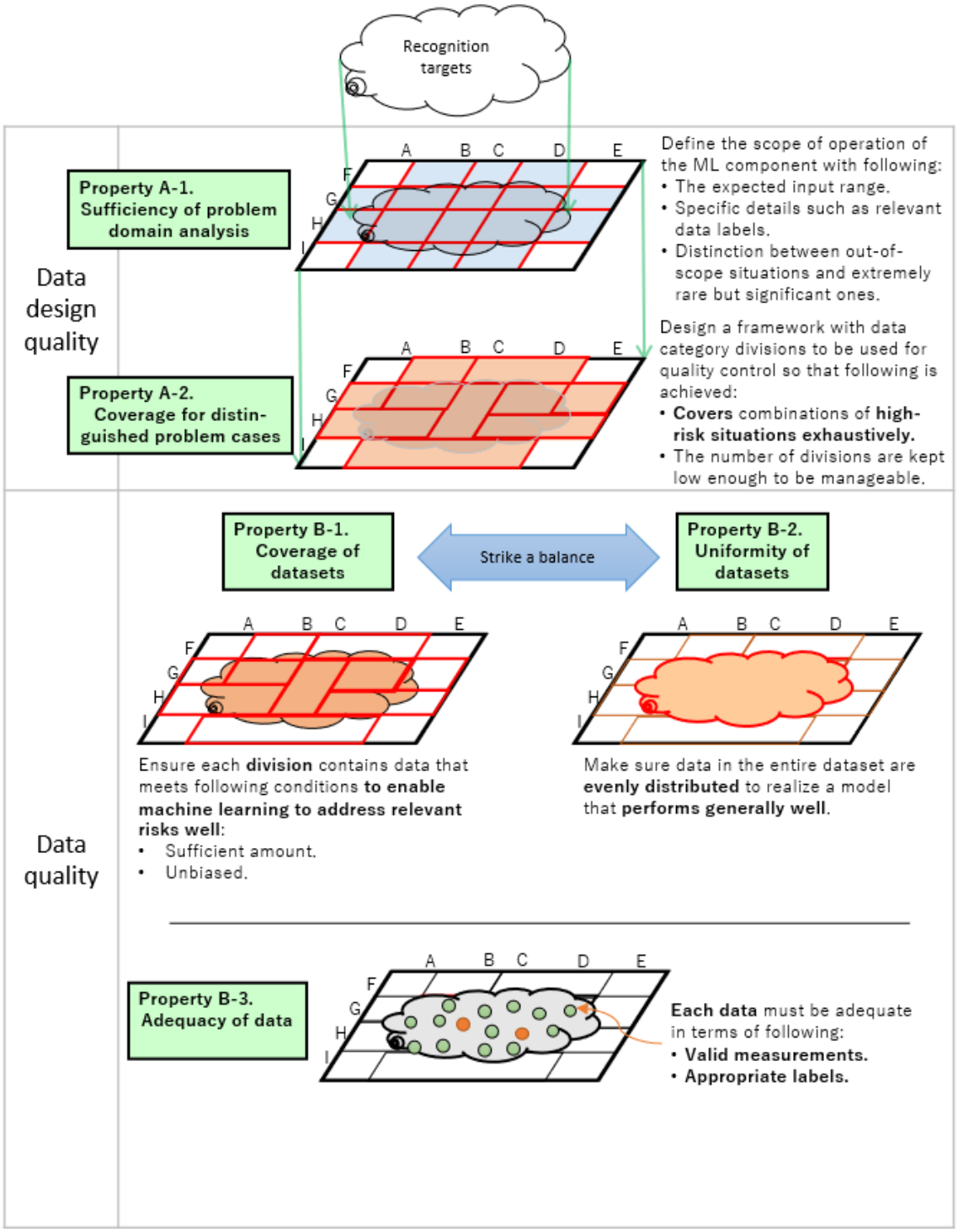

Figure 3: Internal quality characteristics focused (1)

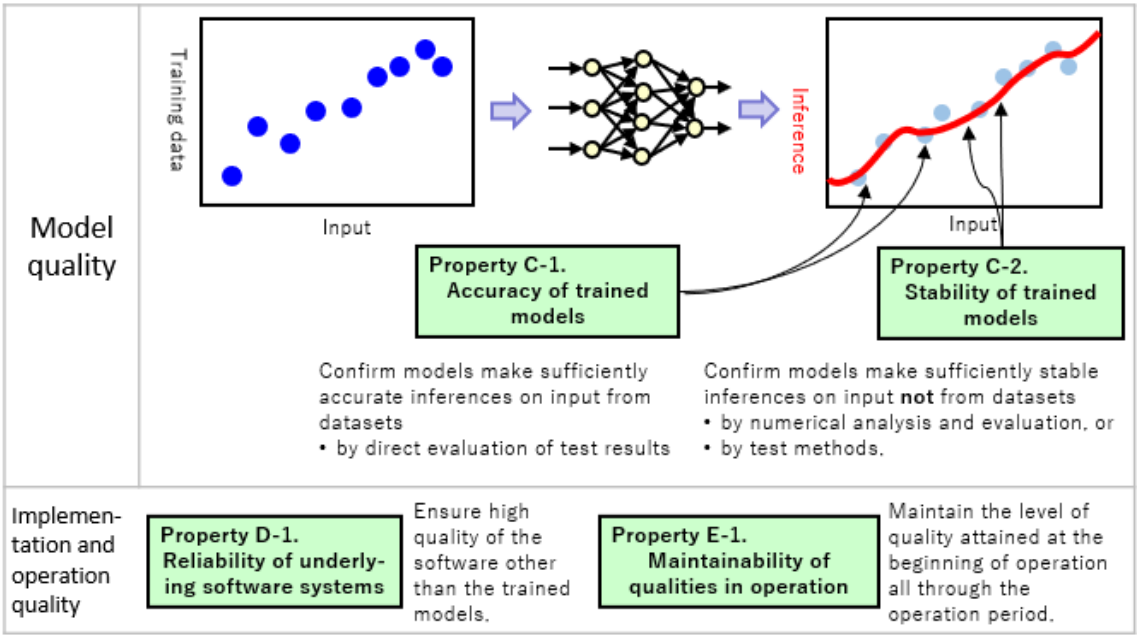

Figure 4: Internal quality characteristics focused (2)

1.7.1. A-1: Sufficiency of problem domain analysis

First, we define *sufficiency of problem domain analysis* as the fact that the analysis of the nature of the actual data, expected to be input to the machine learning component during operation, corresponds to the real-world usage of the machine learning application system, and that the results of the analysis cover all expected usage situations.

There are different ways of setting a specific *axis* for data analysis. The basic concept of the Guideline envisions a concept of feature tree in conventional software product line engineering and a more simplified method of classifying and organizing *the axis* as itemized independent conditions and capturing specific uses as their combination (See the following examples). Moreover, we conduct at this stage both top-down analysis from the request side such as risk analysis, failure mode analysis and bottom-up analysis consisting of preliminary data analysis in the Proof of Concept (PoC) stage to sufficiently sort out possible situations in which external qualities change or deteriorate.

(Example 1)

We can write down the following situations which a self-driving vehicle running outdoor might encounter in machine learning components that recognize traffic lights based on images. (Please be reminded that this example is not sufficiently comprehensive for actual application)

Significance of display: Green, yellow, red
Time zones: Day, night
Weather conditions: Sunny, cloudy, light rain, heavy rain, snow, fog, etc.

Others:

For example, the situation in which “the traffic light is red on a clear night” is one *combination of situations*.

(Example 2)

We can write down how a machine learning component that predicts the sales trend of a retail store is used as follows (this example is not comprehensive, either).

Day: Monday, middle day of week, Friday, Saturday, Holidays
Weather conditions: Sunny, cloudy, light rain, heavy rain, snow
Time zones: Morning, before noon, noon, afternoon, evening, night, midnight
Season: Spring, summer, fall, winter
Neighboring events: Yes/No

In this example, one combination could be “there is a neighboring event on a holiday morning under winter clear sky”.

The sufficiency of this requirement analysis deals with analysis of risk factors in conventional software and test designs to include those risk factors when a black-box test is conducted. It is also an important characteristic that establishes a unit to grasp and check the quality. On the other hand, when machine learning is implemented, minor characteristics above a certain level may be left to learning processing in the training stage, and a person who implements the system may not give detailed instructions on how to judge individual conditions. We can say that these are the biggest benefits. Moreover, in some cases, the implementation by machine learning shows better performance than that by humans. From this viewpoint, the configuration of *details to be included in requirement analysis* becomes a very important point in examining an implementation strategy of machine learning components taking the quality into consideration.

Furthermore, when this type of analysis is made, we may recognize a situation which we do not encounter in real life and an extremely rare situation in which it is not practical or it is not required to respond to guarantee operation (e.g. snowfall in summer) or a rare situation that does not appear in training data to be collected but it is required to respond to (e.g. snowfall in spring in the Tokyo area). Specifically, it is very important to distinguish those two situations in any system for which safety is required and the distinction is directly connected to design of safety/robustness of the whole system. At this stage of considering the *sufficiency of problem domain analysis*, it is also important to identify situations that cannot happen in real life and eliminate them from examinations in later stages, in the process of identifying rare situations or cases difficult to be found based only on data collected from real life that require response and designing corresponding systems.

A specific concept of configuring these situations will be explained in Section 6.1 in more

detail.

1.7.2. A-2: Sufficiency of data design

Assuming sufficiency of problem domain analysis, sufficient examinations of data design in response to various situations to which systems must respond, for collecting and organizing sufficient training data and test data, are required as *coverage for distinguished problem cases*.

In an extremely simple system, it is enough to mention that corresponding data to all *combinations of situations* identified in the problem domain analysis described in the previous section included in training datasets and test datasets. However, if a situation in which a system is envisioned to be used is complex, the number of possible combinations becomes enormous. Therefore, it is not practical to cover all combinations with datasets. For example, only in the simple case mentioned in the above Example 2, the total number of combinations is 700. In a real case, this number is envisioned to reach 10,000. In this case, it is required to check completeness with a rough granularity level that covers several situations to sufficiently deal with combinations of detailed situations in which any danger or reduction in performance is likely to occur. There is a concept of *coverage criteria* applied to design of tests in the field of software engineering which aims to achieve practical and sufficient operational completeness by selecting appropriate means for each application.

1.7.3. B-1: Coverage of datasets

Next, we require that enough data, especially, test data, are given to each *combination of situations to be supported* designed in the previous section without any missing situation, to be called as *coverage of datasets*. Although the property mainly concerns datasets for testing and validation for the purpose of quality management, to achieve targeted levels of qualities, it matters also on datasets for training.

When a regular software is developed, the details of all features in real world which software operation depends on is grasped in any point from problem domain analysis to implementation and reflected ultimately as conditional branching or calculation formula in programs. However, when machine learning components are built, more minute situations than a certain degree are not grasped explicitly as *feature quantity* or *ground truth labels* of training datasets and are only included implicitly in training datasets. They will be reflected in final operations throughout the training stage of machine learning. The purpose of configurating this axis of characteristic is to guarantee that the shortage of learning due to the shortage of data or any oversight of learning in specific conditions due to biased data does not occur in any situation or case identified in requirement analysis or data design.

(Example 1)

In the case of recognition of images of traffic lights mentioned earlier, this characteristic means that all forms of traffic lights in each prefecture, their height and distance between them, and road conditions around them are included as data without any bias and that training is not carried out based only on limited data of a specific city.

(Example 2)

When image recognition AI that aims to recognize *cats* from small animals around town is to be built and individual characteristics of cats such as breed and size are excluded from recognition, this characteristic means that sufficiently diverse images of *cats* and *other small animals such as dogs* are made available as data and that training does not only target specific breeds (e.g. calico cats) in environments where the product is expected to be used.

1.7.4. B-2: Uniformity of datasets

A concept contrary to *coverage* mentioned earlier is *uniformity of datasets* in relation to the overall assumed input data. When each situation or case in datasets is extracted in accordance with the frequency of its occurrence in whole data to be input, data are considered as uniform (Figure 5). Balance between this uniformity and coverage sometimes needs proper attention and evaluation. The prediction accuracy of machine learning technologies is generally supposed to improve by using samples extracted uniformly in relation to input environments as training datasets. However, the coverage of situations explained in the previous property may be emphasized depending on actual application and quality characteristics required therefor. It is necessary to consider priority or compromise between coverage and overall uniformity carefully.

When safety is strongly requested, sufficient training data are required for high-risk situations that must be avoided by making correct judgments. If such a rare case occurs and you intend to train data by ensuring enough data in relation to that rare case and maintaining the uniformity to all other situations, the necessary amount of data might be enormous. In this case, priority may be given to training of rare, high-risk situations.

On the other hand, when overall performance, AI performance is requested, certainty of inference results in other cases might deteriorate by giving more priority to the training of rare cases than the actual probability of occurrence, thereby resulting in the deteriorated average performance as a whole. In this case, standards for coverage for each situation mentioned in the previous section are not appropriate.

Moreover, when fairness is strongly requested, *what kind of fairness* is requested may change our decision. For example, you may choose to give the cases artificially-equal training by artificially processing data (selection or addition of data) or let the cases learn at random in line with the distribution of extracted training data.

Since the two viewpoints mentioned in Section 1.7.3 and this Section may be compatible or incompatible, you may need to adjust appropriate training data to strike a balance at an adequate level. Moreover, different characteristics may be sought in the training stage and the quality test

stage.

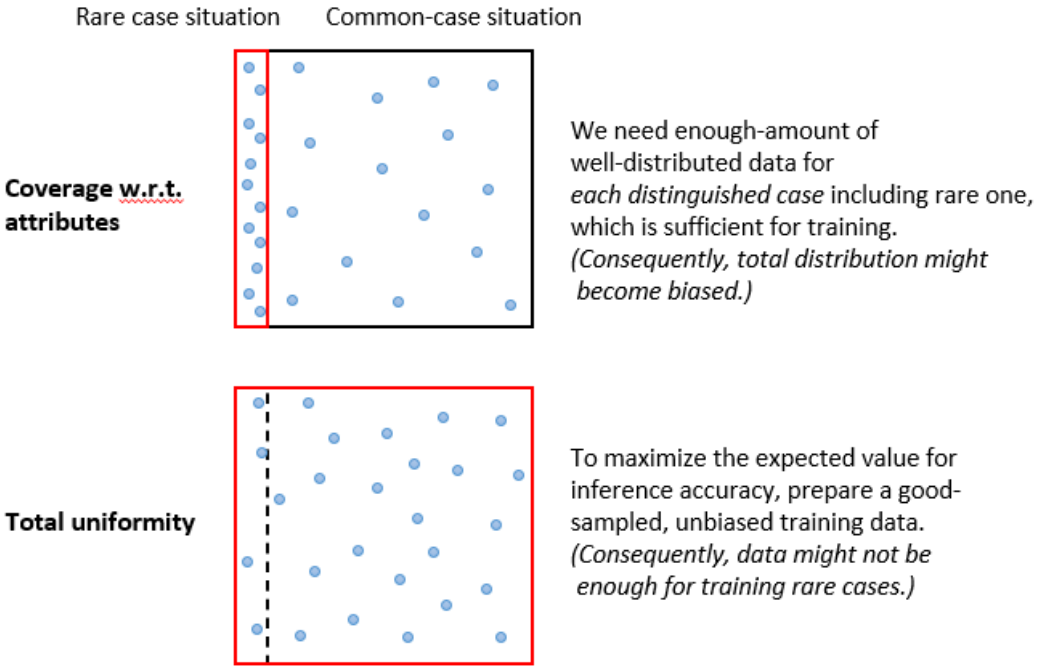

Figure 5: Relationship between coverage and uniformity

(Example 1)

For example, if snowfall may affect the performance of signal image recognition in an self-driving vehicle, the necessary training or verification data must be prepared to suppress the effects of false recognition even in snowfall conditions that are expected to occur only one or two days a year in Tokyo, which is set as the operational area. For this reason, the proportion of snow images in the dataset may be increased compared to the actual probability of snow events.

In this case, it may be necessary to prioritize the *coverage of the dataset* over the *uniformity of the dataset*.

(Example 2)

On the other hand, if one or two snowy days a year are strongly trained in sales predictions of a retail store in Tokyo, the prediction performance in other climate conditions may deteriorate and the average profit cannot be maximized. In this case, we might need to give priority to *uniformity of datasets*. Of course, what kind of data is actually prepared as training datasets will finally be examined in the implementation process taking the balance of both internal quality characteristics into consideration.

1.7.5. B-3: Adequacy of data

In contrast to the properties of the distribution of the dataset in B-1 and B-2, the *adequacy of data* refers to the fact that each item of data in the dataset is appropriate for the purpose of training. Adequacy includes not only the absence of errors in the values, but also the absence of data that should not be used for training even if the values themselves are accurate as consistency, the absence of inappropriate modifications to the data as authenticity, and the fact that the data are sufficiently new as currentness. In the context of supervised machine learning, two perspectives are also included: *adequacy of data selection*, which is the adequacy of the measurements to be trained where values that correspond to the input side when the machine learning component is viewed as a function, and *adequacy of labeling*, which is the adequacy of the correct answers added for training where values that correspond to the output side.

The quality of the adequacy of the data are primarily determined according to the requirements definition for the implementation of the machine learning components. In general, the quality needs to be reevaluated when the requirements definition is updated, while there are some aspects that can be judged as general-purpose data (e.g. removal of obviously invalid data, authenticity and traceability of the dataset as a whole). Even in the development process of a single system, when requirements definitions and labeling policies are updated due to trials in the PoC phase or regressions in the production phase, the data need to be rearranged to accommodate the new requirements and policies, so it is important to carefully consider the man-hours and procedures in the development process.

Although the Guideline aims to evaluate validity through inspection of data as much as possible, there are some aspects that are particularly difficult to check from the data itself, such as credibility and traceability, and some aspects, such as uniformity of labeling policies, that should be achieved through process management.

1.7.6. C-1: Correctness of trained models

The term *correctness of trained models* represents that a machine learning component functions as intended upon the input from the training dataset consisting of training data, validation data, and test data.

Normally, the training dataset does not provide sufficient information to reflect all input given from the in-operation environments. Therefore, a trained model might show high performance with respect only to training data, but not to the data other than the training data. In other words, a trained model may overfit to the training data and may not generalize well to unseen data. When a machine learning component is evaluated using only the validation datasets and test datasets, it is not possible to check if the machine learning component functions as intended when given unseen data from the environment. Hence it is important to evaluate not only the correctness but also the stability, which will be explained in the next section.

1.7.7. C-2: Stability of trained models

The term *stability of trained models* represents that given an input that is not included in the training dataset, a machine learning component behaves in an expected way. When a trained model does not satisfy stability, it may not function correctly upon the input data that are not included in the training, validation, or test datasets. For example, a trained model may function incorrectly when the input is perturbed by natural or adversarial noise.

1.7.8. D-1: Reliability of underlying software systems

The property *reliability of underlying software systems* represents that the underlying conventional software (e.g., training programs and prediction or inference programs) functions correctly and reliably. This notion includes various software quality requirements such as correctness of algorithms, time or memory resource constraints, and software security as well.

1.7.9. E-1: Maintainability of qualities in operation

The term *maintainability of qualities in operation* means that internal qualities fulfilled at the time when the operation started is maintained throughout the operation period. The maintainability implies the following three aspects; internal qualities sufficiently adapt to changes in external environments, they do so also to changes in the system feeding input to the AI, and they do not deteriorate due to changes made in trained machine learning models to make such adaptations.

A specific method to realize quality maintenance depends on forms of operation, especially on how to carry out additional learning or iterative training. This point will be described in Section 6.9.1.

1.8. Concept of development process model

1.8.1. Relationship between iterative training and quality management lifecycle

Machine learning based systems broadly apply more adaptative and flexible development processes such as the development of conventional waterfall process, the introduction of preliminary experiment stage called PoC and agile development. Moreover, continuous development that realizes higher quality in operation and responds to environmental changes based on real data obtained in the operation stage is thought to be effective in machine learning.

Hence, the Guideline defines a system lifecycle process that integrates development and operation, which consists not only of early/late-stages of development in which software is

actually built but also of the stages of formulating system specifications and operation after deployment. Especially, the Guideline considers the following two processes important and treats them as a part of initiatives for quality management; grasping quality goals in the problem domain analysis stage prior to system design, and monitoring behaviors, obtaining additional data, and carrying out additional learning during the operation stage.

The whole lifecycle is modeled as a hybrid process integrating agile or iterative development process in which the progress is managed based on quality inspections or tests and the whole top-down V-shaped development process focusing on process management (Figure 6).

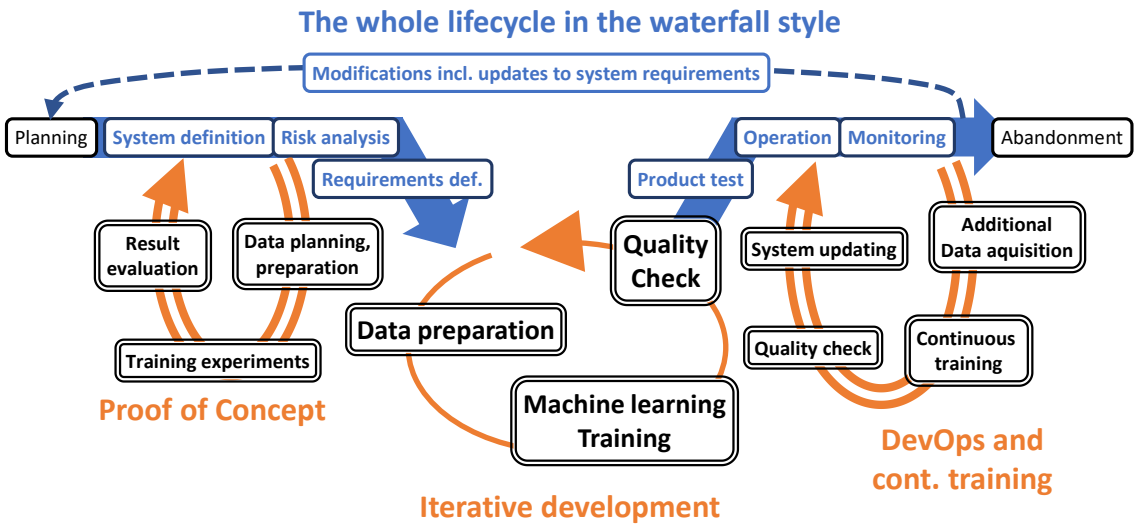

Figure 6: Conceptual diagram of mixed machine learning lifecycle process

Specifically, works to analyze requirements necessary for the whole system quality are regarded as a flow process similar to a top-down process and assumed to be analyzed in detail as an independent process before start of actual system development and data reduction. Reworking and iterative works at this stage suppose that the same level of consistency is ensured as final results even if works are restarted from the middle of the development process. Moreover, the process of obtaining additional data in operation and the monitoring process are positioned as a part of the iterative quality management process based on the concept of DevOps and should be compatible also to the concept of the top-down operation phase in RAMS specifications where necessary.

The stage of so-called PoC development which is seen often in the development of AI is categorized as a part of the process of the prior analysis stage leading to the definitions of requirements as an important quality management stage. This is a concept of identifying and categorizing requirements for quality again, when knowledge and data obtained from PoC are utilized, so that quality management in the main development stage and free explorative development in the PoC stage are balanced. In the actual development process, a trouble of re-categorization can be saved by partially introducing the concept of quality management of the main development stage in the PoC stage.

The lifecycle process in the Guideline is a *reference model*, and development processes carried out by respective developers can be designed as their own. This reference model intends to help readers understand individual stages of the development process created in accordance with different circumstances *in comparison to* the Guideline and find out how quality management technologies included in each chapter of the Guideline correspond to development stages taken care thereby.

(Note)

The development process of AI has been often categorized as a non-waterfall iterative process. When iterative training is given, various works such as addition of collected data, change of selection and adjustment of parameters are carried out other than changes in algorithm implementation codes. Sometimes, the goals are achieved by modifying the principles for implementation and specifications based on those changes, giving training in accordance with new specifications and improving the accuracy. On the other hand, *gates of quality* are often established in the form of check before operation and inspections on orders, even if products that require quality adopt a circular development process model in reality.

A model shown in Figure 7 exemplifies, based on the development process of AI understood as a iterative system, its relation with the mixed process shown in Figure 6 which the Guideline basically assumes.

To be more specific, several development tests conducted until policies for implementation (also called specifications for implementation) are concretized are mapped to several circulations in the PoC stage to examine the quality. Then, one or several tests immediately prior to the operation stage leading toward training/quality tests and release based on the final specifications are positioned as *main development stage* in the mixed process.

Occasionally, the specifications need to be modified again after going through the main development stage. From the perspective of waterfall process, this is a step back from the implementation stage to the requirement analysis stage, but from the perspective of iterate process, we do not need to consider it as a serious step back, because *it was found out that the product was still in the PoC stage*. The PoC development stage has an aspect of implementation forecast and knowledge obtained in this stage is reflected implicitly in the main development stage. Therefore, even if this process is considered as a waterfall type, efforts for implementation in the PoC development stage are not in vain.

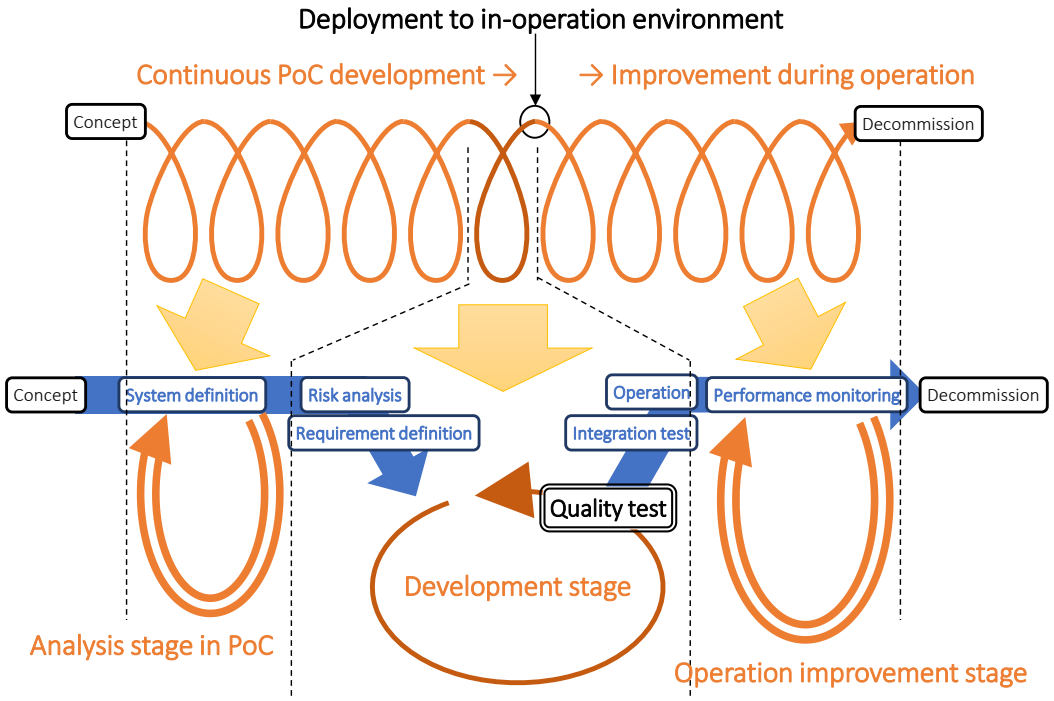

Figure 7: Relationship between the circular process modeling and the model in the Guideline (example)

(Note 2) Section 4.1.1 (Page 49) describes changes in operational situations and the concept of operation with various gates as an example of application.

1.8.2. Development process by multiple stakeholders

In the actual development of AI, works may be divided in various forms. For example, a service provider itself may take care all of planning, development, and operation. In another case, it may outsource only the training stage. Yet in another case, it may entrust stages from system design to a model's development, training, and incorporation into the system. In the quality management process defined in the Guideline, a development entruster responsible for product planning plays a central role and quality management activities are carried out by everyone including developers, etc. Management works for individual stages in the process are shared upon agreement between workers. As a result, a development entruster or development entrustee may take responsibility for setting qualities in use and external qualities based on the purpose of applied systems depending on the form of division[4]. In any case, it is important to

[4] In cases where a requirement for validation that *certain performance indicators (accuracy, etc.) should be achieved on data given by a development entruster* is established as a form of order requirements common

reach agreement to ensure sufficient quality for end users together, and clearly reflect the share of costs in the contract so that activities in the quality management process that continue until the operation starts are not disrupted.

Sections 5.2 and 5.3 explains some sharing models and remarks when several entities involve in development such as consignment development and delta development.

1.9. Relations with other documents and rules

1.9.1. Social principles on human-centric AI

In Japan, the Cabinet Office has defined an ideal relationship between operators and developers of machine learning based systems and the society and public in the form of "Social Principles on Human-centric AI" [22].

In relation to those Principles, the Guideline are primarily positioned as non-binding guidance which sorts out technical matters developers themselves refer to and practice in order to improve the quality and prevent safety and security from being compromised, when operators and developers fully understand those Principles and actually develop machine learning based systems and machine learning components therein (Figure 8). Moreover, the Guideline are, together with other guidelines and future international standards, also positioned as an element that constitutes "Social Principles of Human-centric AI" mentioned therein.

particularly in the development of AI, we need to take note that the development entruster is responsible for ensuring so-called *data quality* where *the data is appropriate for learning consistent with the purpose of products* as explained in Sections 1.7.1 to 1.7.4 of the Guideline. In cases where the development entruster is not capable of validating or ensuring data quality, it is necessary to explicitly entrust a part of the work in the form of *design support work* and consider a possibility that the system cannot be built due to the lack of data quality when it is actually trained and the work is returned to the development entruster.

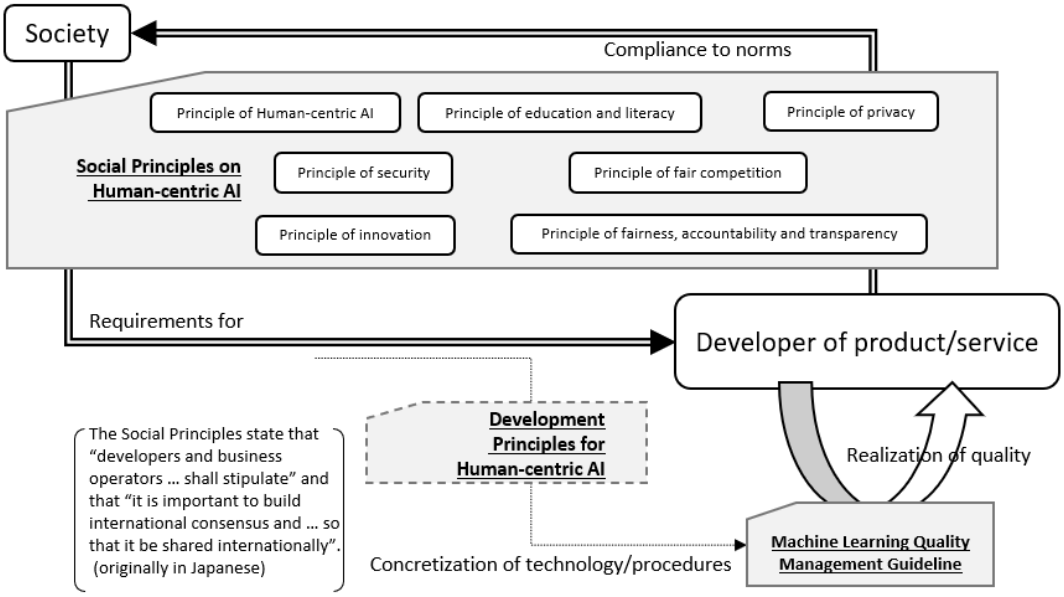

Figure 8: Relation with *the Social Principles on Human-centric AI*

1.9.2. AI-related rules and guidelines of foreign governments and international organizations concerning AI technology

In addition to the above principles, norms for social nature, safety and ethics concerning the development and use of AI technology have been documented recently in different forms [21][26]. As for their relationship with the Guideline, those social norms are positioned under verbalized documents as in the case of the Social Principles of Human-centric AI mentioned in the previous section and categorized to present one specific method for realizing some of them in system development.

1.10. Structure of the Guideline

The structure of the remaining parts of the Guideline is as follows.

- Chapter 2 sorts out the scope of the Guideline and their relationship with existing standards again.
- Chapter 3 sorts out the details about qualities in use mentioned in Section 1.5 including decisions on leveling.
- Chapter 4 shows reference models about the development processes whose outline was explained in Section 1.8.
- Chapter 5 mentions specific methods of operating and applying the Guideline.
- Chapter 6 sorts out the internal quality characteristics mentioned in Section 1.7 in more detail.

- Chapter 7 sorts out possible ways of realizing each internal quality characteristic described in Chapter 6.
- Chapter 8 details issues pertaining to the management of fairness, one of the external qualities.
- Chapter 9 describes the management of privacy, another external quality.
- Chapter 10 elaborates on approaches to security issues in the machine learning quality management.
- Chapter 11 shows reference information such as the relationship with other guidelines.
- Chapter 12 shows the process of analysis and concept until the Review Committee of the Guideline clarifies the internal qualities listed in Section 1.7 (and Chapter 6) as reference information.

(Note)
The Guideline can be read in the following ways (other than reading in order from the beginning).

Read Chapter 5 to understand the outline of the procedures for processes.

Read Chapter 3 to determine the quality level referred to in the development process.

Refer to each section of Chapter 6 (and the table in Section 13.1) to clarify check items required at each level.

Refer to each section of Chapter 7 as needed to grasp specific concepts of the above check items and applicable technology.

The definitions of the development stage, etc. referred to in each section are summarized in Chapter 4.

2. Overview

2.1. Scope of the Guideline

2.1.1. Products and systems subject to the Guideline

Products and services subject to quality management in the Guideline are those that use machine learning technology for the building of some of their software components among all information-processing systems such as industrial products, consumer equipment and information services (hereinafter referred to as *machine learning based system* in the Guideline).

The Guideline mainly considers about machine learning components implemented in *supervised learning* although the basic approach can be also applied to other methods of implementation such as unsupervised learning, semi-supervised learning and reinforcement learning. Future revisions of the Guideline will consider specific methods of handling for them.

2.1.2. Products and services subject to quality management

The quality characteristics which the Guideline set goals, manage, guarantee is the external quality corresponding to the effect on the system which implements the software component built with the machine learning technology (hereinafter referred to as *machine learning component*) that is the internal component constructing the machine learning based system.

On the other hand, quality management mentioned in the Guideline directly target internal qualities machine learning components have.

We consider that qualities in use of the whole system are realized comprehensively by external qualities of elementary parts of that system. Moreover, external qualities of each component depend on the quality of internal components or parts and internal quality characteristics which are the internal quality of those components themselves. Quality management is realized by sorting out the requirements for machine learning components as internal quality characteristics and by managing the process toward the achievement thereof (Figure 9). However, among system components, software and hardware other than machine learning components are explained only within the necessary scope of their relationship to the Guideline for the purpose thereof.

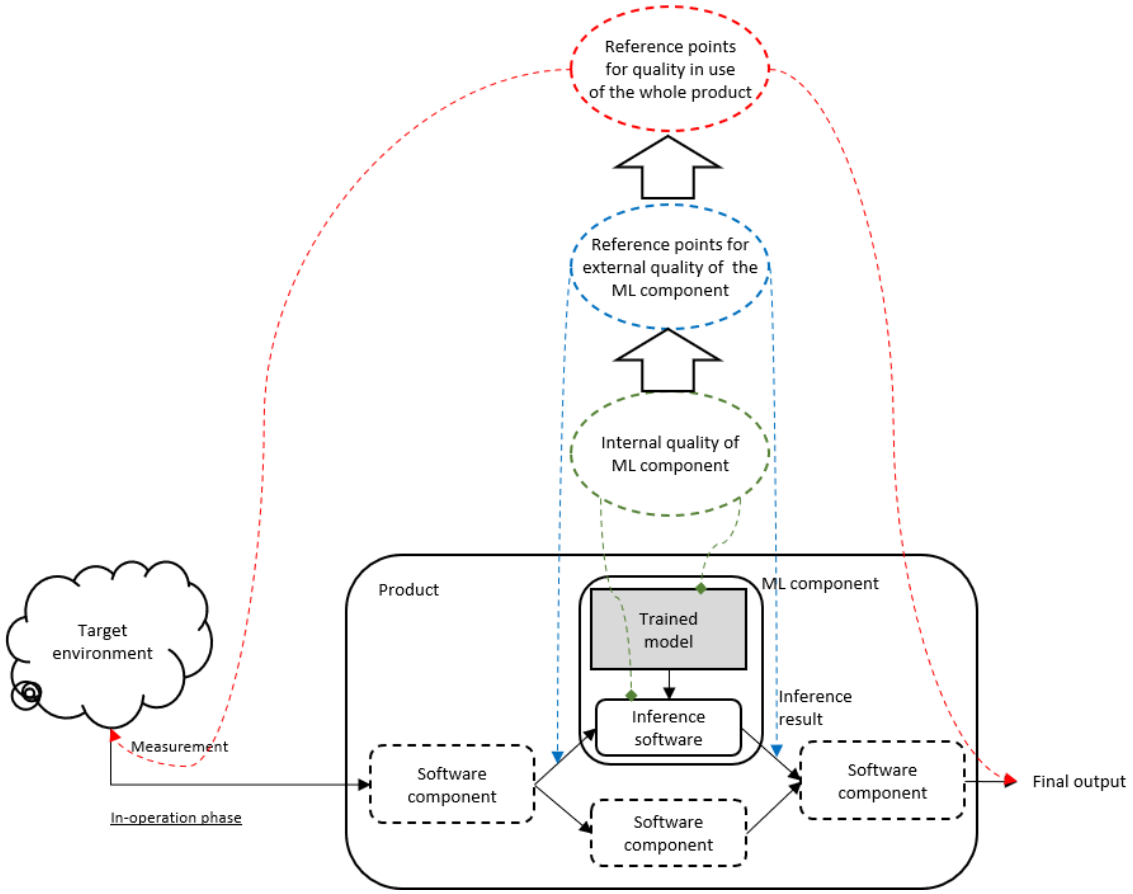

Figure 9: Concept of quality requirement of whole AI-based system and means for realizing them

2.1.3. Scope of quality management

In the Guideline, the term *quality management* refers to a process of quality activities across the lifecycle from the planning stage to the operational stage of machine learning based systems and includes such meanings as setting of quality goals, planning, confirmation, quality assurance and management. Although the meaning of *quality management* is wider than quality assurance or management of general software, it does not include organization planning and management of responsibilities or resources which are included in the meaning of quality management defined in ISO9001.

2.2. Relationship between the Guideline and the other standards for system quality

Mainly, this part sorts out the relationship between the Guideline and the existing international standards for information system quality. Also see Section 1.9 for their relationship with guidelines equivalent to superordinate concept such as sociality.

The Guideline present quality management methods for a variety of applications of machine

learning and define systems that require safety as complement or extension of some of conventional standards for functional safety (IEC 61508-3 [13], IEC61508-4 [14], etc.).

- For systems that strongly require functional safety, functional safety standard IEC61508-1 [12] or other equivalent standards are primarily applied.
- For ensuring the required safety level of the machine learning component in the targeted system, the Guideline sorts out the issues and methods for ensuring safety that is unique to the machine learning and propose the methodology for complementing or replacing the methods introduced in IEC 61508-3 [13] comparing with the conventional software.

2.2.1. Security standard ISO/IEC 15408

The Guideline and information system security standards such as ISO/IEC 15408 [3] are independent, and they should be applied simultaneously when needed. Information security is a mandatory requirement for systems that require integrity and availability to realize safety included in the Guideline, but the basic countermeasures defined in those standards are also applicable to machine leaning based systems.

2.2.2. Software quality model ISO/IEC 25000 series

The ISO/IEC series which define software quality models, especially ISO/IEC 25010 [7], can be partially applicable to machine leaning based systems.

Since product quality concerning software components are analyzed from the viewpoint of conventional software, quality properties sorted out in the above standards are supposed to be achieved also in machine learning components. However, since there are differences from product quality and analysis and decomposition into components focused in the Guideline, we assume that there is no clear relationship between them. Although there is a proposal to improve the above standards to machine learning and AI at an international level, we continue to discuss this matter in view of future standardization.

2.3. Definitions of terms

This section shows definitions of terms used in the Guideline. Some of the terms on artificial intelligence and machine learning are subjects of discussions in ISO/IEC 22989 [4]. Definitions of those terms in the Guideline are to be aligned after the discussions conclude.

2.3.1. Terms related to machine leaning based system structure

2.3.1.1. Machine leaning based systems/systems using machine learning technology

A system containing software components (machine learning components, 2.3.1.2) implemented by applying machine learning technology.
(Note) The common terminology “machine learning systems” refers to machine leaning based systems or machine learning components (2.3.1.2) depending on the context.

2.3.1.2. Machine learning component

A software component implemented by applying machine learning technology. A machine learning component realize the functions of trained machine learning models (2.3.1.5) as software. Generally, this component consists of prediction/inference software component (2.3.1.6) implemented as software and trained machine learning model (2.3.1.5) incorporated as fixed input.

2.3.1.3. Machine learning algorithms

An algorithm that provides the calculation method for prediction/inference of machine learning and for obtaining said calculation method through training. Each of the calculation methods correspond to prediction/inference software component (2.3.1.6) and trained machine learning model (2.3.1.7).

There are a lot of types of machine learning algorithms, such as neural networks, support vector machines and decision trees and so on. Appropriate algorithms are selected based on the type and purpose of knowledge which machine learning components are to obtain.

2.3.1.4. Hyperparameter

A setting value input in training software component (2.3.1.7) to execute machine learning training. It may need to be adjusted in accordance with the progress of training. In some cases, a hyperparameter may not exist, adjustment of hyperparameter may be omitted intentionally, or a technique to adjust it automatically may be used.

(Note) As regards the diversity of mathematical structure that can be subject to machine learning, which scope is fixed as *machine learning algorithms* prior to the start of training and which scope is set/adjusted during training as *hyperparameters* are arbitral depending on how to implement software and selection of policies for training by developers. In some cases, the structure of calculation formula itself is treated as data and subject to automatic adjustment for optimization as a part of hyperparameters.

2.3.1.5. Trained model/trained machine learning model/knowledge base/trained parameter

Information required as output for training that defines functional operations of

machine learning.

(Note 1) *Machine learning model* in the Guideline refers to trained machine learning model or in-processing data for obtaining said model.

(Note 2) What is simply called *parameter* in the context of machine learning technology often refers to this trained machine learning model or numerical data included therein.

(Note 3) Mathematical structures such as neural networks and Bayesian networks are sometimes called as *graphical model* or *statistical model*. Accordingly, the terms such as *selection of machine learning model*, *model design* and *untrained model* may be used for the selection of machine learning algorithms and the setting of their parameters.

(Note 4) In literatures, the term trained model is used in comparison to trained parameter either as *output of model training* or as trained machine learning model implemented as specific software. In the Guideline, the term trained model is used for the former meaning while the latter is called *machine learning component*, to clearly distinguish output from training itself and pre-implemented prediction/inference software component.

(Note 5) When the context on machine learning or AI is clear, there may be no problem in calling it simply as trained model.

2.3.1.6. Prediction/inference software component

A part of machine learning components used during operation which is fixed as the implementation of software of machine learning models. This component receives a trained machine learning model (2.3.1.5) as static or semi-static input and data obtained from real environments as dynamic input.

2.3.1.7. Training software component

A component to generate trained machine learning models (2.3.1.5) using training datasets. This component implicitly corresponds to prediction/inference software component (2.3.1.6) as the implementation based on different aspects of the same machine learning algorithms so that they are usually used as a pair.

Although a training software component is not included in machine learning components used during operation in many cases, there are the cases (e.g. a system that runs the online learning during operation and reinforcement learning) where the training software component is included as a part.

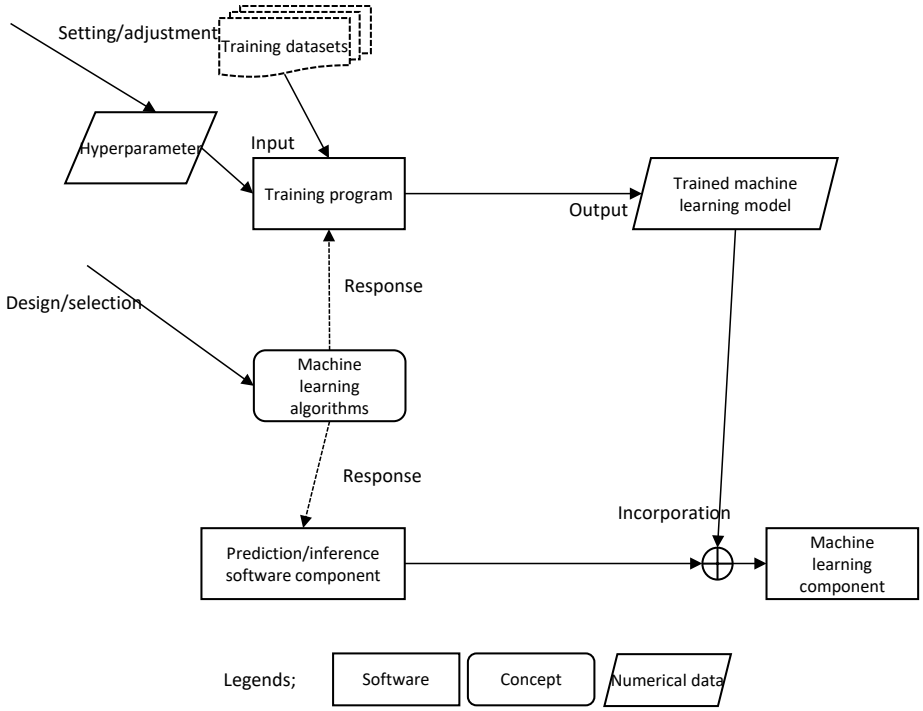

Figure 10: Relationship between terms concerning the structure of machine learning based systems

2.3.2. Terms related to stakeholders of development and their roles

2.3.2.1. Service provider

An entity that uses machine learning based systems for its own purpose or for selling or providing services for customers. Moreover, as regards a business structure in which software providers, etc. plan and develop in advance systems in anticipation of services provided or used by others and sell them as packaged or customized systems, those entities that engage in planning and development are treated according to the service providers.

2.3.2.2. Self-development entity

A product development entity designs and implements the machine learning components.

2.3.2.3. Development entruster

A product development entity that asks others to implement machine learning components regardless of the form of contract (service/entrustment).

2.3.2.4. Development entrustee

An entity that implements machine learning components upon request from the development entruster.

2.3.2.5. (Development) stakeholders

All stakeholders that engage in development and operation of machine learning based systems, including production operators, self-development entities, development entrusters and development entrustees.

2.3.3. Terms related to quality

2.3.3.1. Harm avoidance (Safety)

A quality characteristic of avoiding negative effects such as human damage, economic loss and opportunity loss on operators, users of products or third parties caused by undesirable judgements of machine learning components. The improvement of *safety* corresponds to a concept of *risk reduction* in the safety field.
[Defined in Sections 1.5.1 and 3.1 in the Guideline]

2.3.3.2. AI performance

This characteristic allows machine learning components to output expected by machine learning based systems and their users at a higher precision or probability than the average in the long run. AI performance is evaluated based on comprehensive performance rather than whether the output is acceptable or not. (Defined in Sections 1.5.2 and 3.2 in the Guideline.)

2.3.3.3. Fairness

The output or distribution of machine learning components is not affected by differences in some of the attributes belong to the people or others or the effect is kept sufficiently low.
(See Sections 1.5.3 and 3.3 of the Guideline)

2.3.3.4. Attack resistance

This characteristic prevents machine learning components or machine learning based systems from reacting contrary to the operator's expectations in line with an attacker's intention in relation to input data or external environments built intentionally by the attacker.

2.3.3.5. Ethicalness

The behavior of machine learning based systems is appropriate in the human-centered society.

2.3.3.6. Robustness

This characteristic allows systems to maintain their performance level under any circumstance.

2.3.4. Terms related to development process

2.3.4.1. System lifecycle process

A process management model looking down the flow from the planning of systems to the end of operation and disposal.

[Source of definition: ISO/IEC/IEEE 15288[2]]

2.3.4.2. Agile development process

A collective term of agile and value-driven development approaches.

(Note) This term originates from the Agile Manifest announced in 2001 and can be utilized in various forms such as scrum and XP.

2.3.4.3. PoC (Proof of Concept)

Preliminary development activities carried out with the aim of validating feasibility of ideas to be realized and solution for the problem instead of being realized as products.

2.3.4.4. Systems engineering

An approach across several fields to deploy balanced system solutions in response to diversified needs of stakeholders. It applies both management process and technological process and strikes a balance between them to reduce risks that affect the success in projects.

2.3.4.5. RAMS (Specification and demonstration of Reliability, Availability, Maintainability and Safety)

In the Guideline, RAMS refers to a concept of the comprehensive system lifecycle process specified in the standard IEC 62278 [15] (EN 50126) about the reliability management process in the field of railway service.

(Note) *The RAMS standards* may refer generally to IEC 62278 itself.

2.3.5. Terms related to use environment

2.3.5.1. Environment

In the Guideline, this term is used in three contexts: 1) External environment, 2) computing environments, and 3) operational environment.

2.3.5.2. External environment

The physical environment or cyberspace to be the source of input to the machine learning based system, and there is interference from the environment to the systems or from the systems to the environment.

2.3.5.3. Open environment

The external environment whose system operation is greatly affected by the condition of the people other than users or the things such as nature.

2.3.5.4. Environmental condition

A characteristic which distinguishes the differences in conditions of external environment. From the viewpoint of machine learning quality management, we focus especially on changing characteristics such as the status of hazard and risks.

2.3.5.5. Computing environment

Software execution environment where machine learning components, prediction/inference software components and training software components are executed. It may include hardware environment, operating systems and middleware on which computing environment is based depending on situations.

2.3.5.6. Operational environment

Operational environment includes computing environment and operational structure of humans.

2.3.5.7. In-operation environment

Production environment where machine learning based systems are provided.

2.3.5.8. Runtime (computing) environment

Computing environment of computers, clouds and IoT devices where machine learning based systems including machine learning components are operated in in-operation environment.

2.3.5.9. Development environment

Computing environment where machine learning components are built and modified and such modifications do not affect actual operation directly, and operational environment including computing environment thereof.

2.3.6. Terms related to data used for building machine learning

2.3.6.1. Training dataset

A set of data used for training of machine learning models in the iterative training phase.

2.3.6.2. Training data (instance)

Data included in training datasets.
(Note) Generally, the term *data* is used for both single instance and a set of instances. In the Guideline, the term *data* is used only for the former, while the term *dataset* is used for the latter, to avoid this ambiguity. That is, the term *dataset* is used for the case that it is treated as one set which can be the target of discussion about the distribution and the total quantity of a set. On the other hand, the term *data* is used to refer to individual data points or discrete data points before it is captured as a block. The relationship between these terms is described by expressions of the relationship between set and element such as *added to dataset* and *included in dataset*.

2.3.6.3. Validation dataset

A set of data used for evaluating convergence of machine learning models in the iterative training phase.

2.3.6.4. Test dataset

An set of data used by machine learning models built as input in tests as one or more means for checking the desired quality and performance in the quality check/assurance phase.

2.3.6.5. Adversarial examples

Data built with intention so that the inference results to be different from assumptions or intuition are output when it is input into the machine learning components.

2.3.6.6. Overfitting

Trained machine learning models are adapted excessively to training data and it becomes impossible for them to return desired output in response to input data other than training data.

2.3.6.7. Attribute

Each itemized element to analyze and classify the characteristics of environmental conditions in ML problem domain analysis.

2.3.6.8. Value

Types of specific characteristics of environment included in each attribute classified in the ML problem domain analysis.

2.3.6.9. Labels

Identifiers belonging to data used for classification problems in supervised learning, indicating correct classes which they belong to.

2.3.7. Other terms

2.3.7.1. Software regression

In software engineering, this term means that, when software is improved, it stops to operate as expected in response to input which did not cause problem before.

(Note) In machine learning and statistical analysis, the term regression is often used as regression analysis.

2.3.7.2. SIL (Safety Integrity Level)

The classification of levels related to the achievement of functional safety of products as specified in IEC 61508 [12][13][14].

2.3.7.3. KPI (Key Performance Indicator)

An indicator which quantifies the attainment level of functional requirements to be attained by output from machine learning components through machine learning based systems.

2.3.7.4. Continuous learning

A form of operation to collect data and carry out additional training for machine learning during operation and to update trained machine learning models when necessary.

(Note) This is a concept including not only *online learning* described below but also *additional offline learning*.

2.3.7.5. Online learning

A form of system implementation in which additional training is carried out without going through the validation process in development environment and its results are reflected in operation environment.

(Note) In the Guideline, a form of carrying out additional learning in development environment and updating model parameters by such means as firmware updates after going through the assurance process without online learning is called *additional offline learning*.

3. Levels for external quality characteristics

As described in Section 1.5 of the Guideline, the Guideline defines five aspects of external qualities for machine learning components contained in machine learning based systems. In this section, we set target levels of quality for each aspect. The procedures for determining target levels in development will be provided in Section 5.1.2.

3.1. Safety

For the safety aspects, quality levels are classified into seven *AI safety levels* (AISL 4, AISL 3, AISL 2, AISL 1, AISL 0.2, AISL 0.1, AISL 0). Throughout the Guideline, causes of the safety risks include both physical hazards such as injury of humans and economic hazards as well.

The required AISL shall be determined using Table 1 and Table 2, depending on the nature of corresponding hazards[5]. If the cell containing the" *****" in Table 1 is referred, the functional safety standard IEC 61508-1 [12] or any similar standard for specified application fields is consulted first, safety function is assigned for the components in the system, and the SIL 4–1 assigned to the machine learning component in question or any equivalent assignments in other related standards is translated to AISL 4–1 respectively. Optionally, even in the other cases, stakeholders can apply risk analysis based on IEC 61508-1 or others and assign AISL based on the SIL evaluation result.

In the meantime, the Guideline does not expect AISL 4 or 3 is directly assigned to any machine learning component. In that case, the overall system design should be reconsidered to reduce risks caused by the machine learning components in question.

Table 1: Estimation of AI safety levels for human-related risks

Expected severity of hazards/possibility of avoidance	Unavoidable	Avoidable via human oversight	Human check or manual operation always needed
Simultaneous deaths of multiple people	*	*	*
Death or injury of one person	*	*	*
Injury causing permanent disability	*	AISL 2	AISL 1
Serious injury	*	AISL 1	AISL 0.2

[5] Each AISL is described to correspond roughly to safety integrity levels 4~1 of the standard IEC 61508 of functional safety. Furthermore, to precisely recognize safety requirement strength to be categorized as *No SIL applicable*, the corresponding AISLs are divided into three classes as 0.2, 0.1 and 0.

Minor injury	AISL 1	AISL 0.2	AISL 0.1
Minor injury (where possible victims can avoid the hazard easily)	AISL 0.2	AISL 0.2	AISL 0.1
No damage is expected	AISL 0	AISL 0	AISL 0

Table 2: Estimation of AI safety levels for economic risks

Expected severity of hazards/possibility of avoidance	Unavoidable	Avoidable via human oversight	Human check or manual operation always needed
Damages affecting continuity of business entities	(AISL 4)	(AISL 3)	AISL 2
Serious damage that undermines business operations	(AISL 3)	AISL 2	AISL 1
Considerable/specific damage	AISL 2	AISL 1	AISL 0.2
Minor loss of profit	AISL 1	AISL 0.2	AISL 0.1
No damage is expected	AISL 0.1	AISL 0	AISL 0

(Note) AISL 0.1, 0.2, 1, 2, 3, 4 may be considered to correspond to Sensor Performance Classes A to F in IEC 62998 [18] respectively. AISL 0.1 to 3 may also be considered to correspond to Performance Levels PL_a to PL_e in ISO 13849 [1] respectively.

3.2. AI performance

For the AI performance aspects, quality levels of components are assigned by the agreement between stakeholders based on the following criteria.

- AIPL 2 (mandatory requirements):
  - When it is indispensable or strongly expected that the product or service satisfies certain performance indicators (e.g. accuracy, precision or recall) for the system's operation.
  - When the fulfillment of the certain performance indicators is clear stated as requirement in contracts.
- AIPL 1 (best-effort requirements):
  - When certain performance indicators are identified as an factor for satisfying purpose of the product, but AIPL 2 is not applicable. For example, when shorter

development period is beneficial, or when gradual improvement of performance to a satisfactory level during in-operation phase is acceptable.

- AIPL 0
  - When performance indicators are not specified at the beginning of development, and discovery of a performance indicator itself is the purpose of development.
  - When the development completes at the PoC stage.

3.3. Fairness

Quality levels of components for fairness aspects are determined based on the following criteria.

- AIFL 2 (mandatory requirements)
  - When a certain level of fair treatment is required according to the laws, regulation or equivalent de-facto social guidelines.
  - When the product deals with personal data and its output directly affects right of individuals.
- AIFL 1 (best-effort requirements)
  - When it is possible to define the requirements that there is no bias into the product or the services.
  - When there is possible demand for the explanation of the fair treatment provided by machine learning component (or AI in general) that might affect social acceptability or operation of the product.
- AIFL 0
  - When there are no identifiable requirements for fairness of the product or service.
  - When the product or service would be affirmatively accepted even if their outputs are unfair or non-uniform, in consideration with other factors such as performance.
  - When requirement for fairness is understood as *high performance regardless of input class*. In this case, it will be interpreted as a requirement for diversity support of input classes, related to other properties (safety or performance).
    (For example, in an application to detect some mechanical defects, if a detection rate for some specific defect class is relatively high, there are no need to *decrease* the detection rate to match with those for other unfavored classes).

3.4. Privacy

When the data subject is a natural person linked to personal data, protection of the personal data shall ensure that right to privacy of the data subject is not threatened. When dealing with the right of the data subject, the provisions for personal data protection processing referred to by national and regional laws and regulations shall be followed, and at the same time organizational measures shall be applied. In addition, further technical measures should be adopted to achieve the level of personal data protection expected from an ethical point of view and to reduce the threat of personal data leakage from the artefacts of machine learning component development (Section 9.3).

- AIPrL 2
  - When the product or service represents an explicit and potential threat to the protection of data subjects and requires personal data protection processing and organizational measures and application of technical measures to reduce threat of re-identification as well, aimed at compliance with applicable laws and regulations.

- AIPrL 1
  - When the product or service represents an explicit and potential threat to the protection of data subjects and requires application of personal data protection processing and organizational measures aimed at compliance with applicable laws and regulations.

- AIPrL 0
  - When there is no data subject protection requirement for the product or service in question.
  - When the product or service in question represents an explicit and potential threat to the protection of data subjects and requires application of organizational measures aimed at compliance with the applicable laws and regulations that stipulate the protection.

3.5. AI security

For AI security, no level specific to this external quality is set at this time. The basic approach is to conduct risk assessments for usage and development situations according to the strength of the requirements of the above four quality characteristics, and to implement countermeasures when necessary.

4. Reference models for development processes

This chapter explains the development process reference models for machine learning based systems which the Guideline refer to on the premise of discussions.

The reference models mentioned in this Chapter do not force developers to adopt a certain development method of machine learning based systems. Rather, they aim to add referable names to internal components and development processes of general machine learning based systems and compare recommendations provided for in the Guideline with actual development processes. By this way, the respective unique configurations and development processes can be compared with this model.

As shown in Figure 6 (page 26), the Guideline roughly divides the development process of machine learning components (and the principal parts of machine learning based systems that contain them) into three phases.

4.1. PoC trial phase

In the earliest stage of system development, *the PoC trial phase* as a development lifecycle process is sorted out as the preparatory stage to identify a possibility of performance attained by a system, quality degradation risks and its possible cause in advance of defining the functions of the whole process. Quality management activities in this phase are *sorted out as important preparatory works for quality management activities in the main development phase* in the context of quality management defined in the Guideline. From the viewpoint of data science, analytical activities in this phase are closely associated with subsequent risk analysis and requirements analysis and their results are often used in the subsequent main development phase. In order to attain *sufficiency of requirements analysis* (Section 6.1), it is essential to make efforts in the PoC trial phase. The Guideline stipulate that the adequacy of those activities is checked again in each step of the main development phase ultimately. This allows quality management activities to be conducted on a trial-and-error basis in the PoC phase without restrictions.

4.1.1. Handling of PoC phase including trial operation

Moreover, *PoC* does not only mean simple data collection and analysis, trial training, and build-out of models but includes what is tested in an in-operation environment, although final products will be used in different situations for example, under monitoring by humans. Furthermore, the operation stage may be divided so that quality requirements continuously shift to different operational conditions. The Guideline considers such type of development with physical operation divided in several stages as equivalent to the whole development lifecycle including *the main development phase* in relation to the system building stage in each stage and

treat all achievements in the early operational stage as knowledge obtained from the PoC stage in later stages of development (Figure 11).

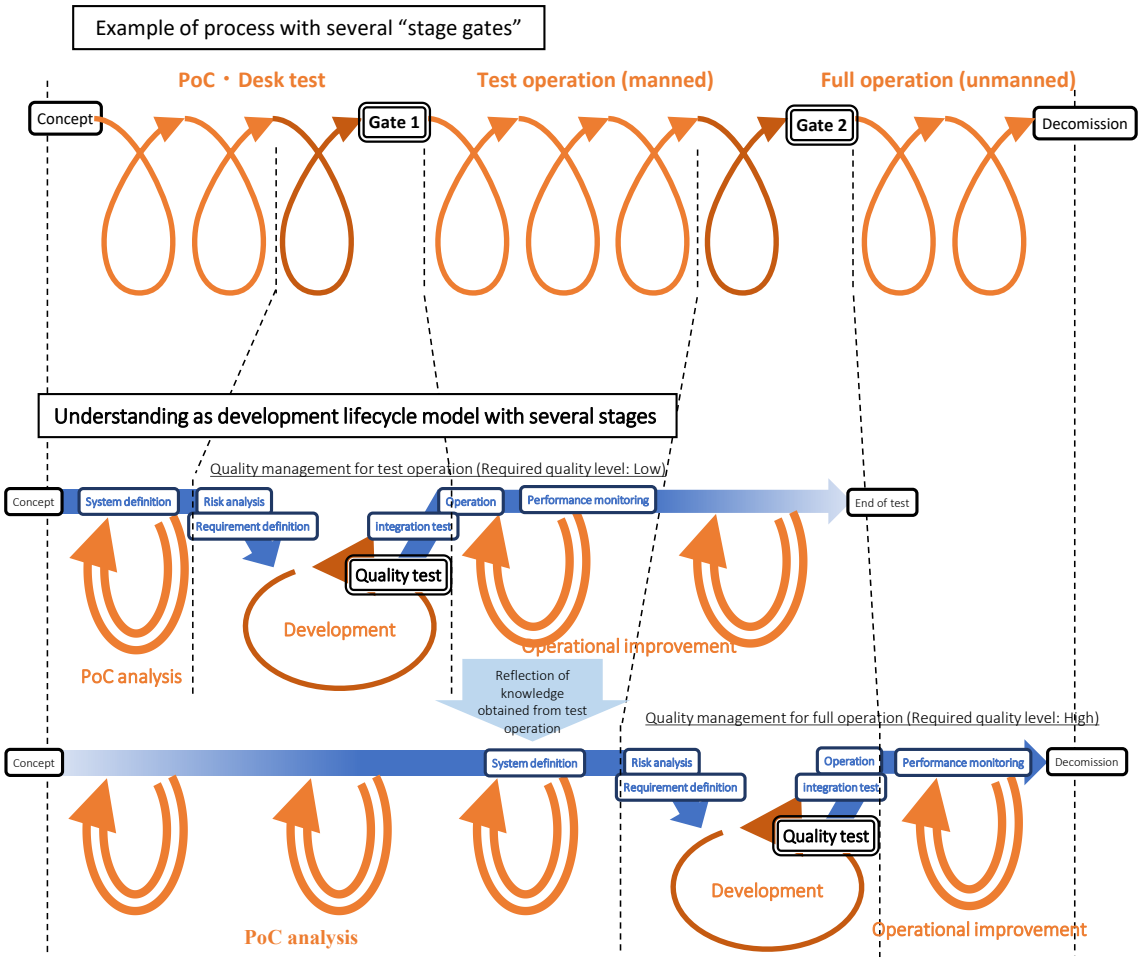

Figure 11: Handling of development process with several operational stages (example)

4.2. Main development phase

The main development phase is a process from when system's functional requirements are confirmed by means of the efforts in the PoC trial phase to immediately prior to the in-operation phase. This phase includes system definition equivalent to very early stages of development of conventional systems and integrated tests prior to final operation and shipment.

In the main development phase, a lifecycle which mixes a part of conventional waterfall early-stages of development and iterative development processes specific to machine learning components is defined as a model. To be more specific, system requirements for each component including functional definitions and risks analysis of the overall system, analysis and decomposition of effects of components and machine learning components are decided. A series

of processes which sets the goals for the last integrated tests prior to shipment are sorted out pursuant to the conventional waterfall model. Moreover, components of software and hardware other than machine learning components are assumed to adopt the conventional waterfall development model or agile development process[6] capable of ensuring the same level of quality as needed in late-stages of development. On the other hand, a development process of specific machine learning model equivalent to *late-stages of development* of machine learning components defines the iterative development process model again in the following section.

4.2.1. Machine learning model building phase

After risks which machine learning components have to deal with as well as quality requirements are specified in the main development phase, a machine learning model is built and its external qualities are tested using indicators for internal qualities and this process is repeated until the model passes a test. This is called *machine learning model building phase*.

A model which breaks down this phase into more specific process is shown in Figure 12. Each process of this model aims to correspond it to the descriptions in the Guideline by comparing a development process specific to each developer with this model so that it is not to restrain the development process unambiguously.

[6] *Agile development process* here means test-driven development or another non-waterfall style of development process which includes quality assurance activities comparable to those of waterfall processes so that expected levels of quality is reasonably explainable. It does not include all forms of non-waterfall development.

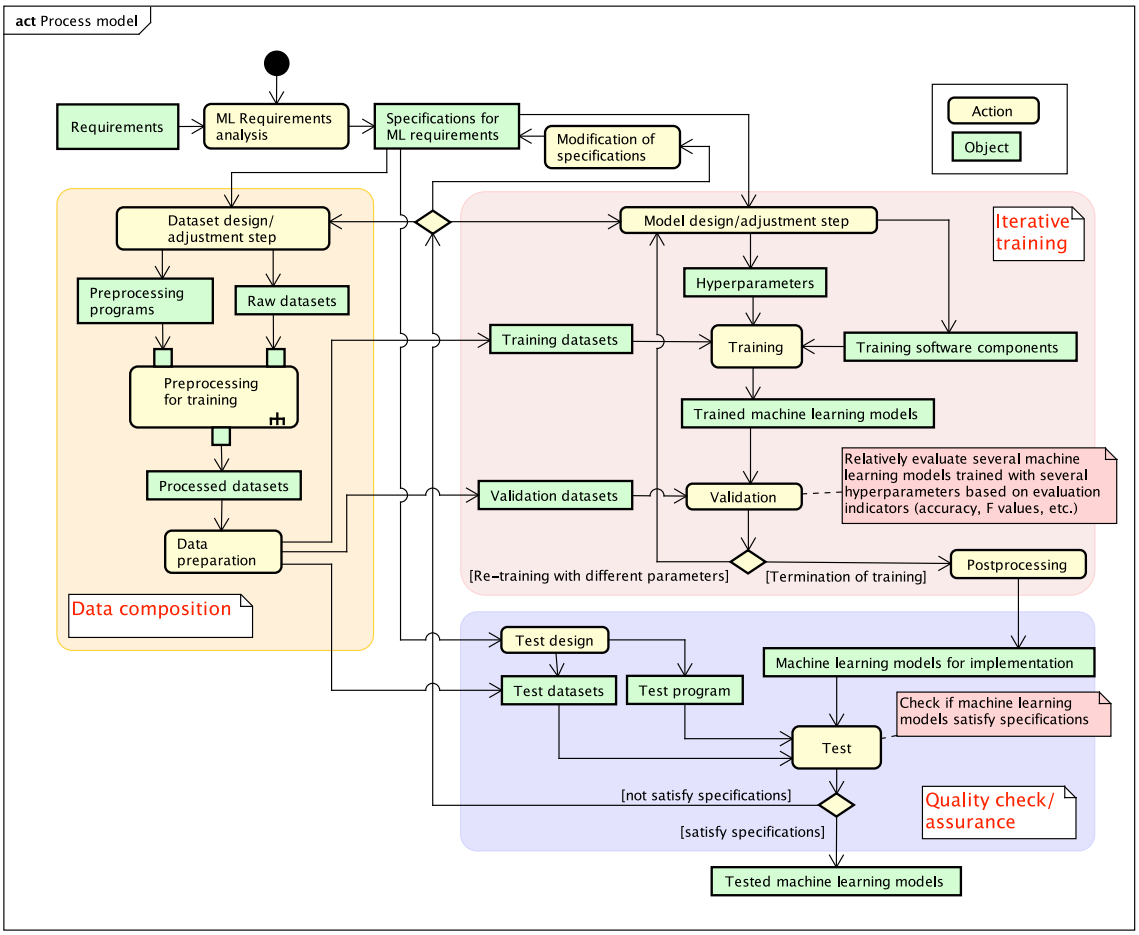

Figure 12: Process model in the stage of machine learning building

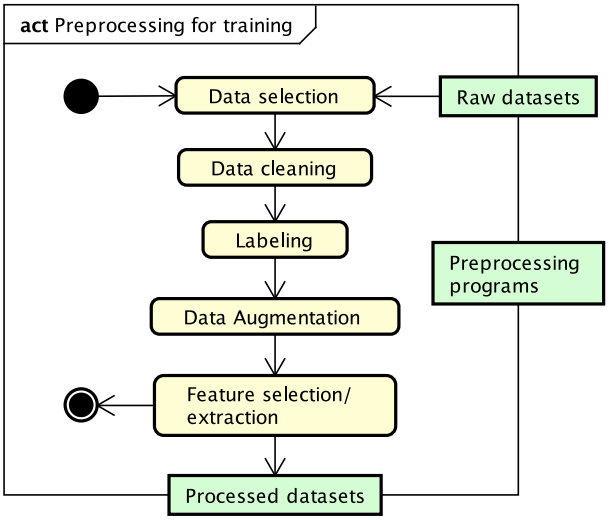

Figure 13: Example model of preprocessing for training

4.2.1.1. ML requirements analysis phase

In the *machine learning requirements analysis phase*, conceptual quality requirements for machine learning components are concretized by sorting out characteristics of data input to or output from machine learning components and re-organized as specific requirements for build data in the ML model building phase. In the actual development of machine learning based systems, performance requirements and quality requirements are specified in many cases based on data characteristics or specific datasets. By explicitly including this analysis phase, it becomes possible to sort out a specific method of managing data quality and the relationship with other components in terms of quality management. This phase is closely related to internal qualities mentioned in Sections 6.1 and 6.2.

4.2.1.2. Training data composition phase

In the training data composition phase, training, validation, or test datasets are generated in accordance with specifications for ML requirements. This phase is explained below in line with *composition of training data* shown in the upper left side of Figure 12.

4.2.1.2.1. Dataset design and adjustment step

In the dataset design and adjustment step, sufficient data are collected for each case in accordance with the specifications for ML requirements and design is carried out to generate datasets suitable for training. Output of this step is unprocessed datasets and preprocessing programs. In cases where it is found that trained machine learning models do not satisfy the specifications after going through training, the process may return to this step to adjust the design of datasets.

4.2.1.2.2. Preprocessing step for training

In the preprocessing step for training, raw datasets generated in the dataset design/adjustment step in Section 4.2.1.2.1 are processed into datasets suitable for training. As Figure 13 shows, there are cases where each process is performed in random order, in parallel, or repeatedly.

4.2.1.2.2.1. Data selection

Datasets used for training, validation and tests are selected considering coverage and uniformity from a large quantity of raw datasets. Since coverage and uniformity are traded off, the balance between them should follow the specifications for ML requirements.

4.2.1.2.2.2. Data cleaning

Noise removal, data forming, complement of missing values and removal of outliers are carried out so that original characteristics of the datasets can be trained.

4.2.1.2.2.3. Data Augmentation

In order to ensure the amount of training data or test data and to prevent overfitting, data variants are generated and added to datasets. For example, controls such as rotation, scale-down, reverse, change in luminosity and background replacement of image data are carried out to generate variants.

4.2.1.2.2.4. Label addition

Labels for supervised learning that show correct answers suitable to achieve requirements are added to the dataset.

4.2.1.2.2.5. Feature extraction and selection

Feature extraction is performed to add new features useful for model training, and feature selection is performed to reduce unnecessary features. For example, feature extraction uses techniques such as calculating frequency components by Fourier transform and finding the principal components of a plurality of features. If useful features are obtained, in many cases more unnecessary features can be reduced by feature selection.

4.2.1.2.3. Data preparation step

In the data preparation step, processed datasets generated are divided into training, validation, and test datasets. Training datasets and validation datasets may be replaced during iterative training.

4.2.1.3. Iterative training phase

In the iterative training phase, machine learning models are designed in accordance with the specifications for ML requirements and training and validations are repeated using datasets created in the training data composition phase. This phase is explained below with reference to the flow of *iterative training* shown in the top right of Figure 12.

4.2.1.3.1. Model design and adjustment step

In the model design and adjustment step, hyperparameters necessary for training including the structure of learning models, learning algorithms and various parameters are designed in accordance with the specifications for ML requirements and training software components are created. Hyperparameters may be adjusted in response to the results of validations and tests.

4.2.1.3.2. Training step

Machine learning models are trained using training datasets based on designed hyperparameters.

4.2.1.3.3. Validation step

Validation datasets are input into trained machine learning models to evaluate models based on evaluation indicators such as accuracy, precision, recall, F values, etc. of learning models. Generally, several machine learning models are trained with several hyperparameters and evaluated relatively to judge their adequacy.

4.2.1.3.4. Postprocessing step

In the postprocessing step, models are transformed to adapt trained machine learning models to in-operation environment of inference servers, edge devices, etc. and learning models for implementation are generated.

The following are some specific cases.

- Transformation of models appropriate for in-operation environment by computational performance
  - Changes of calculation precision
  - Compression and speed-up of models such as distillation, quantization, etc.
- Optimization of models appropriate for in-operation environment for efficient inference
  - Compiling of models such as parallelization, vectorization, etc.

The figure of a process model envisioned in the Guideline expects this postprocessing to be placed immediately after the iterative training phase and the quality check and assurance phase to be carried out under conditions close to the time of in-operation. However, postprocessing is often carried out as a part of the process of building the whole system after the quality check and assurance phase of machine learning components in actual development. In this case, the quality checked in the quality check and assurance phase may change or deteriorate at the time of in-operation, and it may become necessary to re-validate numerical equivalence in the test stage of the whole system.

4.2.1.4. Quality check and assurance phase

In the quality check and assurance phase, tests are designed in accordance with the specifications for ML requirements to carry out evaluations tests of learning models for implementation. This phase is explained below with reference to the flow of *quality check and assurance* shown in the bottom right of Figure 12.

4.2.1.4.1. Test design step

In the test design step, programs required for tests are created and test data are added in accordance with the specifications for ML requirements. Test data to be added may include cases where data are generated by means of pre-processing data augmentation and cases where trained models generate data which tend to make erroneous inferences.

4.2.1.4.2. Test step

In this step, the accuracy of learning models for implementation is evaluated based on regular indicators such as accuracy, precision, recall, F values, etc. using test datasets, and its stability is evaluated using other data such as data made by adding noises to ones in the datasets. In cases where any evaluation result does not satisfy the specifications for ML requirements, the process returns to the previous phase to adjust learning models and training datasets and modify the specifications for ML requirements.

4.2.2. System building and integration test phase

This stage is not part of the machine learning building, but it is described here due to the order of the procedures.

In an ideal development situation, all quality requirements for machine learning components are pre-arranged through the PoC trial phase, and all achievements can be confirmed up to the machine learning building phase.

It is hoped that no problems will occur at the subsequent stage of system building and integration testing.

In actual system development, complex real environment requirements analysis may not be perfect, unexpected interactions may occur with other components, and the effects of defects in other components may spread. Quality defects in machine learning components due to these various factors often occur during the so-called integration testing stage. Moreover, especially in a large-scale and complicated system, the prior data are inevitably insufficient as test data, and in many cases, inspection in the actual environment/system after integration is indispensable. In addition, it is conceivable to reproduce the inspection for rare cases that cannot be prepared in advance in the data composition phase at the stage of integration test.

In any of these cases, if a test fails, ML requirements analyses are partially modified and the whole machine learning model building phase is repeated again. In order to avoid this enormous amount of repeated work, it is desirable to accurately record the content of quality management activities conducted in each phase of each stage of machine learning building so that the effect of modifications can be grasped correctly and partial modifications are made without fail.

4.3. Quality monitoring and operation phase

After the system is deployed in actual operation, the activity to maintain the performance by continuously monitoring the quality at the operation stage and making necessary corrections is clearly positioned as the *quality monitoring and operation phase*. Although this phase is placed outside the development phase in a more restricted sense in the waterfall V-shape development model, it already exists in the concept of system lifecycle such as the RAMS standards.

We consider that, in many cases, the quality needs to be managed during operation, from the viewpoint of 1) When qualitative requirements analysis based on prior data does not capture the actual environment and 2) When an environment at the time of operation is different from the environment when prior data were prepared, in machine learning based systems.

In various applications of machine learning, there are various patterns to update trained machine learning models in actual operation, because they cannot be classified uniformly. Therefore, the Guideline classify such patterns into two categories.

1) Pattern of carrying out re-training of machine learning in development environment and deploying it in operation environment after conducting tests.
2) Pattern of carrying out additional learning automatically in operation environment so that the system is self-adapted.

Section 6.9 summarizes the principles for these patterns individually.

5. How to apply the Guideline

5.1. Basic application process

In this section, we overview the process of applying the Guideline.

Although this section focuses on the build-out of machine learning components, it includes processes such as analysis of whole systems which is deemed to have been carried out from the past in the scope necessary for explanation. Therefore, in cases where a specific development process has been built for the adaption to international standards and a person in charge of development can explain its adequacy, modifications such as addition of any necessary stage listed in this section based on that existing process may be made.

5.1.1. Identification of functions in charge in machine learning component system

This paragraph can be skipped in cases where the functions which machine learning components should fulfill within the system such as safety are fully identified in conventional system development processes (e.g., IEC 61508 [12]).

In cases where qualities in use of overall machine learning systems have been unidentified, qualities in use and external qualities of machine learning components are identified through the following process.

5.1.1.1. Prior examination on safety functions/check of applicable standards

- Whether a machine learning based system requires safety functions over a certain level which is roughly higher than SIL 1 or non-negligible personal injury is envisioned is examined in advance.
- In cases where safety functions are required, applicable functions should be selected from IEC 61508[12][13][14] or standards of each application field and follow that process.

5.1.1.2. Identification of system function requirements on purpose and goals

- The purpose of using the system, the scope of envisioned use environment and KPI to be achieved are identified at an overview level.

5.1.1.3. Examination on risk scenarios related to system use

- Examine a possibility that the system's functional requirements are not achieved, or

the system becomes incompliant to the purpose of use or social requests and list damages that may be caused in that case and other disadvantages. Respond to risk assessments.

5.1.1.4. Examination of requirements for characteristics of qualities in use of overall system

- Examine the quality required when the system is in use using quality metrics of any system.
  - As an example of means when there is no appropriate option, the three axes of external quality characteristics listed in Chapter 3 shall be applied to qualities in use to judge which disadvantage examined in the previous section is applicable. Then, the degree of severity thereof is judged based on the standards mentioned in each section of Chapter 3 to correspond them to the quality levels.
  - The maximum level should be identified for each of the three axes, and they shall be the level of qualities in use which the system should achieve.
- However, in cases where existing safety standards are adopted in Section 5.1.1.1, a safety level in relation to physical damages shall be decided based on quality indicators such as functional safety, etc. applicable to the existing standards.

5.1.1.5. Identification of components that contribute to the achievement of system component design and functional requirements and quality characteristics during use

- A combination of system components is designed to differentiate functions achieved by machine components and conventional software from those achieved by machine learning components.
- Moreover, it is analyzed which component of system the achievement of the characteristics of qualities in use depends on.

5.1.2. Identification of required level for achieving external qualities of machine learning components

- A level of achievement of external qualities is identified by means of the following procedures. At this time, machine learning components should specifically contribute to qualities in use.
- As regards risk avoidance, when any conventional standard for functional safety is applied to risks of physical or human damages, an analysis based on the conventional standard has priority, and a functional safety level to be achieved by machine learning

components as software based thereon shall be replaced with a level of achievement required for safety of machine learning components.

- Other cases are defined as follows:
  - Basically, a characteristic level of qualities in use required for the whole system shall be a level of achievement of external qualities required for machine learning components.
  - In cases where undesirable output from machine learning components is monitored and corrected such as modification by overwriting the output by means of software components (Figure 9) processed serially or parallelly with machine learning components and sufficient quality is judged to be ensured for said software components by means of any conventional method, a quality characteristic level required for machine learning components is one level lower than the whole system level.
  - When machine learning components do not contribute to or interfere with the achievement of the quality characteristic level required for the whole system at all, no level of achievement of said quality characteristic, Level 0, should be set for machine learning components.

5.1.3. Identification of level required for internal qualities of machine learning components

- For each item of the internal quality characteristics listed in sections of Chapter 6, a level required for each characteristic is deducted from the level of achievement required for the determined external quality characteristics described in the preceding paragraph in accordance with the relationship set in Chapter 6.

5.1.4. Realization of internal qualities of machine learning components

- A method of realizing the level required for the internal quality characteristics mentioned in the preceding paragraph is examined in accordance with each section of Chapter 7.
- The realization of each characteristic is assured by means of the technologies and processes listed in each paragraph and other technologies deemed to be equivalent.

5.2. (Informative) Entrusting AI developments

The content of this section is informative.

The work explained in the previous section sometimes cannot be covered within one organization, thus requiring an entrusting of development. This section mentions points development entrusters and development entrustees should pay attention to when they work

together, especially in dealing with AI specific aspects. Regarding the contract issues around the ownership of IP and sharing of compensation for damage, please see Section 11.1.1 “Contract guidelines for AI of the Ministry of Economy, Trade and Industry”.

5.2.1. Exploratory approach

When an entrusting of development starts for some machine learning component, it is often difficult for entruster to clearly provide KPI and acceptance criteria to entrustee. Therefore, an exploratory approach is required. The difficulties of this approach is similar to the problems we encounter when applying so-called agile development to the product development process that requires *strict cost management and quality*. Thus, enterprise agile practices are expected to be beneficial.

For example, the PoC phase described in Section 0 can be viewed as *Inception* phase as described in DA (Disciplined Agile [7]). In this phase, the development entruster and the development entrustee should build a consensus on the purpose of the PoC phase of AI such as PoC exit criteria and deliverables to the next phase. The deliverable list (e.g., decision on test strategy and basic architecture) recommended in *Inception* phase in DA can be utilized for this activity. It is highly important to set goals for the PoC stage gate. To prevent any missing out important points, and to appropriately promote subsequent phases, the following two aspects should be covered in the PoC phase. i.e., *requirement clarification* and *feasibility study* (see Figure 14). For both activities, two perspectives should be taken care: *Terminate condition of PoC* and *PoC outcome to be provided to the entruster*.

[7] A summary of best practices for businesses to adopt agile development. It was acquired by PMI (Project Management Institute) in 2019. https://disciplinedagiledelivery.com/

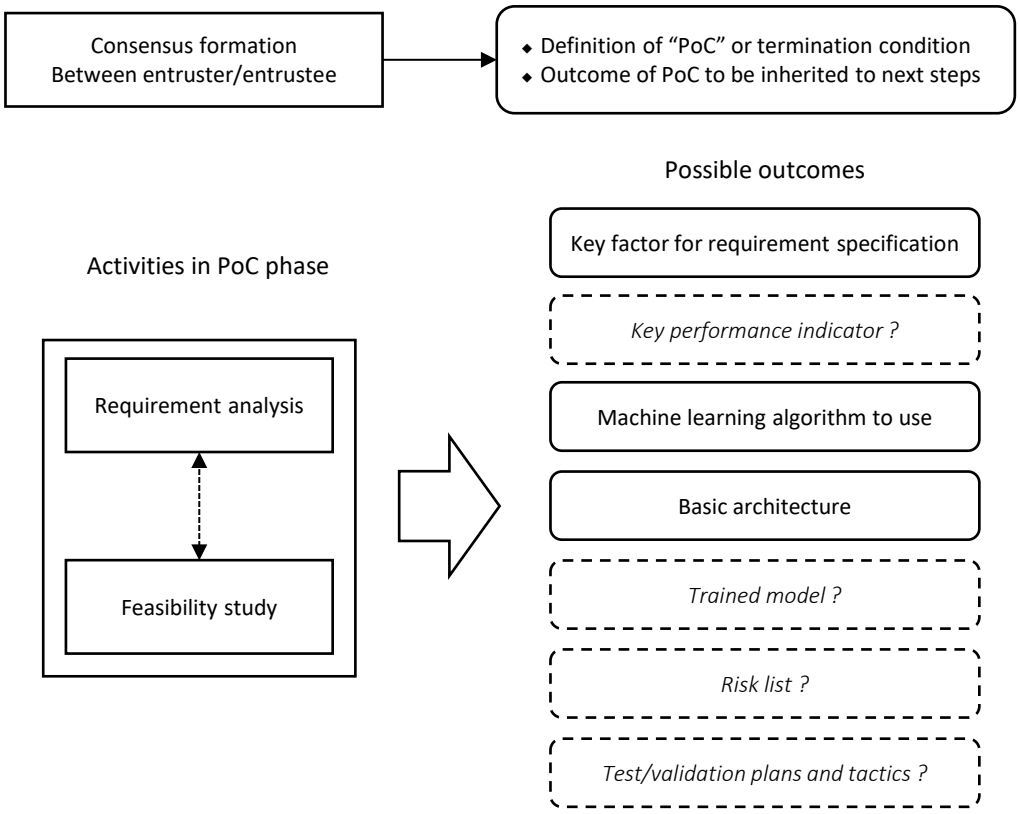

Figure 14: Activities and deliverables in the PoC phase

5.2.2. Role clarification of entrusters and entrustees

The clarification of roles to be performed by development entrusters and development entrustees, for each work in the process described in Section 5.1, will protect both side from risks and prevent overlook of necessary works, leading to a quality assurance of machine learning components.

One example of developing tailor-made AI from scratch is shown below. Terms XXX, YYY, etc. are to be filled in appropriately.

Table 3: Example division of roles

Step	Task for development entruster	Task for development entrustee
1. Safety standards	To examine primarily checking applicability of existing standards.	To confirm the examined result.

2. Functional requirements of the system	To describe/provide in natural language document, based on the requirements of XXX. Qualitative expressions of qualities in use to be included.	To review provided document to clarify the purpose and goals.
3. Risk scenario	To prepare risk list for some similar products YYY Presentation of constraints.	To analyze based on some established method of ZZZ, Update where necessary.
4. Qualities in use/external qualities	To present priorities and constraints between quality aspects (e.g., quality characteristics indispensable for the given product).	To assert quality levels to be targeted. If applicable, use three external quality characteristics.
5. Design of system components	To review and confirm.	To perform system design. For reviews by the entruster, provide materials AAA.
6. Levels for external quality aspects	To review and confirm.	To identify and propose target levels.
7. Internal qualities	To review and confirm.	To examine and propose the levels of target quality aspects and methods for achieving these.
8. Realization of internal qualities	To provide dataset sources. To review test reports.	To perform quality enhance methods examined above and report.

The details in the table would differ depending on the project, and the contents once decided may change, when the work is repeated.

(Example 1)

In some cases, no specific KPI nor quality target can be presented by entruster, and *using particular training dataset* may be the only clear request as a practical measure (the condition defined in the contract). In this case, additional training data are likely to be required as a result of conducting an analysis on internal quality related to data described in Section 6. It is desirable for both sides to recognize such a risk and reach an agreement on necessary arrangements (their roles) in advance, considering Point 7 in the above table.

(Example 2)

When the development entruster *lacks understanding* or *has ambiguity* about the quality to be achieved by the machine learning components in the early stages, such activities as

Presentation of priority and constraints in the above table can be hardly made. In this case, the development entrustee side may present an initial idea as to qualities in use (in a bottom-up manner taking the outline of system design into consideration) in the first PoC phase. In the subsequent main development phase, the roles are to be updated as the entruster vision becomes clearer.

In the case of the PoC phase, the content of each step in Table 3 may change, but entire step is not unnecessary. On the other hand, in the main development phase, the loop within Step 8 is usually sufficient, if the goals for the PoC stage gate have been set and cleared appropriately.

However, depending on the projects, the goals for the PoC stage gate must be dropped or compromised. This means that there may be no clear separation between the PoC phase and the main development phase, leading to uncertainty of the outlook for quality management. In this case, the development entruster shall take the risk, and it is usually left to the development entruster's judgment, whether to accept it and what to do with the work division for mitigating the risk.

5.2.3. Notes on determining the detailed roles

We explain here the results of extracting and examining the items to be noted from the perspective of AI with reference to "Guide to Quality Assurance in Connected World" [212] of IPA/SEC, to concretize works and make arrangements in the development lifecycle as described in the previous section.

1) System and organization that can fulfill accountability

In AI development, quality evidence is often based on the process perspective only. It is so important to clarify the definition including *who can approve it*. Also, we should consider underlying risks in the process.

Example: Regarding the training dataset preparation, it is not enough to describe the preparation of raw data provided by entruster. We should care about who will perform preprocessing (e.g., data cleansing) and who and how will confirm that those preparation has been appropriately achieved.

It is important to consider and specify *compromises and logics both sides can agree on* in advance in the face of time and cost constraints.

2) Test

Even if the training of models is completely up to the entrustee side, the entruster usually must have deep understanding of the content in the following phase, i.e., *quality check and assurance*. For that purpose, it is desirable to not only define evidence but also make arrangements as to the method and scope of tests, and what to give up in terms of effectiveness

and efficiency, so that both sides feel convincing as much as possible when the contract made.

It is important that both sides cooperate for the purpose of *final quality improvement*. For example, *any act just focusing on clearance of verification*, such as using test datasets presented by entruster for training, must be strictly avoided.

3) Quality management during operation

Regardless of immediacy of model update, continuous quality management is indispensable for the machine learning component even after its release, as described in Section 4.3 “Quality monitoring and operation phase”. Therefore, it is necessary to include the design and implementation of optimal quality management methods during operation in accordance with system functional requirements in the scope of works.

For example, both sides may discuss on identification of *data to be obtained* for performance evaluation and environmental data which cannot be grasped completely during development, and the system implementation for obtaining and saving those data during operation.

5.3. (informative) Notes on delta development

Existing software components are often recycled not only in machine learning based systems but also in systems and services which use software components. In machine learning, additional learning is applied to certain data to customize it based on trained models. It is complicated to guarantee the quality of recycled products.

As a fundamental principle, in order to ensure both conventional functional safety and the quality of machine learning covered in the Guideline, the guidelines for assuring quality in situations where a new system is used such as context is consistent from the analysis of conditions in use through the definition and implementation of functional requirements to tests with regard to the quality of systems which use recycled software components. There are, for example, the following methods to realize this consistency, any of which should be chosen in accordance with a quality level required by the system and the state of components provided.

1. A new quality management process is established in the context of new system including detailed checks of datasets in which recycled machine learning components called older datasets are installed, and quality management is carried out independently from older components.
2. Quality management activities equivalent to those described in the Guideline are carried out with respect to recycled machine learning components. When a level of quality management higher than the requirement of the new system is carried out, especially when we can clearly confirm that an envisioned condition in use is a subset of envisioned situations of older components including the case where they are the same with regard to *sufficiency of problem domain analysis*, it is necessary to check if the quality is not deteriorating due to additional learning as needed with reference to records of original quality management.

It is difficult to apply this method if said components are not conscious of quality from the beginning, since quality management of older components should have been carried out and recorded sufficiently. Moreover, because the analysis of conditions in use implicitly assumes certain context of use, it may be difficult to judge comprehensiveness.

Furthermore, a developer that provides machine learning components as general-purpose parts can improve reusability by completing quality management activities in advance in anticipation of this type of recycling and by clarifying the envisioned quality levels.

3. When a relatively low-quality level is required for application, quality management activities with a focus on the test phase should be examined on the assumption that existing datasets are considered as an unknown black box. For example,
   - For Section 6.1 *Sufficiency of requirement analysis* and Section 6.2 *Coverage for distinguished problem cases*, make a fresh analysis for a new system to set a new quality targets.
   - For Section 6.3 *Coverage of datasets* and Section 6.4 *Uniformity of datasets*, prepare new test datasets based on new result of requirement analysis, and check achievements of these characteristics to a certain degree during the test phase.
   - For Section 6.6 *Correctness of trained model* and Section 6.7 *Stability of trained model*, if additional learning is carried out, make an evaluation in the validation/test phase using performance indicators obtained during additional learning, and using training datasets added based on new requirement analysis. When additional learning is not carried out, make a validation based on new result of requirement analysis in the test phase, or to perform evaluation during integration tests on the whole system.

   However, the current descriptions of Chapters 6 and 7 reveal that it is difficult to carry out sufficient quality management activities by using only sampling-type tests of training results without analyzing datasets used for training. It is practical at this stage to obtain detailed information on older datasets and combine it with a white box approach described in the previous paragraph.

6. Requirements for quality assurance

This chapter sets the following nine internal qualities as the quality management characteristics to manage quality of machine learning components mainly for the purpose of the achievement of the two qualities in use, *safety* and *AI performance*. Section 12.1 explains the analyses leading to the setting of these internal qualities. On *fairness*, we will present requirements specific to it in Chapter 8 as well as in this chapter. On *privacy*, we will show requirements specific to it in Chapter 9.

6.1. A-1: Sufficiency of problem domain analysis

6.1.1. General

Sufficiency of problem domain analysis means that usages of a target machine learning based system are analyzed sufficiently and every requirement for the system is captured as described in Section 1.7.1.

Problem domain analysis is important especially in the early stage of development of a conventional software which is used for a safety application. The main purpose of requirements analysis for the development of machine learning based systems in the Guideline are as follows.

1. Sufficient identification of cases in which risk management is needed, mainly in applications requiring safety.
2. Sufficient identification of attributes of objects for which inequality is not allowed, mainly in applications requiring fairness.
3. Sufficient analysis of real world in order to verify that training datasets and test datasets are comprehensive and appropriate extractions of the real world, to all requirements including AI performance.

Sufficiency of requirement analysis is always required not only for machine learning based systems but also equipment and services controlled by software.

However, if any situation where machine learning based systems are used is overlooked at this stage, there are few opportunities for taking note of an error from the data collection stage to the actual training and test stages, so that it is likely to cause malfunction that occurs for the first time in the stage of final system test in real environment or the actual operation stage.

On the other hand, if the details of all possible situations where machine learning based systems are used are analyzed including minute differences, the analysis results can be implemented as a regular software component, and there is no merit in using a machine learning technology.

To put it another way, it is impossible to make comprehensive and thorough analyses in advance for such applied systems. That is why there is demand for having systems acquire

knowledge through machine learning in any way.

From these two viewpoints, it is extremely important to appropriately set *levels of detail of requirements analyses* in machine learning based systems in terms of both quality assurance and feasibility and efficiency of implementation.

They correspond to the judgments of *what do humans need to analyze* and *what should we have machine learning acquire*. The maintenance of appropriate balance between these two viewpoints is an important goal of requirements analysis in quality management of machine learning.

6.1.2. Approaches

The two goals of this internal quality aspects described in this paragraph are

- Clarifying what is required for machine learning components; and
- Clarifying the limited scope which machine learning components have to deal with.

6.1.2.1. Extracting and listing attributes (characteristic viewpoints) and attribute values (specific features)

First, inputs in real world which can be extracted as specific characteristics are organized from the aspects listed below. Each identified viewpoint will be referred to as *attribute*, while the specific features belonging to the viewpoint as *an attribute value*.

The following examples can be given as attributes and their viewpoints.

1. Attributes that characterize the differences in output required for machine learning components as functional requirements.
   - In supervised machine learning, *supervision labels* are different.
   - For example, in the recognition of numeric characters, 10 distinctions *from 0 to 9*. In anomaly detection, whether there is any anomality is detected.
2. Attributes whose output is the same but require machine learning components to explicitly respond at a functional specification level.
   - One example is a case where, in character recognition, the specifications are set that each of *l and ℓ* and *l and /* should *be recognized correctly*.
   - Another example is that, when the specification says that various kinds of obstructions such as *pedestrians*, *bicycles* and *automobiles with crossing movements* are to be avoided in automated driving, each kind of object should be identified as a separate attribute value. Or, *gender* and *age groups* in the scoring for recruitment that requires fairness apply to this paragraph.
3. Attributes for which deterioration risks of performance expected for machine learning

components are likely to be different significantly in real operation

- Some examples include *climate*, *distinction of day and night*, *backlight and front light* and *movement speed* in the case of outdoor object recognition.

4. Elements whose expected output is the same due to the characteristics of machine learning, but it is difficult to find a common ground by the model after learning or they can be recognized as different internally.
   - For example, in the case of outdoor object recognition, different characteristics of objects may be captured in the daytime and nighttime, or another example is the distinction of the horizontal and vertical installation of traffic lights.
   - Another example is that different forms of handwritten numbers such as *1 and /* are categorized into several characteristics which differ greatly depending on countries of origin. In order to recognize all different types, it is necessary to intentionally include them at least in training datasets and test datasets as different general populations, even if they are not clarified in the functional specifications.
5. Characteristics which make it difficult to specify the method of *evenly collecting* learning data as a process when data are collected.
6. Differences which humans can easily capture, though there are no differences as described above, e.g. kinds of animals, skin color, etc.
7. Object which humans cannot recognize or explain their characteristics with word, but it is possible to extract enough samples without bias. For example, it is difficult to analyze all possible ways of describing the number 8 in advance. However, if people do not intentionally write various styles of the number 8 and we have a sufficient number of randomly extracted samples, they are expected to be unbiased samples.
8. Cases where it is extremely easy to mechanically cover and complement differences. For example, once required specifications are established, it is possible to mechanically generate samples of parallel shift of images, color changes and expansion of rotations within a certain range, and intentionally synthesize training data and test data.

We make up a list for candidates of possible attributes from requirements analyzed on the overall system, and then we should determine whether each attribute is picked up for explicit target of management by the developers or is *left to machine learning* without picked up. To determine that, the level of risks related to values of each attribute and the type of the application using machine learning should be considered. Results of PoC trials may be useful as a reference. In general, the characteristics shown earlier in the above list 1 to 3 are highly likely to be extracted as attributes. In contrast, it is often reasonable to *leave the characteristics* shown later 7 to 8 for machine learning without extracting them as attributes. However, the diversity of data types corresponding to attributes *left to* machine learning will be examined in the third internal quality *coverage of datasets*, so that it is necessary to take this point into consideration in the analysis, too.

6.1.2.2. Combinations of attributes that should be excluded

Impossible combinations of attribute values should also be examined. The purpose of examination is to distinguish cases which should be excluded as functional requirements such as *snow in summer in Tokyo* from cases which should be handled as functional requirements such as *snow coverage in winter* which occurs very rarely. It is important to make this distinction, because it will become impossible to determine whether lack of data is acceptable and to explain *the quality of quality management method itself* in subsequent quality management.

To be more specific, after attributes and attribute values corresponding thereto are listed, impossible combinations of attribute values are selected based on system requirements.

6.1.3. Requirements for quality levels

For sufficiency of requirement analysis, it is required to handle the following three levels in accordance with the level of external quality.

The correspondence between external qualities and internal qualities regarding the required level is as follows.

- *Safety*
  - AISL 0.1: Lv 1 or above
  - AISL 0.2: Lv 2 or above
  - AISL 1: Lv 3
  - AISL 2~4: More than Lv 3 (some requirements will be added to Lv 3)
- *AI performance*
  - AIPL 1: Lv 1 or above
  - AIPL 2: Lv 2 or above
- *Fairness*
  - AIFL 1: Lv 1 or above
  - AIFL 2: Lv 2 or above

The requirements for each level of internal qualities are as follows.

- Lv 1
  - Examine and record the major cause of possible deterioration of quality.
  - Based on the examination results, design data and reflect it in necessary attributes.
- Lv 2
  - Analyze risks of deterioration of quality in use in overall system and their impact with a certain level of engineering coverage and record the results in documents.
  - Analyze if any measure is required for each of those risks, and analyze attributes related to the risk which are contained in an input to machine learning components.

- Analyze and record the application-specific characteristics of environments which will generate machine learning input, with regards to the difficulty for machine learning and other aspects.
- Examine sets of attributes and attribute values, based on the results of those analysis and record the background of such decisions.

– Lv 3
- The following activities are carried out in addition to those listed in Lv 2.
- Investigate documents on own past examination results and those of others with regard to elements to be extracted as characteristics of system environment and record the background of examinations leading to the extraction of necessary subsets.
- Investigate past examination results in line with application fields of systems with regard to deterioration risks of qualities in use of overall systems and record the examination results including the background of selection.
- Moreover, extract deterioration risks of qualities in use of overall systems using engineering analysis such as Fault Tree Analysis and record their results.

6.2. A-2: Sufficiency of data design

6.2.1. General

On the basis of the sufficiency of requirements analysis described in the previous paragraph, it requires careful consideration of data design as *coverage for distinguished problem cases* in order to secure sufficient training data and test data with respect to various situations systems need to respond to. More specifically, the number and details of combinations of attribute values focused in the stage from training data preparation to testing process is considered at this stage.

In an ideal situation, for example, if there are sufficient data corresponding to all combinations of attribute values (direct products of attributes) for combinations of attributes that are supposed sufficient as described in the previous paragraph, it can be said that it covers all possible situations in the real world. However, the number of attributes may reach 10 or more in actual system development so that the number of combinations of attribute values may often reach thousands or millions. In this case, it is crucial to consider appropriate *coverage* and *design* for quality management in this paragraph.

It is important to focus on two viewpoints in actual quality management. First, combinations of attributes that may cause malfunction or misjudgment need to be reliability addressed at the training or testing stage. Second, at the same time, it is necessary to cover as much as possible all situations which the machine learning based system to be implemented may encounter during operation from the viewpoint of both the quality of training and the quality of deliverables.

Such a problem has been addressed as a task of *test design* in developing conventional

software. However, in machine learning where not only the testing process but also the implementation process is performed based on datasets, it is unique that the same task is required in the implementation process.

Some practical solutions to the excessive number of combinations for test design have been already presented in conventional software engineering. The Guideline aims to provide practical and sufficient training and quality tests by applying such existing knowledge to machine learning applications.

6.2.2. Approaches

Firstly, if the number of attributes to be recognized is very small and the total number of all attributes where the impossible cases described in the previous paragraph are deducted is only 10 to 20, their combinations are named as *cases*, and they are checked if all cases are included in test datasets and training datasets in later stages.

On the other hand, if there are too many combinations when the total of those attributes are used or sufficient data cannot be acquired especially in rare cases, combinations of some attribute values should be extracted under certain standards and set as *cases* to cover all those combinations. For example, in the example of traffic lights listed in Example 1 in Section 1.7.1, such combinations as *daytime green*, *daytime yellow*, *nighttime red*, *daytime rainfall*, *nighttime rainfall* and *red signal in rain* are extracted and called as *cases* to make sure that data applicable to these combinations is included in test datasets. Even if all attributes cannot be covered completely, it is possible to intend to achieve a certain degree of *coverage* by avoiding cases where high-risk situations such as *nighttime rainfall* are not included at all in datasets or eliminating cases where *red signal* is not included in training at all. When considering such *simplified coverage*, the data of *red signal and daytime rainfall* are included in the several cases such as *daytime rainfall*, *daytime red signal* and *red signal and rainfall* at the same time.

Furthermore, completeness should be guaranteed more specifically in applications that require high quality by introducing mathematical *coverage criteria* to such extraction works. In the case of black box tests in software engineering, methods such as the sampling rate of random sampling, the orthogonal table and pair-wise test and the *coverage criteria* based on them are known, we should select appropriate means for each application.

6.2.3. Requirements for quality levels

The handling of the following three levels is required for coverage of distinguished problem cases in accordance with a level of each external quality.

The correspondence between external qualities level and the required level of this internal qualities are as follows:

- *Safety*
  - AISL 0.1:Lv 1 or above
  - AISL 0.2: Lv 2 or above
  - AISL 1: Lv 3
  - AISL 2~4: Requirements to be added to Lv 3 should be examined.
- *AI performance*
  - AIPL 1: Lv 1 or above
  - AIPL 2: Lv 2 or above
- *Fairness*
  - AIFL 1: Lv 1 or above
  - AIFL 2: Lv 2 or above

The requirements for each required level for the above internal quality are as follows.

- Lv 1
  - Set cases for each of attributes corresponding to major risk factors.
  - Moreover, set cases corresponding to combinations of composite risk factors.
  - Furthermore, extract attributes of differences in particularly-important environmental factors and prepare cases corresponding to combinations with serious risk factors.
- Lv 2
  - Satisfy all requirements listed in Lv 1.
  - Particularly-important risk factors should satisfy, in principle, the standards for pair-wise coverage. To be more specific, a case of combining *an attribute value of combination of those factors* and *individual attribute values included in all attributes other than those to which the attribute value belongs* should be included.
- Lv 3
  - Based on engineering consideration, set standards for coverage of attributes and establish sets of combinations of attribute values that satisfied standards for coverage as cases.
  - The level of strictness of the standards for coverage such as pair-wise coverage, triple-wise coverage, etc. should be set taking into account system usage and risk severity. Standards can be set individually for each risk where necessary.

6.3. B-1: Coverage of datasets

6.3.1. General

The term *coverage of datasets* means that enough data are given to each of the cases covered by establishing the standards as described in the previous paragraph without omissions for the input possibilities corresponding to each of the cases.

When conventional software is developed, the details of all characteristics in real world which software operations depend on are captured at any stage from the machine learning requirements analysis phase to the implementation phase and reflected in software components in the form of conditional branching or calculation formula. On the other hand, when a machine learning component is built, differences in details exceeding a certain level are not captured as attributes of ML requirements analysis or *labels* at the time of training but are reflected in ultimate operations through the training phase of machine learning as a distribution of datasets used for learning, as described in Section 6.1. The purpose of establishing this characteristic axis is to guarantee that no inappropriate learning behavior occurs due to lack of data with regard to characteristics of those unidentified details as *attribute values*.

6.3.2. Approaches

The main purposes of this quality aspects are to achieve sufficient level of *quantity* and *coverage over input situation*. However, since there are characteristics whose minor differences have not been identified as attributes, the latter coverage often has no choice but to depend much on the process of collecting and processing data. On the other hand, when the standards for coverage are introduced in Section 6.2, it would be possible to test if attributes which *are not included in cases* are distributed without bias.

Moreover, as regard the former quantity, it is important to appropriately define granularity of how to set the cases described in the previous section, but it is possible that sufficient data for specific cases cannot be obtained, for example, in rare cases where the frequency of occurrence is low. In those cases, it may become necessary to take measures in the whole development process for evaluating the response status by means of tests on the whole system by partially abandoning *coverage of datasets* for cases that lack specific data and after covering looser coverage standard.

6.3.3. Requirements for quality levels

The correspondence between external qualities level and the required level of this internal qualities are as follows:

- *Safety*
  - AISL 0.1: Lv 1 or above
  - AISL 0.2: Lv 2 or above
  - AISL 1: Lv 3
  - AISL 2~4: Requirements to be added to Lv 3 should be examined.
- *AI performance*
  - AIPL 1: Lv 1 or above
  - AIPL 2: Lv 2 or above

- *Fairness*
  - AIFL 1: Lv 2 or above
  - AIFL 2: Lv 3
    note: We consider that this quality characteristic is important in dealing with data bias, which is an important factor that impairs fairness.

The requirements for each required level for the above internal quality are as follows.

- Lv 1
  - Consider the source and method of acquiring test datasets to ensure that no bias is found in application situations.
  - Extract samples without bias from original data for each case to ensure that no bias is found.
  - Record activities carried out to prevent bias from entering.
  - Check that there are sufficient training data and test data for each analyzed case in the training phase, validation phase, and so on.
  - When sufficient training data cannot be acquired for any case, review and loosen the coverage standards and record what should be checked individually by system integration tests in line with the original standards.
- Lv 2
  - The following activities are carried out in addition to those listed in Lv 1.
  - Grasp an approximate probability of occurrence for each attribute value or each case.
  - Check if acquired data are not deviated from the distribution.
  - Positive check other than acquisition methods should be made regarding the coverage of the data included in each case.
    - For example, in each case, when there is any attribute not included in that case, extract the distribution related to attribute and check if there is no significant bias.
- Lv 3
  - Acquire certain indicators for coverage of data included in each case in addition to those listed in Lv 2.
    - For example, check if there is no correlation between data other than attribute values included in combinations of cases using feature extraction or any other technique.
    - Or consider an expected distribution of attributes not included in each case, and analyze and record differences.

6.4. B-2: Uniformity of datasets

6.4.1. General

Evaluating only the *coverage of each case* in the previous paragraph does not always mean that all datasets are good sampling of the overall environment expressed by input data. When the probability of occurrence differs significantly from case to case, simply preparing samples for each case will generate datasets with a large bias as a whole, which may significantly impair performance, especially in terms of AI performance. On the other hand, when performance is required for rare cases that probability of occurrence is very low, it cannot be generally coexist the preparation of the practical amount of uniform data without bias for all input and the preparation of the sufficient amount of data for rare cases. For example, when 100 training data are required for an event that occurs at a frequency of one millionth, it is not usually acceptable that the number of cases of all unbiased data is 100 million. From this viewpoint, it is considered that the uniformity in this paragraph needs to make an appropriate compromise in some cases, contrary to the coverage in the previous paragraph.

In general, it is required to prepare sufficient training data for combinations of attribute values with risks which should be avoided by making correct judgments when safety is strongly sought. When such a risk occurs on rare occasions, an enormous amount of data might be required for training all other cases with sufficient *amount of data*. In such case, it is quite conceivable to *focus* on training of particularly rare risky cases.

On the other hand, when overall performance (AI performance) is required, by focusing on training rare cases more than the actual probability of occurrence, the inference accuracy deteriorates in other cases, and it may also worsen average performance. In such case, it is not always appropriate to deeply explore the coverage of detailed cases in the previous section.

Moreover, when fairness is strongly required, it may change whether the training should be artificially equivalent between cases or should be randomly trained according to the distribution of extracted training datasets, depending on a type of fairness required.

6.4.2. Approaches

The basic concept itself is to focus on cases of *overall* in the topic of coverage in Section 6.3.2. While taking care not to bias in the process of acquiring whole datasets, it is necessary to monitor the frequency of occurrence of each attribute value as appropriate.

Rather, as stated in general, in this section, it is important to consider how to coexist with the coverage in the previous section and data design.

6.4.3. Requirements for quality levels

The correspondence between external qualities level and the required level of this internal qualities are as follows:

- *Safety*
  - AISL 0.1: Lv S1 or above
  - AISL 0.2/1: Lv S2 or above
  - AISL 2~4: Requirements to be added to Lv 2 will be examined.
- *AI performance*
  - AIPL 1: Lv E1 or above
  - AIPL 2: Lv E2 or above
- *Fairness*
  - AIFL 1/2: Lv E2 or above

The requirements for each required level for the above internal quality are as follows.

- Lv E1
  - Same as Lv 1 in *Coverage of datasets* in the previous section.
- Lv E2
  - Same as Lv 2 in *Coverage of datasets* in the previous section. However, assumed probabilities of occurrence are compared with the whole sets of assumed events.
- Lv S1
  - Regarding the amount of data for each case considered in L1 of the previous section, explicitly check if there are sufficient amount of data for risk cases.
  - When data of rare cases is insufficient for training, comparing the amount of the whole sets of training data with a probability of occurrence of rare cases, consider focusing on learning of rare cases. However, especially when Lv E2 is required, with prioritized, the impact of reduced training of other cases on whole system quality should be considered.
- Lv S2
  - In addition to what is listed in Lv S1, estimate and design in advance the amount of data of each case, based on the estimated probability of occurrence for each risk event/case.

6.5. B-3: Adequacy of data

6.5.1. General

Adequacy of data means that the data used in the machine learning training and testing process are free of errors and inappropriate data, and include a variety of internal perspectives.

The data quality defined by SQuaRE (ISO/IEC 25012) [8] can be classified in terms of data-specific and system-dependent aspects used as follows. Of these, the underlined terms are considered to be particularly important as the source data for building machine learning.

- Data-specific quality characteristics
  - ✧ (1) Accuracy
  - ✧ (2) Completeness
  - ✧ (3) Consistency
  - ✧ (4) Credibility
  - ✧ (5) Currentness
- Data quality characteristics from both data-specific and system-dependent perspectives
  - ✧ Accessibility
  - ✧ Compliance
  - ✧ Confidentiality
  - ✧ Efficiency
  - ✧ (6) Precision
  - ✧ (7) Traceability
  - ✧ Understandability
- Data quality characteristics from the system-dependent perspectives
  - ✧ Availability
  - ✧ Portability
  - ✧ Recoverability

Also, in terms of the machine learning process, there are two distinct factors:

A) Data Selection Appropriateness:

Existence of each data point in the dataset is appropriate for the policy fixed by the coverage and uniformity of the dataset (B-1, B-2). For example, each data point does not contain measurement errors or is not outliers that should be removed.

[(1) Accuracy, (2) Completeness, (4) Credibility, (5) Currentness, (6) Precision, (7) Traceability]

B) Appropriateness of labeling:

For each data point in the dataset, the information added in the dataset preparation stage is appropriate.
[(2) Completeness, (3) Consistency, (4) Credibility]

These aspects are generally considered to correspond to the properties of SQuaRE as shown in square brackets.

Authenticity, which is often pointed out as data quality in the field of artificial intelligence, and is often important from the perspective of security, can be considered to correspond to (4)

Credibility and (7) *Traceability* in the above quality characteristics.

6.5.2. Approaches

Although the validity of the data is ultimately aimed at *obtaining good machine learning components*, from the perspective of quality management, it is basically considered to be evaluated against the requirements for the data expected by the system, which are defined by the internal quality characteristics from A-1 to B-2.

In the specific evaluation of this quality characteristic, it is necessary to evaluate various viewpoints in the inspection of the data itself as well as in the management of the construction process.

6.5.2.1. Unification and scrutiny of labeling policies

Although the labels to be assigned as teacher data themselves are specified in A-1 Sufficiency of problem domain analysis, there can be various fluctuations, confusions, and ambiguities in the actual data. For example, in image feature detection, the size and distance of objects to be extracted as labels, and the handling of blocking conditions of overlapping objects need to be clarified in terms of functional requirements. If these points of view are not consistent among workers, fluctuations in the labels may lead to reduced accuracy in the training process and inaccurate inspections in the testing process.

In addition, when the requirements are extended due to repeated trials in the PoC phase or rework due to insufficient accuracy, the labeling itself needs to be extended or modified. However, relabeling work is generally expensive, and even here, there may be inconsistencies in labeling among workers. Furthermore, when acquiring additional data, the environmental conditions of the data itself may not be consistent.

From these perspectives, it is important to conduct a thorough study as early as possible in the PoC phase, to solidify the labeling policy in as much detail as possible, and to record it in writing, in order to ensure traceability of quality control.

6.5.2.2. Consistency checking and rechecking of datasets

In the construction of a machine learning system, it may be possible to use existing pre-measured datasets or labeled datasets. However, since the validity of data is determined by its consistency with the requirements, the validity of existing data must be re-evaluated whenever the functional requirements or the assumption of the operating environment is changed. In

addition, when data collection or labeling work is outsourced, it is necessary to conduct inspections for acceptance from the viewpoint of quality control. How such inspections are conducted needs to be thoroughly discussed in advance, even before the process is established.

6.5.2.3. Handling the long tail and determining mismeasurement and outliers

It is possible to automatically screen for a certain amount of data being out of tendency from other data by using statistical analysis or other methods. However, whether such rare data should be taken as a meaningful training target, such as the long tail, or should be dismissed as outliers/mismeasured values may depend on the nature of the problem and the content of the individual data and may also depend on the priorities of external quality of safety and AI performance. Without a clear policy or decision-making process in this regard, the data selection and labeling process will produce ambiguities.

6.5.2.4. Addressing data contamination (security and authenticity)

When training and test data can contain intentional errors or biases, or when there are malicious modifications to the measurement environment (e.g., interference with sensors), the functionality of the final system may be severely affected. Such errors are not detected in the testing process if there is contamination in the test data, and in general cannot be prevented sufficiently by system testing.

From this point of view, not only the information security protection of the data itself, but also the physical security and diversity of the data acquisition environment must be ensured from a process management perspective, and it is also important to keep records of such quality assurance activities.

In addition, currently there is no general-purpose method to detect data contamination before training, especially in the case of procuring data from external sources. From this point of view, a system that requires a certain level of quality may have to rely on the credibility and traceability of the dataset itself or the provider to prevent data contamination

6.5.2.5. Currentness of the data

In machine learning applications, performance often degrades as time passes since the acquisition of training data. Managing the currentness of the training data is important to prevent such quality degradation. On the other hand, the requirement for currentness often conflicts with the amount of data available for training, and there can be a trade-off between currentness and completeness, especially when rare cases need to be handled. From this point

of view, it is necessary either to consider the policy of currentness in advance, or to identify a reasonable requirement level at the PoC stage.

6.5.2.6. Processes, organizations, and systems

As mentioned above, in ensuring the validity of data, there are many factors that depend not only on the specific handling of individual data but also on the overall construction process and the system, including auditing. In addition, in the actual construction of a machine learning system, handling concrete data in the process of judging the validity of data often leads to more detailed requirements definition, which may have a cascading effect on the internal quality characteristic A-1 in the previous stage. In the Guideline, this kind of work is classified as trial and error in the PoC stage. However, when this kind of circulatory work occurs, it is especially important to verify whether the final outcome is consistent. On the other hand, it is not practical to redo all data inspections from scratch in every trial. From this perspective, it is considered that quality control requires the creation of a system and structure to properly manage the process, including the management of changes in policies during development.

6.5.3. Requirements for each quality level

The correspondence between the external characteristic level and the requirement level of this internal characteristic is as follows.

- *Safety*
  - AISL 0.1: Lv 1 or above
  - AISL 0.2: Lv 2 or above
  - AISL 1: Lv 3
  - AISL 2∼4: Requirements to be added to Lv 2 will be examined.
- *AI performance*
  - AIPL 1: Lv 1 or above
  - AIPL 2: Lv 2 or above
- *Fairness*
  - AIFL 1: Lv 1 or above
  - AIFL 2: Lv 2 or above

The requirements for each required level for the above internal quality are as follows.

- Lv 1
  - General:

- ✧ Properly examine and confirm whether the source of the data is appropriate for the problem.
- ✧ Organize the labeling policy.
- ✧ Review and summarize the criteria for labeling and outlier removal in advance.
- ✧ Determine whether the criteria are appropriate in the context of the given data, and if necessary, review and recheck the criteria.
- ✧ When labeled data are used, the validity of existing labels should be examined in advance and confirmed by pre-testing, etc. if necessary.

➢ Fluctuation of labels:
- ✧ A consistent standard should be established among workers to judge labels, or a double check is to be performed.

➢ Data contamination:
- ✧ Consider the impact and likelihood of contamination of data sources.

➢ Currentness:
- ✧ Consider in advance whether the dataset contains data from an inappropriate time period, according to the characteristics of the problem.

– Lv 2

➢ In addition to Lv 1, take the following actions

➢ General:
- ✧ Incorporate the data preparation stage into the quality control process and manage it properly.
- ✧ If data are procured externally, incorporate data preparation methods, processing methods, quality control processes, and security controls into the requirements.

➢ Labeling Policy:
- ✧ Establish a control process to eliminate variations in labeling by workers.
- ✧ Establish a process to manage label changes when data attribute definitions are changed.

➢ Label Fluctuations:
- ✧ Review and document in advance the range of acceptable label variability.
- ✧ Record labelling decisions regarding fluctuations that occur during production.

➢ Data Contamination:
- ✧ Consider methods to inspect training data to the extent possible.
- ✧ Adversarial Examples should be addressed.
- ✧ Consider a design that detects anomalies at runtime.
  See Chapter 10 AI security for more details.

➢ Currentness:
- ✧ Integrate and manage the data preparation phase into the quality control process.

- A posteriori inspection:
  - If possible, consider applying input impact analysis, neuron firing status, and other internal information analysis, and manually eliminate obvious errors to the extent possible.

– Lv3
  - Take the following actions in addition to Lv2.
- Labeling policy:
  - Conduct and record a risk analysis of the impact of label design.
- Confirmation of label data removal:
  - Double-check at receiving inspection at the time of outsourcing, or set up and inspect the audit process in advance.
- Data Contamination:
  - Conduct and document a risk analysis of data contamination.

6.6. C-1: Correctness of trained models

6.6.1. General

The term *correctness of trained models* represents that a machine learning component functions as expected upon the input from the training dataset consisting of training data, test data, and validation data.

6.6.2. Approaches

The correctness of trained models is usually evaluated in terms of quantitative measures, such as accuracy, precision, recall, or F-value. These measures may overfit to the training dataset, that is, they may perform well for the training data, but badly for the data that are not included in the training dataset. Therefore, in the test phase, the correctness needs to be evaluated using a test dataset that is disjoint from the training dataset, or more generally, by applying cross validation.

Data scientists should select concrete techniques suited to the application and need to explain their selection. In Section 7.6 we will explain some examples of techniques that can be used to test the correctness of a trained model.

6.6.3. Requirements for quality levels

The relationships between these internal qualities and the required levels for external qualities are as follows:

- *Safety*
  - AISL 0.1: Lv 1 or above
  - AISL 0.2: Lv 2 or above
  - AISL 1: Lv 3
  - AISL 2~4: Requirements to be added to Lv 3 will be examined in future.
- *AI performance*
  - AIPL 1: Lv 1 or above
  - AIPL 2: Lv 2 or above
- *Fairness*
  - AIFL 1: Lv 1 or above
  - AIFL 2: Lv 2 or above

The requirements for each quality level for the above internal quality are as follows.

- Lv 1
  - Prepare a test dataset by taking into account the *coverage of data* and the past experiences, e.g., obtained in the PoC stage.
  - Prepare a training dataset in an analogous way to the test dataset. Note that the training dataset may not necessarily follow the same distribution as the test dataset.
  - Decide and record how to deal with the incorrect behavior of a trained model (e.g., false negative or false positive in the test) before the validation phase.
  - If a machine learning component is required to satisfy fairness, decide and record the methods and criteria to evaluate fairness before the validation phase.
- Lv 2
  - All the requirements listed in Lv 1.
  - Decide and explain methods and criteria to validate the trained model (e.g., accuracy and its threshold) before the validation phase.
  - Test the trained model using the given test dataset and additional test data (e.g., generated by data augmentation techniques).
  - If possible, analyze internal information on the trained model (e.g., the neuron coverage to evaluate the adequacy of testing).
- Lv 3
  - All the requirements listed in Lv 2.
  - Perform the validation/testing of the whole system (e.g., integration tests), especially by focusing on risky cases.

6.7. C-2: Stability of trained models

6.7.1. General

The stability of trained models indicates that the machine learning components respond as expected to inputs that are not part of the datasets. Low stability leads to poor prediction performance for unknown inputs (poor AI performance) and increased risk due to potential serious misjudgments (poor safety). Therefore, it is important to evaluate stability, especially when safety is required. The following two issues are well known regarding stability.

- The trained model can function incorrectly when the input is far from the training data. Typically, this may be caused when the model is overfit to the training dataset.
- The trained model can behave significantly differently when a small amount of noise is added to the input to the model. Such input noise can be either random noise in nature (e.g. dirty camera lens) or adversarial perturbation caused by malicious attacks. For example, those attacks are triggered by data poisoning in the training dataset, and by certain tricks on physical objects (e.g. attachment of tiny stickers to road signs).

6.7.2. Approaches

Stability can be evaluated and improved mainly in the following three phases in the machine learning lifecycle. Some useful techniques will be explained in Section 7.6.2.

- Training phase: A trained model's overfitting to the training dataset can be avoided, for example, by separating the validation dataset from the training dataset, by regularization techniques, by evaluating the impact of small input noise, and by monitoring the training process.
- Evaluation phase: Correctness measures (e.g., accuracy, precision, recall) can be evaluated by using synthetic data that are obtained by adding random/adversarial noise to test data.
- Operation phase: By monitoring the input to the trained model, some of adversarial input may be detected and eliminated.

6.7.3. Requirements for each quality level

It is required to record the methods applied to improve the stability of a trained model and the evaluation results of the stability. In the list below, neighboring data refer to data generated by adding small perturbation noise to original data. Examples of concrete techniques will be explained in Section 7.6.2.

The relationships between these internal qualities and the required levels for external

qualities are explained as follows:

- *Safety*
  - AISL 0.1: Lv 1 or above
  - AISL 0.2: Lv 2 or above
  - AISL 1: Lv 3
  - AISL 2~4: Requirements to be added to Lv 3 will be examined in future.
- *AI performance*
  - AIPL 1: Lv 1 or above
  - AIPL 2: Lv 2 or above
- *Fairness*
  - AIFL 1: Lv 1 or above
  - AIFL 2: Lv 2 or above

The requirements for each required level for the above internal quality are as follows.

- Lv 1: Record the concrete techniques applied to improve the generalization performance of a trained model (e.g., cross validation and regularization are widely used to prevent overfitting to the training data).
  - For stability at Lv 1, we recommend to apply some widely-accepted techniques for avoiding overtraining and overfitting, such as cross validation and regularization.
- Lv 2: Record the evaluation results of stability by using neighboring data.
  - In the evaluation of stability at Lv2, we require to use some synthetic data obtained by adding a small amount of noise to the training datasets. In particular, it is recommended to apply techniques to prevent attacks based on adversarial examples. Such techniques include robustness evaluation using adversarial examples, adversarial training to train a robust model, and dynamic detection of adversarial examples. Currently, these new methods are still being studied and developed in academic research but might be applied to system development more effectively in the future.
- Lv 3: Provide some formal guarantee to the stability for data outside datasets.
  - At Lv 3, we require to formally guarantee a certain level of stability for such data. For example, methods for certifying adversarial robustness have been studied recently and might be used in system development in the future.

6.8. D-1: Reliability of underlying software systems

6.8.1. General

The term *reliability of underlying software systems* represents that the underlying conventional software (e.g., training programs and prediction/inference programs) functions

correctly. This notion includes the software quality requirements such as the correctness of algorithms, the time or memory resource constraints, and the software security. When electronic hardware such as MPU is developed, this notion also includes the hardware reliability.

In many applications of machine learning systems, open-source software is used to train machine learning models and to develop conventional software systems. Although the quality of open-source software may not be guaranteed, the developers of machine learning systems should be responsible for ensuring sufficient quality even when using open-source software.

Furthermore, the correctness of software should usually be tested under the same environment as actual operation environments. When the test environment is different from actual operation environments, it is necessary to evaluate the differences between them. In the development of machine learning systems, however, there are often significant differences between an environment used for training (e.g., computing environment with cloud and GPU) and an actual operational environment (e.g., built-in computer). Even the behaviors of numerical calculations (e.g., the accuracy of floating-point computation) can change between them. In some cases, numerical processing itself may be changed, for example, by model compression. When implementing a machine learning system, the developer needs to cope with these environmental differences that may affect the AI performance of the system.

6.8.2. Approaches

Quality management methods for conventional software systems can be applied in order to ensure the reliability of the underlying software system.

As regard the use of open-source implementations, appropriate quality assurance measures such as those listed below should be taken to achieve the safety and reliability of the system.

Moreover, when the in-operation environment is different from the environment in the training process (e.g., when performing model compression), the Guideline recommends to test the software that reproduces the same calculations as the in-operation environment in the test phase. If this is difficult, the developer should test the whole system to check if the quality deterioration falls within the acceptable range.

6.8.3. Requirements for quality levels

- *Safety*
  - AISL 0.1: Lv 1 or above
  - AISL 0.2: Lv 2 or above
  - AISL 1: Lv 3
  - AISL 2~4: Requirements to be added to Lv 3 will be examined in future.
- *AI performance*
  - AIPL 1: Lv 1 or above
  - AIPL 2: Lv 2 or above

- *Fairness*
  - AIFL 1: Lv 1 or above
  - AIFL 2: Lv 2 or above

- Lv 1
  - Select reliable software used for the machine learning system, and record the process of this selection.
  - Monitor the system's operation to check and update the selected software.
  - Examine in advance the impact of differences between the environment in the training/test phases and the environment in the actual operation phase.
- Lv 2
  - Evaluate the reliability of the software used for the system by testing.
  - If possible, use software whose reliability is SIL 1 or equivalent.
  - Prepare for the maintenance of software during its operation.
  - In the validation and test phases, conduct validation tests in an environment that simulates the environment used in the actual operation phase. Alternatively, validate the consistency of operations of the trained model between in the test phase and the actual operation phase.
- Lv 3
  - Check the quality of software for SIL 1 (or a higher SIL level when required by the system).
  - Perform testing or formal verification of the behaviors of the trained model in an actual environment.
  - Check the consistency of those models and operations in an actual environment in any stage after integration tests.

6.9. E-1: Maintainability of qualities in operation

6.9.1. General

The term *maintainability of qualities in operation* means that internal qualities satisfied at the beginning of operation are maintained throughout operation. This concept means that internal qualities can fully respond to changes in operational environments outside the system and that any change in trained machine learning models do not cause unnecessary deterioration of quality.

A specific method of realizing the maintainability of qualities in operation depends largely on forms of operation, in particular, how to carry out additional learning and retraining. The Guideline envision the following two patterns of additional learning and retraining.

(a) Cases where the results of additional learning are reflected in an operation environment for the first time after going through additional learning and quality tests outside (developmental environment) of actual systems in service (operational environment) through explicit updates of software. Solutions such as automatic driving envisioned now belong to this pattern.

(b) Cases where trained models are updated on a real-time basis during operation and updates complete without any human tests in an operational environment. Online learning used to process streaming data, language recognition and application of conversations in projects in which development and operation are integrated belong to this pattern.

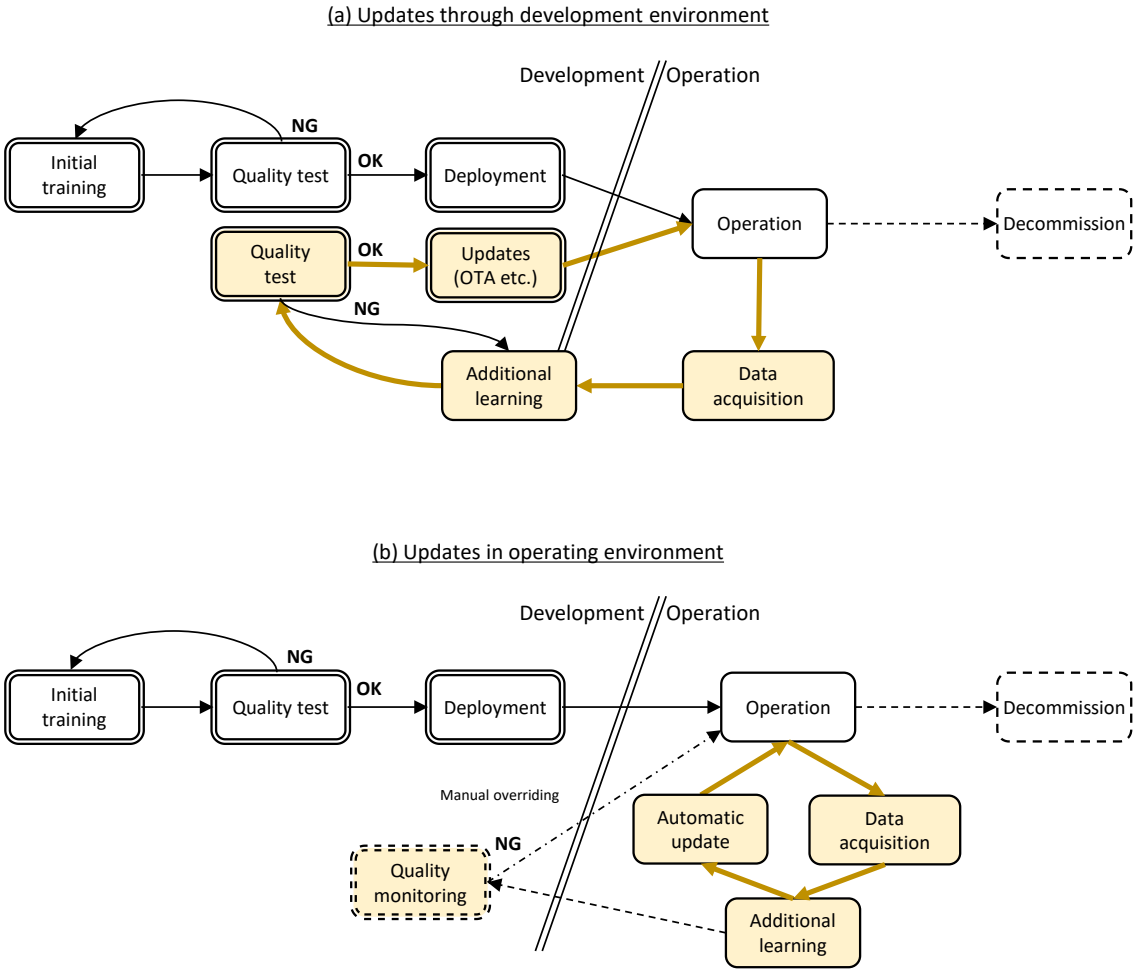

Figure 15: Forms of updates of machine learning components in operation

As described in the above figure, quality tests are always carried out before updates in the update pattern in the development environment described in (a). If the content of tests is the same as those conducted in the test phase of initial development, a certain level of quality can

be maintained. On the other hand, in the update pattern of the operational environment described in (b), there is a higher risk that a model whose quality has been deteriorated is reflected in operation, because updates are fully automatic. In this case, it is important to incorporate a quality monitoring mechanism and a mechanism dealing with deterioration in an operational system.

Moreover, as regards maintainability of qualities in operation, in addition to the deterioration of overall performance, it may be necessary to give attention to a problem that a machine learning component makes a misjudgment on specific input after its update[8], although it used to draw out correct answers prior to the update. We might think that it is enough if the external qualities described in Section 1.5 improve or are maintained as a whole, but in actual industrial fields, it may be difficult to accept that any component which was implemented and used to operate correctly stops operating properly at a later stage. On the other hand, additional learning inevitably draws out different output values from past input due to the characteristics of machine learning based systems. Therefore, it is necessary to examine in advance how to handle additional learning in the form of operational guidelines.

Moreover, though the Guideline do not cover directly, it is required to give consideration to legal positions of contracts and privacy of data in real environment used for additional learning, when operation is examined. From the viewpoint of quality management, it is desirable to save all data used for training including additional learning and use them for validating and monitoring the quality and verifying performance deterioration when a software component used to give correct answers. However, some restrictions on handling of data may be imposed from the viewpoint of privacy in actual applications. When data which are prerequisite for quality management at the time of designing a machine learning based system is not available at the time of its operation, the overall logic of system quality assurance may collapse. This is the reason why the identification of available and reservable data is an important factor to examine an operational system.

6.9.2. Approaches

There are the two patterns described earlier depending on whether quality tests are conducted by a developer, etc. when a system is updated.

(a) Cases where the quality check process comes before updates

- In advance, estimate the frequency of updates or examine judging criteria for a necessity of updates.
- Analyze the update process required at the time of operation in the design stage and

[8] In software engineering, this problem is often called “regression”. The content mentioned in this section is particularly regarded as “regression test”. However, “regression” is used in totally different contexts in the field of machine learning. Therefore, we intend to avoid the use of those terms in the Guideline.

envision and analyze its outline to design specific procedures prior to the beginning of operation. It is necessary to take note of at least the following points.

  - How to collect available data from an operational environment and deploy such data in a developmental environment to be updated.
  - Preprocessing method of data for additional training and method of placing filters and labels.
  - Range of data used for additional training and model updating. How to eliminate old data especially when environmental changes are expected due to the passage of time.
- Examine prior to the beginning of operation a method of quality tests at the time of updates, especially, judging criteria for acceptable updates (or a method of decision-making).
- It may be necessary to decide in advance how to handle some specific cases where the quality deteriorates depending on its application.

(b) Cases where updates are made without going through the quality check process in a developmental environment

- Envision in advance a possibility of notable quality deterioration caused by additional learning and the range of impact on systems if such deterioration occurs.
- When said impact can be unacceptable, examine a technical or operational treatment to accommodate risks for overall system caused by quality deterioration of learning models due to additional learning within the acceptable range and incorporate this treatment into operation. For example, there are the following ways.
  - Technically limit the range of output values and use surrounding system components (software) to prevent any deviation from the envisioned normal range.
  - Rewind learning and stop or suspend operation based on performance monitoring from the outside of an operational environment.
- If there is a possibility of monitoring the quality prior to updates in an operational environment and operational systems, such monitoring is deemed to be useful for quality management (see Figure 16).

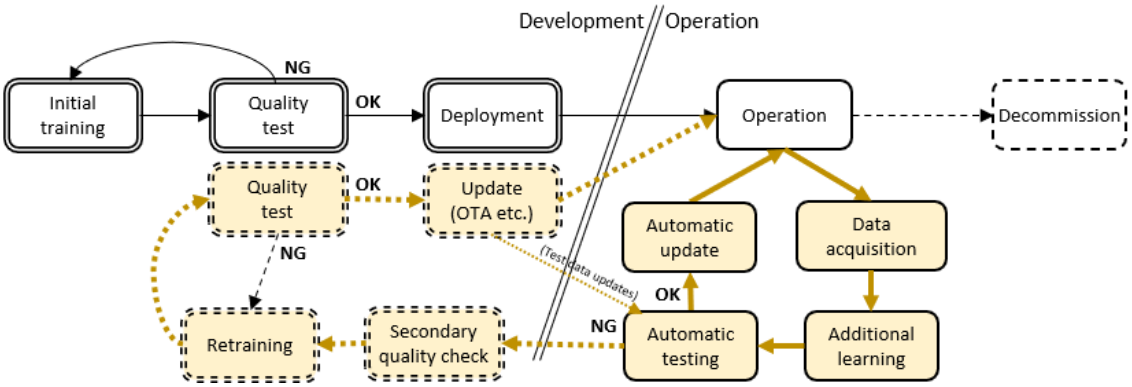

Figure 16: Possibility of automatic quality monitoring in operational environment

6.9.3. Requirements for quality levels

The relationship between required external characteristic levels and levels required for these internal characteristics are explained as follows:

- *Safety*
  - AISL 0.1: Lv 1 or above
  - AISL 0.2: Lv 2 or above
  - AISL 1: Lv 3
  - AISL 2~4: Requirements to be added to Lv 3 should be examined.
- *AI performance*
  - AIPL 1: Lv 1 or above
  - AIPL 2: Lv 2 or above
- *Fairness*
  - AIFL 1: Lv 2 or above
  - AIFL 2: Lv 3 or above

The requirements for each required level for the above internal characteristics are as follows.

- Lv 1
  - Examine in advance how to respond to notable system quality deterioration caused by changes in external environment.
  - In the case where online learning is given, examine in advance the impact of unexpected quality deterioration and take measures from the system side such as the limitation of operation range if necessary.
  - When additional learning is given off-line, quality management in line with the previous seven paragraphs should be introduced.
- Lv 2
  - Monitor system quality deterioration and misjudgments by comparing with

operation results within the range permitted by system use. It is necessary to sufficiently examine factors other than product quality such as privacy at the time of monitoring.

- When online learning is given, regularly monitor additional learning results by any method. When any deviation from the requirements for performance is found as a result of monitoring, an immediate handling should be taken.
- When additional learning is given off-line, conduct *regression tests on quality deterioration* with test datasets used in the system development stage to check if the quality has been maintained prior to updates. Update test datasets using the same method used in the system development stage where necessary.

– Lv 3

- Make sure to establish measures for monitoring system quality, including an operational system, compatible to privacy.
- When online learning is given, before the results of additional learning are reflected on systems, implement a mechanism to check quality to some extent within those systems so that updates are suspended if it becomes impossible to ignore unexpected quality deterioration. Make sure to ensure measures for making updates and modifications off-line.
- When additional learning is given off-line, the quality should be managed using data collected from operation, test datasets used for the initial system building and test datasets updated on a regular basis using the same method.

7. Technologies for quality management

7.1. A-1: Sufficiency of problem domain analysis

A problem domain analysis for machine learning called ML problem domain analysis in the Guideline is a step in the development process. Diversity of data input to machine learning components is analyzed from various viewpoints in regard to the real-world situations, in which machine learning systems or services are used. In particular, the analysis aims to concretize the identified risks or divergence in system requirements as attributes of data employed for training and testing of the machine learning products. The analysis activities are basically similar to the risk analysis, hazard analysis, or problem domain analysis, all of which are subjects in Systems Engineering. These existing technologies, or the body of knowledge, can be helpful in conducting the ML problem domain analysis.

The ML problem domain analysis, viewed from Software Engineering, is a part of initial stages of software development processes, in which software engineers often rely on trial-and-error styles of the development. In the Guideline, the steps are basically meant to establish a piece of information necessary to achieve required quality levels of machine learning products and require activities such as what data scientists conduct in developing PoC systems with their informal know-hows concerning with the quality aspects of machine learning products.

7.1.1. Initial hints

Recent standard textbooks on Software Engineering (e.g., [145]) may present an overview of the analysis methods/steps that are employed in software development processes. Because safety analysis, or safety, is one of the primary foci in the system analysis steps, several modeling notations or methodologies have been proposed and studied. They include Causal Loop Diagram and STAMP/STAP [211]. Furthermore, the methodologies are equipped with the other detailed notations such as Fault Tree Analysis, Fault Mode and Effects Analysis, Loop Diagram, Feature Tree. In the Guideline, such results are concretized as characteristics of data used to build, in particular, machine learning components. Therefore, the Guideline recommends that Feature Tree is suitable as the modeling tool for the final deliverables obtained using the other modeling notations.

There might be two major problems if the analyses follow the mentioned approach; (a) *modeling of input incidents, especially of risk factors*, (b) *design of the problem structure to be concretized as data characteristics*. This section mainly refers to those concepts specific to machine learning based systems and machine learning components from these two viewpoints.

7.1.2. Modeling of risk factors in input space

Identifying risk factors mostly involves trial-and-error activities without any systematic means, and thus may often be conducted through brainstorming sessions by various stakeholders. It is, indeed, an analysis process regarding to Known –Unknowns. How well the risk identification is done is difficult to exhibit with quantitative measures, but may be confirmed only through the inspection by human experts.

As an example of such brainstorming activities, NASA Hazard Analysis Process, published by the National Aeronautics and Space Administration (NASA) [82], presents an outlined list of what should be considered in the early stage of the hazard analysis.

1) Standard Hazard List
2) Historical experience/documentation from legacy systems
3) Your engineering training and experience

This literature also presents the *NASA Generic Hazards List* consisting of 26 items as the initial standard hazard list. Although these items include some hazard causes specific to particular applications, not causing any problem in the others, this list can be general purposed as the causes of hazard are associated with mechanical operations of systems operating in outdoor environments. The list serves as a starting point for analyzing other systems.

Moreover, 2) is similar to what data scientists conducted in the past in order to build machine learning based systems. Since findings on similar systems in the past or early versions of the same systems are valuable, it is desirable to be utilized actively. In addition, a PoC system is recommended to act as a testbed so that it helps the PoC developer understand those characteristics of final machine learning products. Finally, 3) can be considered as a form of utilizing what is called *domain knowledge*. It would be desirable to be incorporated in the body of knowledge for the users as well as the vendors involved in the machine learning software business. The NASA publication promotes to conduct the analysis activities based on these viewpoints, and then presents a conclusion that hazard analysis *Requires rational to justify hazard classification*.

The Guideline recommends that developers or engineers conduct the modeling of risk factors basically from the above three viewpoints together with those findings collected from the PoC stage and record precisely the modeling process and modeling results as well. Specifically, the followings are compiled into an integrated list of identified risk factors.

- Utilization of basic knowledge on the existing functional safety design, the existing hazard lists in each application domain, or brainstorming results based on the above NASA Hazard List;
- Utilization of analysis cases on prior systems and similar machine learning based systems;
- Introduction of domain knowledge by brainstorming with users (e.g. entrusters in contracted developments); and

– Knowledge about data employed preliminarily and exceptional cases identified in the trials of the training in the PoC stage.

7.1.3. Design of problem structure as characteristics of data

This stage of the development process produces, as an output deliverable, a piece of information on the attachment of attributes to data used for machine learning. It is affected inevitably by the characteristics of the next process, machine learning and training. Considering the training process that follows, the main activity of the process is to identify *attributes which are handled distinctively throughout the machine learning process,* which consists of the following two aspects.

4) Problematic if data are not recognized as *distinct* with respect to the differences in the output result values or risks; and

5) Problematic if data, resulting in the same output results, are likely to be understood as *different* due to their input values and characteristics of machine learning models.

4) is drawn mainly from the system specifications and the hazard analysis. A difference in output values is a difference in labels of training data, and risk incidents with different levels can basically be enumerated once the cause of hazard is identified.

On the other hand, 5) is dependent on application domains, for example, depending on differences in background images or character fonts for image recognition tasks. Moreover, 5) may be affected by preprocessing or architecture network of machine learning models. It is necessary to forecast a final implementation form to some extent in order to conduct such an analysis. This is called *Implementation Forecast* in software engineering because the information relating to the final implementation is indeed necessary to conduct the analysis at the early stage of the development. That may result in distorting risk analysis if carried out inappropriately, but it is inevitable for the quality management of machine learning components. Specifically, the activities are concerned with the implementation consideration made in the PoC development stage, and the analysis of PoC behavior using data.

In order to extract viewpoints with which the forecast is actually carried out, the three points listed in the previous section are considered applicable in view of *a risk that AI built makes different judgments in similar situations*. In the future, it would be desirable to combine 2) and 3) in the previous section and 5) in this section to accumulate system analysis results for each application area to some extent and compile them as detailed standard hazard lists in each area.

7.2. A-2: Sufficiency of data design

7.2.1. General

The internal quality in this section has two purposes, which follows the analyses described in Section 7.1. First is to prepare data needed at the time of training and testing so as to avoid *an unknown situation which a machine learning based system has not been learnt nor tested*. Second is to either decimate or integrate systematically an enormous number of combinations of prepared attributes so that building machine learning software of a required quality level is practical.

If a system intends to solve a simple problem and the number of combinations of attributes is a few dozen, it is possible to consider the amount of data that achieves a certain recognition level for each attribute. (e.g., Only 10 characters x 2 fonts need to be recognized in number recognition and there is no need to take into account differences in paper quality. Then, there are 20 combinations.) If each attribute has such an amount of data, sufficient and comprehensive training can be practiced. However, since the number of combinations of attributes often exceed several tens of thousands if calculated in a naive manner, it is not practical to prepare data representing all combinations for testing. In a certain case, training data or testing data may be a few or even null as a result of breaking up attribute combinations. It would then be impossible to rely on the Law of large numbers to make a judgment such as the convergence of training.

Combination testing is developed in conventional software engineering for mitigating the complexity problem. The method, employing the notion of *test coverage*, establishes the degree to which combinations of attribute are detailed and the degree to which combinations of test condition are covered, yet devises them separately. The method is often used to evaluate test data in the conventional software testing, and is also expected to be effective for the case of machine learning, because training is data-centric. For example, [57] presents examples to apply the combination test technology (t-way) to evaluate dataset for machine learning.

7.3. B-1: Coverage of datasets

Establishing coverage of datasets is an extremely difficult but important problem in quality management of machine learning. A dataset, if a certain anomaly is associated with it, may derive trained machine learning models to exhibit unstable or unfavorable functional behavior in a certain specific envisioned situation, lowering fairness levels.

As the *problem domain analysis* is finished, this section focuses on the issues with small differences in the dataset. The differences are so small that they are easily overlooked at the time of constructing the dataset. The discussions in this section are mostly related to the preliminary stages of data preparation or of data evaluation in the development lifecycle, auxiliary checks may be sometime conducted in the software testing stage as well.

7.3.1. Plan for data acquisition

A data acquisition plan in the data preparation stage is a basis to study the issues relating to the coverage of dataset. In order to construct a dataset that are unbiased in view of operation time circumstances, the plan must refer to the range, period and size of data collection. Specific plans should be considered for each application following outcomes of the brainstorming sessions, which may be those activities, in the stage of analyzing *domain analysis problem*, to utilize prior knowledge about the similar applications so as to examine in advance the degree of diversity.

7.3.2. Pre-flight tests in data scrutinization stage

Auxiliary checking of data distribution belonging to certain attributes is desirable in some cases. These attributes are what, although identified as candidates in the early stages of the analysis, may not be adopted, or what may become latent, because they are composed to be a new attribute. The composed attributes are accompanied with explicit specifications, and in some cases those checks may be conducted mechanically, because data labels are given.

This is not perfect, because attributes overlooked in the brainstorming sessions cannot be recovered. It is worth considering.

Moreover, for some data, what is called data specific feature may be identified in the preprocessing stage, which makes it possible to check the distribution.

7.3.3. Additional tests in testing stage

In a case where the data distribution itself has some problems, what can be re-checked in the test stage is limited. For instance, there may be latent correlations between the adopted features and overlooked unidentified ones. In such cases, analyses of correlation or influence between the input values and their resultant counterparts' output from the trained machine learning model may be helpful to see whether the model behaves as expected. The model may make inferences based on features different from what are assumed. It is worth considering this analysis when extreme safety is a major concern.

7.4. B-2: Uniformity of datasets

The uniformity of datasets is equivalent to considering the coverage of datasets, described in Section 7.3, as a single *case* of all the input data. The measures listed in each section in Section 7.3 are deemed applicable here as well.

7.5. B-3: Adequacy of data

7.5.1. Quality control cycle from the perspective of data

The Guideline mainly examines the lifecycle process from the perspective of the development of machine learning models, but from the perspective of data quality, the lifecycle process of establishing consistency between data collection policies and actual data is also important. From the standpoint of data, there may be a recursive process iterating the following steps.

1) Determination of data collection policy
2) Collection and selection of data in accordance with the policy
3) Validation of both the policy and data

In (1) determining the data collection policy,

- Formulation of a data collection policy that is appropriate for the quality of the target machine learning system
- Advance consideration of the acceptable range of data validity
- Determination of the data validity evaluation method in advance

are necessary.

In the process of (2) data collection and selection, we perform

- Determination of the appropriateness of the selection and collection method,
- Process management of the actual selection and collection methods,
- Collection and consolidation of evidence on selection and collection methods,
- Management of meta-information about the source and traceability of data

in accordance with a defined policy.

In (3) validation, we perform

- Verification of data in accordance with the policy,
- Review of the explanatory nature of the policy itself,
- Feedback to the policy based on the collected data,

And the process continues back to (1) until a sufficiently satisfactory and valid policy and data pair are obtained. Furthermore, this entire process is subject to quality management and assurance, such as:

- Evaluation of the completion of the process itself,
- Evaluation of the skills of the people in charge of executing the process and their appropriate assignment,

- Evaluation of the mechanism for managing the entire process and securing of evidence,
- Evaluation of the system for managing the entire process and ensuring its traceability.

In relation to the development lifecycle described in Section 1.8, such a cycle of data policy improvement is abstractly organized as a repetition of trial and error in the PoC process, but in reality, sometimes the improvement cycle described above is repeated even in actual development process, leading to agile developments. In such a case, it is particularly important to assess whether the requirements definition has been updated together with the data collection policy, or whether it is necessary to reconfirm the contents identified in the previous stages such as internal quality A-1 to B-2, etc. in accordance with such updates, in order to ensure the overall quality. It is also extremely important to ensure that such policy and requirement updates are subject to change management in order to reconfirm and verify quality later on.

7.5.2. Technical support for organizing outliers and corner cases

There are several techniques that can be applied to efficiently extract and organize data that require data validity judgments either as outliers or as rare cases to be addressed.

First, DeepXplore [159] uses the Neuron Coverage technique for corner case extraction. Surprise Adequacy [113] provides a measure of the *surprise* of the inputs to a machine learning model, which also allows corner cases to be extracted from the input values. Tinghui et. al [153] improved on Distance-based Surprise Adequacy to perform corner case analysis. Whether the data extracted in this way should be removed as outliers or treated as important rare cases is to be determined on a case-by-case basis in conjunction with policies. Section 5.3 in Zhang [208] and Section 2.3 in Dong [84] are also helpful.

7.6. C-1/C-2: Correctness and stability of trained models

Checking the correctness and stability of machine learning models is a task corresponding to *unit testing* in conventional software testing. It is because the check is focused on each machine learning component. In this section, we will introduce the following topics.

- Basics of software testing techniques [53], the aspects to be taken into account depending on the characteristics of machine learning components, and recent research works on machine learning component testing
- Potential of technology for testing of stability directly

7.6.1. Testing machine learning components

7.6.1.1. Characteristics of test target

First, there are two distinctive test targets to consider when discussing the quality of machine learning component results: training and learning programs that input training datasets to obtain trained models, and prediction and inference programs of which the resulting trained models specify functional behavior.

Second, there is no clear criterion for deciding whether a computation result is correct, which is called test Oracle problems [58]. For a training and learning program, it is difficult to know in advance the correct value of its computation result, that is, a trained model. In addition, the answers derived by prediction and inference programs are not definite but accompanied by probabilities referring to certainty degrees, which illustrates difficulties to define absolute criteria for correct answers [87]. In general, machine learning programs are classified as non-testable programs [199].

7.6.1.2. Test oracle problems

Software testing technology consist of test input generation methods and test execution result confirmation methods. The first is how to generate test inputs that efficiently achieve the testing purpose. As we will see later, there are some interesting research works on test case generation methods for machine learning components.

The second is concerned with methods for comparing the results of program execution with known correct answers to a given test inputs and is collectively referred to as test oracle techniques. A non-testable program is one in which the correct answer to a given test input is not known, and derived oracles, pseudo-oracles, or partial oracles have been used in software testing so far.

Metamorphic Testing (MT) [76] is a practical method of partial oracles. It was introduced as a testing method for programs in which it is difficult to know in advance the correct value of the computation result for an input, such as numerical computation or translators involving compilers, and was later applied to the testing of operating systems and security systems to use implicit oracles [77]. It is now a standard testing methodology for machine learning components [140]. For some testing purposes, a combination of partial oracles and statistical testing can be useful [142].

7.6.1.3. Two perspectives on testing

In general, when checking a program, positive testing and negative testing are conducted.

Positive testing is to confirm that the test target exhibits the expected functional behavior. For

the sake of explanation, let's assume that a program has pre-conditions and post-conditions. The relation is established such as *when the pre-condition is satisfied, the post-condition is satisfied in the state after the program body is executed*. When the above relationship holds for test input data that satisfies the pre-conditions, the program under test is considered to pass the test.

Negative testing checks whether a program will not fail catastrophically under unexpected conditions. In general, a program is expected to behave reasonably even in response to input that does not satisfy its pre-conditions. For example, the program should not run out of control and terminate abnormally. Checking these cases does not need an application-specific correctness criterion. Such test oracles are called implicit oracles, meaning that there is no explicit test Oracle at all.

Fuzz Testing or fuzzing [131], which uses randomly generated data, while satisfying some constraints, as test input, has been known effective for integration testing of open system software such as operating systems or security systems. This method is sometimes used in negative testing of programs. An appropriate fuzz, a test input data, is generated to meet each test objective.

In testing for machine learning components, positive testing is, in a sense, related to checking the accuracy. We can use data that follow the same statistical characteristics as the training dataset. On the other hand, negative testing includes checking the accuracy in regard to the generalization performance and checking behavior against unexpected inputs in the case of checking the stability or the model robustness. Stability testing includes the case where the input is either corrupted data or adversarial examples. As described below, various fuzzing techniques have been proposed for testing machine learning components, depending on the purpose and perspective of the test.

7.6.1.4. Test coverage metrics

In machine learning component testing, it is important to know an extent to which a target trained model becomes activated for a given test input. The idea corresponds to the test coverage criteria in conventional software testing methods [53]. Neuron Coverage (NC) [159] is, an early proposal for such a structural metric, defined by a frequency of activated neurons. Later, in order to use more detailed information than NC, several coverage metrics are proposed [126]; those are based on structural patterns over activated neurons, such as correlations between activated neurons in different layers. If these structural coverage values are large, it can be assumed that a wide range of the learning model is activated, and thus is checked as well. The test coverage is considered sufficient and thus we may conclude that such test inputs are useful.

In general, it is difficult to define the test inputs used for negative testing which do not satisfy pre-conditions, because the statistical characteristics of training data, which might be regarded as such pre-conditions, are not explicitly stated for machine learning components. Surprise Adequacy [113] provides ways to check how test inputs deviate from the statistical characteristics of the training data. Although the adequacy metrics are originally introduced for

evaluating quality of test inputs, they are also considered to be a kind of indicators referring to degrees of being outliers. Note that these indicators refer to the internal activated states of trained models, and thus that measuring these requires monitoring of execution states of trained models.

7.6.1.5. Automatic generation of test inputs

In general, it is important to establish a method on generating appropriate test input data for each test target. In addition, such an appropriateness depends on the testing purpose, either positive or negative.

When the test target is a training and learning program, the input is a collection of data (a dataset), and the notion of dataset diversity [138] provides a guideline for test input generation methodologies for both positive and negative testing.

In testing of prediction and inference programs, various types of data that were not considered during training stages are expected to input for checking prediction results. The method is sometimes called Test-time Augmentation. For generating test data, there have been various approaches including those to adopt machine learning techniques. They include conventional Data Augmentation methods for image data [116], or methods using Generative Adversarial Networks (GANs) [97]. Furthermore, some methods make use of the adversarial example generation method[144] [210]. In the case of generating a large number of test data, the aforementioned dataset diversity is also useful as a conceptual guideline.

7.6.1.6. Test generation and coverage metrics

In generating test inputs, it is desirable to find useful data that can be used to conduct testing efficiently depending on test purposes. As in conventional software testing techniques, we apply some metrics to control the process of generating data that, for example, improves the structural coverage criteria. This method is collectively called Coverage-Guided Test Generations. There are some research works that use the NC when machine learning components are test targets. The NC is combined with the classical data augmentation method [187], or with the GAN-based data generation method [209].

Some recent works report that the structural coverage may not be useful depending on the purpose and perspective of the test [99]. For example, the correlation between NC and adversary robustness is known to be small [194]. The NC values are more highly correlated with model capacity than with defects in trained models [215 Chapter 4]. For example, experimental results show that the accuracy can be high even if the NC is small, and conversely that the NC value can be large even when the accuracy is poor. In addition, it is known in conventional software testing techniques that the structural test coverage indices have a small correlation with the effectiveness of test suits or the efficiency of detecting faults[103]. This empirical report results seem to be applicable to machine learning components as well[139]. As alternatives to the

structural coverage criteria, a new method is proposed that is based on an error function to be used as metrics[142][215].

Although identifying data for executing corner cases is desirable for effective testing, the relationship between inputs and corner cases is not clear in machine learning components. Therefore, it is necessary to consider a metric that replaces the structural coverage criterion. For example, the method using the error function is used for evaluation metrics [194].

7.6.1.7. Prioritization for Preparing Training Data

In conventional software testing, regression testing involves a large number of test cases. Selecting test cases that are useful for detecting faults early can improve the overall efficiency of test activities. Therefore, when a large number of test cases are available, prioritization is performed [167], and choosing metrics is essential so as to find useful subset of the test cases. For machine learning, the prioritization ordering techniques can be utilized from the perspective of improving efficiency of work for preparing training data, especially of work for labeling [67].

We consider here a task of augmenting an existing training dataset when a lot of data without any tag is available. First, candidate data are picked up in a certain way so as to be put into a data pool. Second, labeling data, selected from the pool, with a tag is done manually, which is time-consuming to require a lot of human work if a number of target data is large. Therefore, if we select data that have characteristics different from the existing training data and label these data only, we can obtain useful training data while reducing the amount of work.

Technically, the method is essentially deciding whether data is useful or not, and it comes down to another problem of devising appropriate indices to be used for the decision. Recent research works propose a number of methods using either external metrics [92] such as Confidence and Gini Impurity, or internal metrics [67]. Choosing an appropriate metric should take into account what different features are desirable for data in regard to the existing training data.

7.6.2. Technologies on stability issues

The technologies on stability issues are broadly divided into the evaluation and improvement. Figure 17 illustrates how each technology is related to the level described in Section 6.7.3.

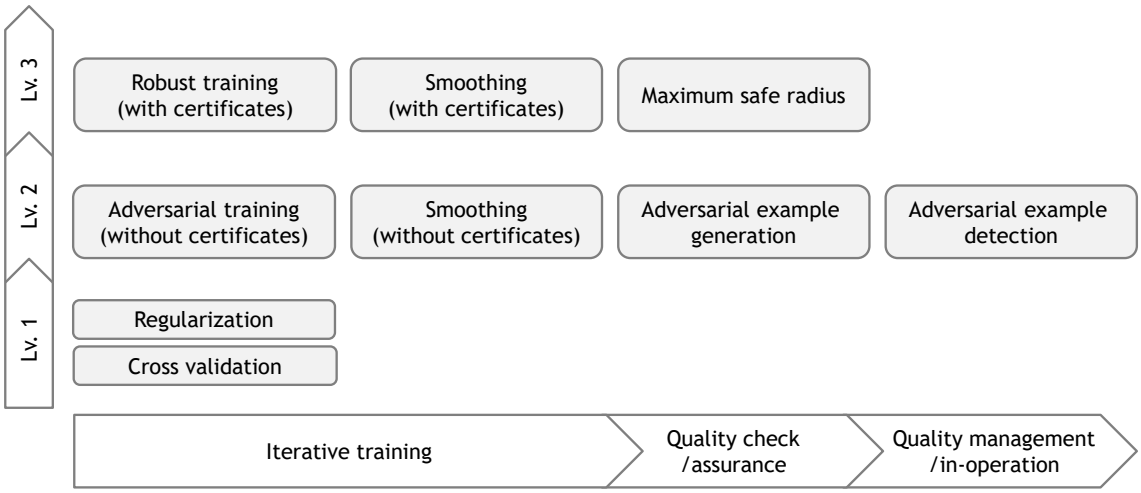

Figure 17: Technologies to evaluate and improve stability (process and levels to which technologies are applied)

7.6.2.1. Cross validation

Cross validation is a classical method to mitigate the over-fitting problem, and thus to improve stability levels. The idea is simply dividing a whole dataset into training, validation, and testing datasets [114]. For example, in K-division cross validation, a whole dataset is divided by K, and $1/K$ of the dataset is used for validation and the remaining $((K-1)/K)$ datasets for training. The roles of datasets, either training or validation are interchanged. Cross validation is recommended to Lv 1.

7.6.2.2. Regularization

Regularization is a methodology to mitigate the over-fitting problem, and thus to improve stability levels. Especially, regularization methods suppress the absolute values of learnt weight parameters not to become excessively large [149]. For example, loss functions may include a regularization term to increase as absolute values of parameters become large). Dropout is another regularization method [181], in which neurons are randomly excluded from the training target during the training process. Dropout may obtain the same effects as cases where several machine learning models are trained simultaneously. Regularization is recommended to Lv 1 if the stability is an issue.

7.6.2.3. Adversarial example generation

Adversarial examples are those data that cause miss-inference in machine learning components. Such data are augmented with a slight perturbation, a semantic noise, that human

eyes are not able to recognize. Various methods to generate adversarial examples have been proposed [71][159]. A stable machine learning component is expected not to cause miss-inference even when the perturbation with the input adversarial data is large. Therefore, the degree of perturbation is used to evaluate the stability. Unfortunately, no practical tool to generate such adversarial examples have been established, this technology will be hopefully applicable to Lv. 2 evaluations in the future.

7.6.2.4. Maximum safe radius

The maximum safe radius refers to the minimum distance between the original data and its adversarial example. Adversarial example generation (Section 7.6.2.3) looks for nearby adversarial examples, while the maximum safe radius guarantees that there is no adversarial example in the neighborhood, inside the hyper-sphere of the maximum safe radius. Methods to accurately calculate the maximum safe radius are proposed in [109][188] that use SMT solvers. However, a cost of accurately calculating the maximum safe radius is so high that the size (e.g., number of neurons) of networks is limited. Alternatively, methods to approximately calculate a safe radius smaller than the maximum safe radius are proposed in [62][198][201]. Moreover, a method to approximately calculate a radius that probabilistically guarantees safety of not 100% is proposed in [197]. These technologies are still in the research stage, but they can guarantee that there is no adversarial example inside the hyper-sphere of the maximum safe radius. Therefore, they are expected to be applicable to Lv.3 in the future.

7.6.2.5. Generalization error bound

The generalization error is the expected value of the incorrect answer rate for all input data that is not only for a particular dataset. In general, it is difficult to measure the rate of incorrect answers because of the large number of input data. However, it is possible to estimate a probabilistic upper bound on the generalization error which *guarantees with probability p % that the generalization error is less than e %* [190]. At present, the method is still in the research phase and the accuracy of its estimate is not yet high. However, it is expected to be applied at Lv.3 in the future as a measure of generalization performance.

7.6.2.6. Adversarial training/robust training

Adversarial training is a learning method, which looks for neighborhood data that are likely to cause miss-inference [128]. Compared to the standard training method, it may be able to improve the resistance of trained machine learning models against adversarial examples. Although this technology is still in the research stage, it is expected to be applied to Lv.2 in the future.

On the other hand, a robust training method is to eliminate neighborhood adversarial examples [200]. An approximate calculation method of the maximum safe radius is given together to guarantee that adversarial examples do not exist near each data point in the training dataset. Although this technology is still in the research stage, it is expected to be applied to Lv.3 in the future.

7.6.2.7. Randomized smoothing

Randomized smoothing is a method to augment data with randomized noise so as to calculate the final result value to be a mean with respect to the noise distribution. The method is reported to improve the stability against the L2-norm adversarial attacks [123]. Furthermore, methods of adding randomized noise are proposed so as to guarantee the correctness of inference results [80][118]. In these methods, the randomized smoothing is applied to the neighborhood of input data. Because the statistically expected value is approximated by an average, multiple running results are needed. Furthermore, the stability degree is increased as the randomized noise is large, which in turn lowers the correctness. Such a trade-off relation must be taken into account. Nevertheless, this technology is expected to be applied to Lv.3 in order to guarantee the stability of inference results.

7.6.2.8. Adversarial example detection

Adversarial examples detection is a method to check, at runtime, whether incoming data are adversarial or not, and is one of the active research areas [127][203]. Although such a technology is still in the research stage, they will be expected to be applied to Lv.2 for protecting potential adversarial examples in machine learning components.

7.7. D-1: Reliability of underlying software system

7.7.1. General

The reliability of underlying software system is an important item even in conventional software. Quality management is expected to be difficult especially in implementation of machine learning AI which uses a large number of libraries including open-source libraries. Open-source software does not usually count on guarantee. Therefore, in relation to end users, open-source users (e.g., developers and operators) are responsible not only for differences in malfunction but also for discovering and monitoring potential errors and, in some cases, making modifications.

Moreover, when a machine learning model is built, it has been revealed that a training process incorporates bugs (e.g., program error) so that the impact of bugs does not appear in

tests, causing quality deterioration [141].

On the other hand, in relation to conventional software, configuration management and quality monitoring of libraries and fundamental software are emphasized and infrastructure and services for configuration management and quality monitoring are improving. Some data included in those services contain information such as libraries specific to machine learning, although they are somewhat limited. These existing infrastructures are deemed worth being utilized in applications of machine learning.

7.7.2. Quality management of open-source software

It is desirable for each business operator to think about how much open-source libraries can be trusted and their quality maintained by itself or outsourced.

It can be expected that supported libraries whose quality is assured and software that went through the quality inspection process are used where necessary in relation to a necessary quality level.

7.7.3. Configuration management and tracking of bug information

For example, Common Platform Enumeration (CPE) [40][213] is used for configuration management of software components as a common ID to list system constituent components, in the security field. There are commercial products that manage versions of software components in use and extract information on their updates based on CPE. A list of vulnerability information related thereto called Common Vulnerability Enumeration (CVE)[41][214] does not include bugs that are not directly related to security vulnerability. However, at least tools related to CPE are very likely to be beneficial to track the latest version of libraries and infrastructure software.

7.7.4. Possibility of specific check thorough testing

Moreover, when constituting software components have any bug, that bug does not always have a direct effect on actual ML results in machine learning components different from conventional software. If software with bugs is placed in a feedback loop of learning or training, a trained model *memorizes* behaviors of said bug and training seems to be superficially successful in some cases. In these cases, it is reported that potential quality deterioration caused by software bugs can be found by analyzing statistical behaviors using any of the metamorphic test technologies listed in Section 7.6.1.

7.7.5. Software update and possible adverse effects on performance and operation

On the other hand, actual software whose configuration is complex often operates

unintentionally even if a developer of software libraries updates it with the intention of improving performance and operation. Depending on how a machine learning based system is configured, behaviors of bugs are sometimes adapted or learnt in the training process. Therefore, when any software constituent component is updated, its operation and performance must be reviewed. In some cases, it is important, to make a judgment to start training over or keep using an older version taking into account vulnerability.

The Guideline cannot recommend any option, because a specific judgment depends on each situation. However, it is important to record the background of each judgment for accountability reasons in order to claim *the proper quality management.*

7.8. E-1: Maintainability of qualities in operation

This section describes technologies to maintain internal qualities satisfied at the commencement of operation throughout the operation period. In order to maintain internal qualities realized at the beginning of operation in spite of changes in external environments that may arise during operation, a machine learning component needs to respond to those changes. As described in Section 6.8.1, there are two operational patterns therefor. One is to make batch-processing updates by returning to the developmental environment and deploying it again. Another is to automatically update the software component when needed or at a high frequency in the operational environment. In the former pattern, monitoring to determine the timing of update and update processing are main elements, while in the latter pattern, automated update processing is a main element. In order to realize this, on-line learning [173] is adopted. Even if updates are made automatically, it is necessary to monitor if automatic updates operate properly and process updates to respond to cases where there is any deviation from normal operation conditions.

Some of monitoring technologies necessary for both patterns are presented here. We focus especially on technologies to detect changes in data distribution over time called concept drift [94]. Moreover, technologies to retrain machine learning models which play the core role of updating trained machine learning models and technologies to create additional training data used therein will be presented.

7.8.1. Monitoring

The relation between newly acquired input data and output (e.g., inference) results may have changed at the time of operation from the relation between them at the time of training due to various factors including changes in external environments. When a machine learning model trained in the design or development stage using such data whose input and output relations change is kept using during operation, its performance (e.g., accuracy) may deteriorate and result in serious damage. Therefore, it is required to continuously monitor behaviors of machine learning based systems and machine learning components for the purpose of checking if the

quality fulfilled at the beginning of operation is maintained throughout the operation period.

There are the following four tasks of monitoring during operation.

- Accuracy monitoring
- KPI monitoring
- Model output monitoring
- Input data monitoring

Accuracy monitoring directly measures the accuracy of trained machine learning models. This monitoring is divided into some patterns in accordance with the method of collecting correct answers compared with inference results of trained machine learning models required for calculating the accuracy. That is, after making an inference, (1) cases where correct answers are acquired automatically after a certain period, (2) cases where correct answers cannot be acquired automatically so that it is necessary to manually put labels and (3) cases where manual labeling is beyond budget or it is impossible to put labels. It is necessary to select an appropriate monitoring method in accordance with the above grouping of cases where correct answers are collected in order to appropriately monitor the accuracy. Moreover, some applications emphasize not only monitoring the accuracies of models but also monitoring from the viewpoint of KPI, *KPI monitoring*, in line with a conversion rate and the benefit of users. In this case, it is required to monitor the consistency between the accuracies of the models and the KPIs of the applications.

Model output monitoring and *input data monitoring* refer to the monitoring of results of inferences made by a trained machine learning model and the monitoring of its input data, respectively. The monitoring methods are divided into human monitoring where 100% sampling, automated monitoring when alert conditions are known and filtering when conditions with higher possibilities of alert are known. Model output monitoring is further categorized into a case where each output inference is checked by experts as in the case of medical diagnostic and a case where all inferences are checked altogether after a certain period of time. In the same way, various conditions may be imposed in the case of monitoring by filtering (For example, false positive is acceptable but false negative is unacceptable). In addition, when input data are monitored, whether an alert is issued in the case of monitoring by filtering or input is disregarded depends on an application.

7.8.2. Concept drift detection methods

Concept drift is one of major causes of the deterioration of the accuracy of trained machine learning models during operation. A variety of monitoring or detection methods have been proposed recently. Concept drift detection methods are categorized as shown in Table 4 in accordance with whether correct-answer labels on data acquired during operation are used [171].

Table 4: Classification and characteristics of concept drift detection methods

Use of label	Method	Characteristics
Labeled detection (Supervised detection)	Sequential analysis	Monitoring of absolute values such as accuracy
	Statistical Process Control	Monitoring of the increase or decrease of error rate (Early detection by finding indicator)
	Window based distribution monitoring	Monitoring of distribution differences between training time and operation time
Unlabeled detection (Unsupervised detection)	Novelty detection/clustering method	Simple detection by clustering
	Multivariate distribution monitoring	Detection by applying statistical hypothesis testing to data distribution
	Model dependent monitoring	Output dependent of models, such as confidence score

Studies of unlabeled detection methods have been carried out actively in recent years. For example, a literature [91] proposes an unknown class detection algorithm (MINAS) in multi-class problem based on clustering methods such as k-means. Moreover, another literature [164] proposes an algorithm of incremental KS test which improved the amount of calculation by developing incremental Kolmogorov-Smirnov test to judge whether samples are generated from the same distribution. Moreover, a literature [121] proposes a method of detecting data changes by using confidence scores associated with output from trained machine learning models without using labels (CDBD).

7.8.3. Retraining

When any change in data distribution or deterioration in the accuracy of a trained model is detected as a result of the monitoring describe above, it is necessary to retrain the machine learning models using datasets which add recent data or replaced recent data with existing training data. Many researches have been made about this retraining. For example, a literature [186] provides a platform design method to automatically determine the appropriate timing of model update in relation to existing machine learning frameworks. Moreover, it has also been proposed a method of applying balanced parameters of both *existing models* and *models that learnt new tasks* to retrain models in order to reduce forgetting of old tasks caused by retraining of a neural network called catastrophic forgetting [119].

7.8.4. Creation of additional training data

There are many cases where labels cannot be collected automatically. In this case, it takes much cost to manually place labels to new data acquired at the time of operation. A study on reduced costs of retraining by reducing the number of data points to which labels are attached has been proposed through a method of reducing labeling work using software equipped with GUI (Graphical User Interface) [191] and active learning have been proposed to solve this problem [172].

8. Fairness

In this chapter, we analyze the social background and problems of quality management regarding fairness, as well as the technical issues specific to fairness. It also summarizes the issues that should be considered as internal quality characteristics to address fairness.

8.1. Background

8.1.1. Social demands and social principles

8.1.1.1. AI social principles, etc.

In recent years, there has been a lot of discussion about *ethical* and fairness issues of AI, and the norms in society of AI. Correspondingly, all the documents on AI social principles referred in Section 1.9 explicitly address requirements for fairness and ethics as important topics. The Council for Comprehensive Innovation Strategy's "Principles for a Human-Centered AI Society"[22] states in the point 6 *Principles of fairness, accountability and transparency* that *under the design concept of AI, all people shall be treated fairly without unfair discrimination on the basis of their race, gender, nationality, age, political beliefs, religion, or other diverse backgrounds*, and requests the sufficient consideration regarding fairness from the design phase of AI system. The OECD AI Principles [24] request AI systems to be designed with human rights and diversity in mind, and to include appropriate *safeguard* mechanisms against the fairness related issues. Furthermore, based on the OECD AI Principles, AI fairness subjects have been mentioned in EU AI White Paper [28] and the European AI High-Level Expert Group's AI Transparency Guidelines [30] ,and is considered to be gaining consensus at the principle level [33][88].

8.1.2. AI governance

According to a report by the Ministry of Economy, Trade and Industry (METI) [33], AI governance is *the design and operation of technical, organizational, and social systems by stakeholders with the aim of maximizing the positive impact of AI while managing the risks arising from its use at a level acceptable to stakeholders*. The principles exemplified in the previous section are generally conceptual, technology-neutral, and cross-cutting statements of goals, while AI governance aims at achieving these goals through activities of the developers, supervisors, and regulatory enforcers, based on various grounds such as legally binding rules, non-legally binding guidelines, and international standards.

8.1.2.1. Legally binding rules

Fairness has a long history in the U.S. legal system and has been the basis for several concepts. More than half a century ago, the Civil Rights Act of 1964 was enacted to prohibit discrimination by race, religion, etc. Title VII of this act addresses discrimination in employment. Through the administration of these laws, two important legal concepts, Disparate impact and Disparate treatment, have been established. However, these laws and other historic laws do not address issues specific to current AI system technology. For example, Facebook's 2019 lawsuit was [9]also based on technology-neutral laws that are not limited to AI.

Recently, in April 2021, the EU announced an AI-specific policy package, including a regulatory framework draft on AI [25]. There has been a lively debate among experts that legal regulations that apply to *a wide range* of AI fields may be a significant impediment to innovation. In terms of the severity of the restrictions, the EU's draft regulation takes a pragmatic, risk-based approach to classifying AI systems according to their intended use. For example, all AI systems those may contravene EU's values, for instance by violating fundamental rights will be classified as *unacceptable risk AI* and prohibited in principle. AI systems that create a high risk to the health and safety or fundamental rights of natural persons are classified as *high-risk AI* and subject to compliance with certain mandatory requirements and an ex-ante conformity assessment. As with the case for GDPR, when the regulation is formally passed after further deliberations, it is expected to apply not only to AI system providers within the EU, but also to AI system providers outside the EU if used within the EU.

8.1.2.2. Non-legally binding guidelines

In general, it can be assumed that non-legally binding guidelines serve two purposes.

1) In cases where legal regulations exist, to specify the means to actually fulfill the requirements (as detailed checklists, etc),

2) In cases where legal restrictions do not apply, such as low-risk classification case, provide the basis for ensuring the quality of the AI products and services.

For example, "The Assessment List For Trustworthy Artificial Intelligence (ALTAI)" [31] by the European High Level AI Expert Group addresses fairness as a *Diversity, Non-discrimination, and Fairness* requirement, and lists the following checklist items as the purpose of *avoiding unfair bias*:

[9] In March 2019, the U.S. Department of Housing and Urban Development filed a lawsuit against Facebook for violating the Fair Housing Act when its targeting ads excluded certain demographics from ad delivery based on race and other characteristics.

- Did you establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design?
- Did you consider diversity and representativeness of end-users and/or subjects in the data?
- Is your definition of fairness[10] commonly used and implemented in any phase of the process of setting up the AI system? (Did you consider other definitions of fairness before choosing this one? etc.)

The Guideline can also be utilized for both the purposes I and II mentioned above.

8.1.2.3. International standard

Reports and standards on ethics, transparency, and fairness of AI are being developed in ISO, IEC, and IEEE. These international standards are described in Section 11.2.

8.1.2.4. Inclusiveness

Addressing fairness is also essentially *achieving diversity*, and *inclusive development, design and deployment* that involves various communities from the beginning of the development process could help prevent further social damage and reduce existing social inequalities. An international multi-stakeholder initiative Global Partnership on Artificial Intelligence was established in June 2020 to achieve responsible, human-centered AI development and use, including this *inclusiveness*.

8.2. Ethics and fairness definitions in the Guideline

At present, we can hardly find out clear-cut definitions of terms such as *ethics* and *fairness*, which are fully agreed in both the social and technical domains. In the Guideline, we define *ethics* as *good behavior in the context of real-world norms* as a property of the field of sociology. This ethics is not limited to fairness but may also include policies on deterring certain forms of judgment, certain kinds of safety consideration, and on dilemmatic decisions. In general, ethical requirements for systems are implicitly presented by society, are derived from social norms that are not always explicitly stated, and serve as constraints and considerations for defining

[10] *Fairness* here is roughly equivalent to the *fairness metrics* of the Guideline.

requirements mainly in the planning and design stages of systems.

On the other hand, we narrowly define *fairness* as *not being treated differently, on some defined basis, due to differences in the inputs that are not defined as requirements*. In discussions at the level of social norms, this definition of fairness is considered to be part of ethics.

The fairness of the machine learning components that the Guideline aims to achieve is defined further narrowly with engineering focus, as a set of identification requirements that are subject to specific fairness guarantees, and that are discovered at the requirements definition stage from requests for ethics, functionality, etc., and are derived by the system provider as specific requirements from the requirements definition.

8.3. Difficulties with fairness

8.3.1. Diversity of requirements

In the previous section, we described the fundamental definition of fairness in the Guideline. However, there are multiple perspectives to explain, assure, and convince people of *being treated equally*, and often, just saying *being fair* is not enough. A fair system would be developed only if fairness is defined in a detailed and mathematical way. In order to make such investigations on fairness, it is crucial to understand the key perspectives of fairness, especially related with various difficulty aspects.

In this section, we will focus on the following two discriminations that have been often referred in the subjects related with the equal employment legislation of the United States, and other discussions about fairness.

(A) Disparate treatment discrimination

(B) Disparate impact discrimination

Disparate treatment discrimination refers to cases where there is some unfair treatment in the process, where fairness is intentionally compromised, and is considered a basic requirement in the employment process.

On the other hand, *Disparate impact discrimination* refers to cases where resulting fairness is compromised, and is sometimes regarded as a problem, such as in the adverse events to minorities. The definition of resulting fairness needs to be defined in detail using target indicators, such as *equal hiring result number*, *equal ratio of hiring to applicants* (between men and women). There are cases where deciding the right fairness index for the ethical requirement require quite a consideration. Furthermore, it is often not enough to simply *not include intentional discrimination in the process*; it may be necessary to intentionally introduce *discriminatory treatment* in order to eliminate discrimination as a result, and this may even involve an intentionally unequal process called affirmative action. From the perspective of engineering for society as a whole, this can be thought of as a situation where feedback control with overshooting is applied toward a target value that is considered to be correct, while from the perspective of AI system implementation, a specific output distribution is required as a

functional requirement based on ethical and fairness considerations.

Regarding the terms *equal opportunity* and *equal treatment* that are often referred to in Japan, there exist multiple levels and need to be clarified case by case. Taking *equal treatment* as an example, it can refer to *strong equality*, meaning that measures are taken to ensure equality (e.g., algorithms) as a system during the development stage, or it can refer to *weak equality*, meaning that intentional inequities are avoided during the development stage. Especially in the field of statistical machine learning, the latter type of equal treatment alone is often insufficient to ensure fairness due to various factors. In addition, some situations such as personnel evaluations, may even require a *justification* that is felt by the individuals concerned that are hardly quantitively defined.

(Reference) Group Fairness and Individual Fairness

In general, fairness demands are concerned with attributes that may cause inequity such as race and gender (attributes requiring consideration). Group fairness is to avoid discrimination among different groups with respect to some attribute value (e.g., disadvantageous treatment of *women* which is a value of the attribute *gender*), while individual fairness seeks for similar individuals to be treated similarly. As described in the following sections, most of the current general-purpose machine learning metrics and measures are based on the premise of group fairness that mainly addresses *attributes that require consideration, generally called sensitive attributes*, and the Guideline also assumes the group fairness perspective unless otherwise noted.

In the case where *justification from the individual person point of view(=individual fairness)* is required, such as personnel evaluations, it is difficult to define general metrics thus measures to satisfy the requirement must be considered for each system correspondingly. As for individual fairness, research on *degree of similarity* using distance learning[102] for example, have been proposed, and we hope to see more of this in the future.

8.3.2. Ambiguous social demands for Fairness

When the demand for fairness in the system is caused by the requirements of laws and regulations even if partially, the concretization level gap between the legal requirements and the technical implementation also becomes an issue. For example, when laws and regulations stipulate that *no disadvantageous treatment shall be given*, it is not always obvious what acts or omissions in automatic processing by machines and their design is considered as disadvantageous treatment. Furthermore, whether *equal treatment* demands on the process how people build AI or on the AI outcome can make a big difference in actual development work. As norms and best practices such as the Guideline become more widespread, it is expected that something like a market view of fairness measures will be formed eventually, but at the moment, it is strongly required to be able to properly explain *how you thought and how you implemented*

it.

In addition, the requirements of laws and regulations are not always self-evident as to what is the object to ensure fairness in other words, *what would be values for the sensitive attribute?* Sometimes it is explicit, such as equal opportunity for *men* and *women*, but more often it is an incomplete enumeration, such as *gender, race, etc.* Even in explicitly defined case, consideration and explanation of the implementation policy will be needed, if the attribute is not necessarily included in the input data, or if the attribute value is subject to estimation and can contain measurement errors.

8.3.3. Embedded inequities in society

When building machine learning components from data, it is important to consider the case where the data itself, such as training datasets, may contain inappropriate biases. Data obtained from the real world may sometimes inherit the existing biases contained in society.

Inequity embedded in society is one reason why a fair system cannot necessarily be created simply by eliminating sensitive attributes through the development process. Therefore, in developing a system that uses machine learning, where fairness is important, fairness requirements should be clarified not only from the deduction of the data itself, but also from a top-down analytical perspective, leading to scrutiny at the data preparation stage. Such activities will also lead to clarification of social demands for the system that are not always specifically stated, also improving its accountability.

> (Reference) COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) Case Study
>
> It was pointed out that the COMPAS system, which is used to estimate recidivism rates as a basis for making decisions on parole for imprisoned prisoners in the United States, may be making significantly unfavorable decisions for black prisoners, and that socially embedded biases such as bias in crime detection may be included in the training data. Whether the point was valid or not depended on how the fairness metrics were defined, which also suggested the importance of defining fairness metrics.

8.3.4. Hidden correlations and proxy variables

When building machine learning components from input data that has a huge number of attributes and information, other attributes that seem unrelated to discrimination, or features of the input that are not clearly identified as attributes, may have hidden correlations with sensitive attributes, resulting in statistical reproduction of discrimination. For example, such correlations between attributes such as name and gender, school name and gender, address and income, may result in unfair output even if the direct contribution of the sensitive attributes to the output is removed. It is also possible that unfair output may be produced as a result even if

the direct contribution of the attribute to the output is removed.

In such a case, even if you try to remove social inequities by data synthesis as described in the previous section, you may not be able to synthesize training data that is appropriate in the real world. For example, if the name of a school has characteristics such as boys' school and girls' school, adding data by artificially synthesizing data with only the gender of the input data reversed will not result in valid data.

8.3.5. Variables requiring careful consideration

When trying to eliminate unfairness in the AI development process, it is sometimes difficult to decide how to treat attribute values that have indirect correlation with sensitive attributes, even if they are correctly identified. For example, in a case such as a lending credit operation, the *amount repayable for the target customer* which is not measurable directly is an important attribute that we want to estimate. In this case, if a measurable attribute which may be an input variable to the system is clearly identified as discrimination cause in the estimation, such as gender, or if it is a variable that can be clearly identified as an input to the target estimation function, such as *loan amount*, the appropriate process would be rather obvious. On the other hand, for example, attributes such as *income* are closely correlated with *repayable amount* by nature, but may also reflect social disparity and discrimination. Therefore, there is a possibility that we may want to use them from the perspective of improving estimation performance, but from a fairness perspective, maybe shouldn't use it. In such a case, it is difficult to determine the strategy for ensuring fairness, and explanations are also important.

8.3.6. Attacks against AI in use

Lastly, let us note that the possible attacks against fair machine learning systems by users should be assumed. For example, in a system that performs continuous learning, an attacker may embed unfairness into the system by continuously feeding biased data, etc., or conversely, an attacker may gain by feeding adversarial data to a machine learning model that is distorted by the process of making it fair.

8.4. Basic approach to ensure fairness

In this chapter, we summarize the approach of fairness quality management in machine learning module that the Guideline address, based on the issues enumerated in the previous section.

8.4.1. Structural model for ensuring fairness

Generally, fairness requirements for products and services between users and service providers concern quality in use and are expressed rather conceptually and only qualitative requirements such as *not to be treated unfairly* or *to be treated equally* are described there. On the other hand, *fairness metrics* (Section 8.5.2.4.2.1), which are dealt with in the literature on machine learning AI techniques, quantify the degree of bias of various input datasets and output results focusing on certain characteristics in the stages of system development & operation. They are numerical index corresponding to the *internal quality index* in the Guideline.

Based on this perspective, the Guideline assumes the following process as a way to ensure *fairness* in machine learning systems.

- Handle qualitatively, high level requirements for *fairness* of usually *equal treatment* type, such as derived from clear social demands or found through *quality in use* discussion.
- Take the risk analysis based approach, which regards the compromise of equal treatment as a risk, to detail the fairness issue.
- As the work progresses, define quantitative fairness metrics on equal outcome (i.e, measurable indicator) if appropriate, and use those metrics to achieve the target.

This approach tries to ensure the appropriateness of metrics selection by means of analysis and design, similarly, to risk analysis-based approaches in functional safety area.

Figure 18 below illustrates a typical flow of this approach. This diagram is not meant to constrain individual development policies or phases, but to provide a model for shifting from qualitative considerations to quantitative methods as following.

(a) Fairness demands at the highest level of abstraction, usually demanded from perspectives such as *justice* or *human rights*, would be thought to derive from matters expressed in key words such as *equality* and *equal treatment*.
(b) From the viewpoint of social rules such as legal system and implicit ethical behavior, there are two cases where *equal treatment* is required, and where *equal outcome* are required in the form of numerical target, as shown in Section 8.3.1

In the latter case, it may mean a response assuming that the treatment is not equal in the actual institutional aspect. For example, in the context of equal employment opportunity for men and women, when there is a positive feedback structure of undesirable(negative) phenomenon, such as that the lack of women's advancement in society hindering the development of women with sufficient social experience, which in turn delays women's advancement in society, a positive feedback may be induced to positive phenomenon by forcibly promoting women’s advancement through numerical targets. In this case, since the achievement of numerical equality of outcomes becomes

the primary objective, all subsequent stages are also attributed to the achievement of numerical equality.

In addition, as mentioned before, *equal treatment* has two levels; a strong requirement to *actively design and implement the system to achieve equal treatment*, and a weak requirement to *avoid intentionally providing unequal treatment*, which is a non-binding, just an effort target. The Guideline mainly deals with the positive activities of the former and does not cover weak requirements at the level of mere effort targets.

(c) Regarding quality in use of the overall system design, and

(d) the external quality, which corresponds to the design of machine learning modules, there are two options in goal setting; numerical (measurable) equality of outcome and equality of treatment. At this stage, if the sensitive attributes to be considered are clear, and if we can envision how fairness with respect to those attributes will manifest itself in the results of the machine learning model, we can define fairness metrics, such as quantitative goals, that can be used in later steps.

(e) Next, in the preliminary preparation stage, where we construct(develop) the machine learning system that we want to train and consider some of the internal qualities (A-1 and A-2 as defined in the Guideline), we have the same two options for setting goals. At this stage, analysis may be conducted based on the actual data collected and other information, and goals may be set.

(f) Finally, in the stage of learning and checking internal quality, there are three possible options: analyzing the statistical distribution of results, monitoring statistical and analytical indicators other than the distribution of results, and explaining the equal treatment from the logical structure of the development.

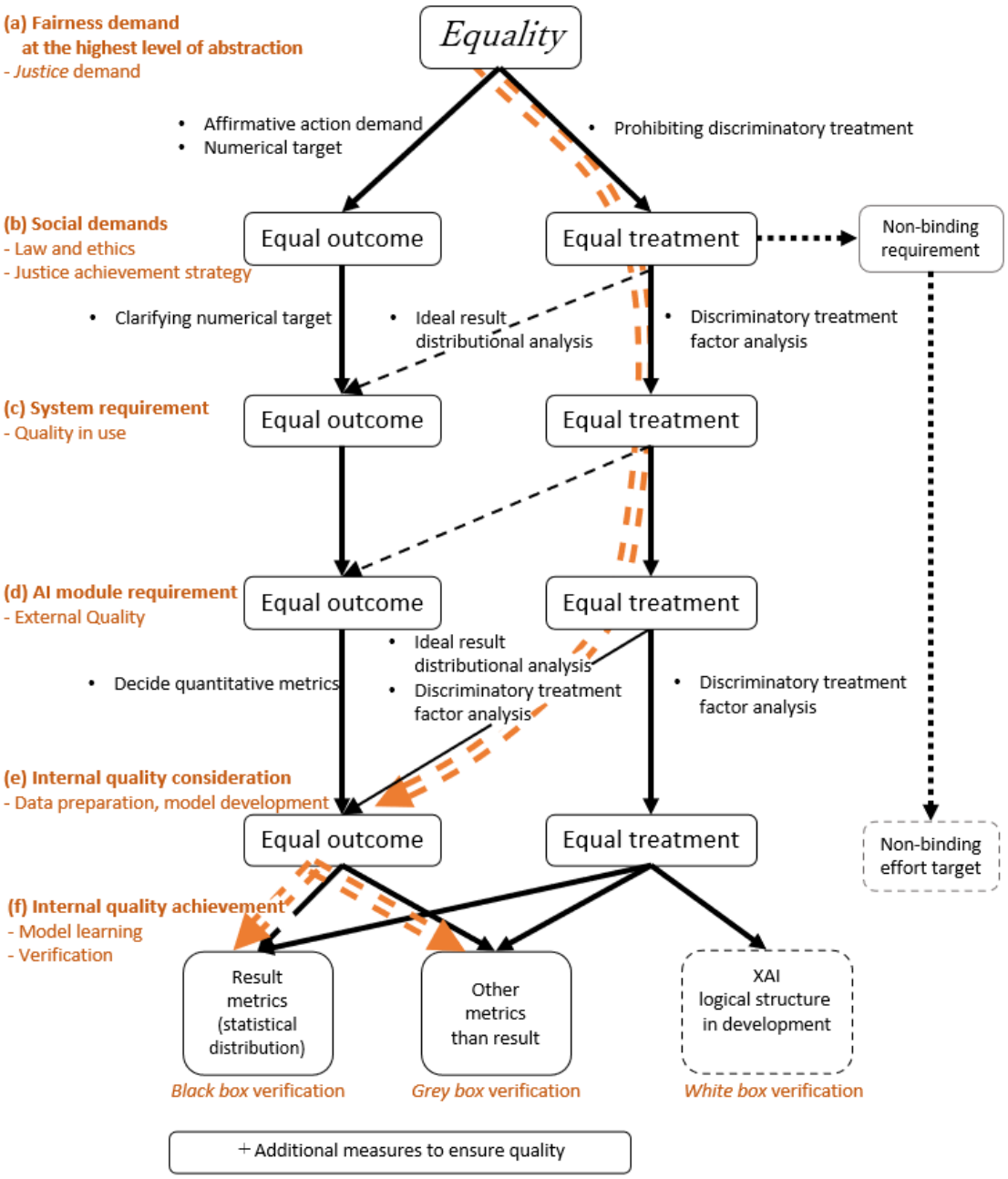

Figure 18: Example of a process structure for ensuring fairness quality

8.4.2. Basic approach to fairness assurance

In the remainder of this chapter, we assume the process shown by the dotted line in Figure 18 will be an approach that can be generally considered in basic fairness management in machine learning modules. Firstly, at the level of external quality, we discuss the fairness to be ensured among features and attributes in accordance with the definition of fairness as shown in

Section 1.5.3 and at the stage of internal quality consideration, we analyze data distribution, etc., and convert them into numerical requirements for bias in outcome, and others such as in dataset if applicable. Then, based on the results of the previous-step analysis, the requirement for bias in outcome (i.e., inequity regarding distribution of results) will be measured at the system development and testing stages, together with other metrics if defined, thus confirming *fairness* we are aiming for.

This is based on the following two considerations. (Point 1) the fairness requirements should be reflected in the distributions of training and test datasets, since machine learning implementations are derived from data. (Point 2) in order to sort out the complex problem structure discussed in Section 8.3, consideration based on actual data at model development stage is necessary, thus keeping the original abstract-level fairness requirements until that stage is important.

Of course, the development processes shown in the figure can be selected according to the situation of the functional requirements of the individual system. For example, in the case of an application where numerical targets are set in advance at stage (b), the implementation will be based on the functional requirements to achieve the numerical targets.

Note that the *avoidance of intentional unfair treatment as a non-binding target* described in the previous section is mapped to a universal goal for product services at the AIFL 0 level. (it is not equivalent to *best effort* as required with AIFL 1)

8.4.3. Considerations for handling data with sensitive attributes

Among personal information, there are *data with various sensitive attribute*, such as race, gender, place of origin, etc., that require careful handling and must avoid unfair treatment resulting from these values. In developing machine learning modules that require fairness, the careful handling of these sensitive attributes (data) is required from the early stages of the process shown in Figure 18.

The Guideline *does not take the position that it becomes automatically fair if it does not handle these sensitive data*. Data obtained from the real world have complex structures and correlations, and it is quite possible that AI built from other input data than sensitive data will eventually be found to have correlations with the sensitive data which are even not included in the input. Furthermore, not using the sensitive attributes in the development, especially in the testing process can be a major disadvantage because it makes it impossible to check and inspect the quality of fairness. This is because sensitive attributes are often the key points for ensuring fairness. Therefore, in order to build a fair machine learning system, we believe that such sensitive data should at least be obtained during development, and sufficient consideration should be given at the stage of system design and development process consideration.

On the other hand, whether or not sensitive data can be obtained during operation phase for the purpose of quality monitoring and additional learning needs to be considered in relation to personal information protection. In addition, there may be cases where it is impossible to handle

sensitive information that requires special consideration, such as race or medical history, even at the system development stage. In these cases, it is necessary to give sufficient consideration to how to eliminate bias and conduct monitoring in the model building and additional learning stages.

8.5. Quality management for fairness

8.5.1. Refinement of fairness requirements as preliminary preparation

If the required fairness target level is assumed to be AIFL 1 or higher, the points described below shall be considered in the requirements definition stage. If multiple factors are related to fairness, the level shall be determined in principle, corresponding to the most demanding factor.

8.5.1.1. Clarification of fairness goals

First of all, the goal of ensuring fairness for the target system should be clarified as follows.

1) Consider the attributes that form the basis for judgments about fairness.

A) Clarify the information such as sensitive attributes that needs to be taken into consideration for fairness, from the target function's point of view. This information is not limited to that which is explicitly identified as an attribute of the actual input data. For example, information such as gender, age, race, and ancestry could be identified as sensitive attributes even though not included in data.

B) If possible, clarify the observable information (e.g., written test scores, department of choice, etc.) that may be used for judgment with sufficient consideration of the influence from the sensitive attributes clarified above.

C) Furthermore, if possible, verbalize the essential characteristics such as potential academic ability, graduation potential, etc. that are the source of the observable attributes to be judged above and are not affected by the sensitive attributes.

Here, it should be noted that the results of the above study and subsequent efforts will vary depending on the scope of fairness that the target system or social activity that includes it should consider, or the assumptions of *indicators to be considered fair/unfair* given in the pre-condition.

(Example)

For example, consider a college admission examination system, and the scoring and pass/fail judgment process of the system is our target of discussion.

From the standpoint that solving the same prepared test questions at the same time shall give examinees a fair chance, the *validity of the answers to the test questions* becomes an essential (abstract level) index that should be used for the final judgment of pass/fail, and whether the *score result* might be an appropriate observable numerical index for it.

On the other hand, as in the case of college entrance exams discussed in the fairness related study[78], the examination system itself may have fairness problems. From their point of view, *potential academic ability* and *possibility of graduation* may be essential ideal criteria for the judgement. In addition, the *score result* would be affected by the disadvantages derived from the poor past educational opportunities, might be strongly correlated racial, income data. They may argue that some corrections considering those aspects might be necessary for a fair pass/fail decision close to the ideal criteria.

2) Clarify the basic handling policy for the sensitive data.

A) Do the sensitive data require that the system not be treated in a discriminatory manner because of the information (equal treatment), or does it require that the results should follow a certain predefined distribution with the specific numerical targets (equal outcome)?

Are there any legal requirements regarding the handling of the information in question, such as those that exist in the areas of equal employment opportunity, equal race, etc.?

B) If the discrimination related to the sensitive data already exists in the real world, how should the AI, which is built based on the data collected from the real world, handle it? Should such discrimination be removed as much as possible, or can a certain amount of residual discrimination be tolerated, or should be reproduced also within AI output based on the system requirements?

C) Clarify the constraints on the available measures to remove residual discriminatory judgments in machine learning results. Is it permissible to introduce intentional corrections to the training data or to intentionally adjust the trained model? It should also be clarified how the decision to stop development will be made if the problem cannot be corrected by the available means.

After making the above decisions, the fairness requirement level (AIFL) is checked again.

8.5.1.2. Modeling dependencies and causal relationships among data attributes

Since data acquired from the real world have complex structures and correlations, machine learning without properly understanding the dependencies and causal relationships between attributes may cause inappropriate for the purpose or inaccurate results.

As we plan to implement the system using machine learning with data rather than logical description by program description, the dependencies and causal relationships among attributes are not completely known (if they were known, the non-AI program could have been written), and it is difficult to grasp them completely even after exhaustive analysis. However, it is very important to investigate especially the following aspects as far in advance as possible.

1) Information paths that may cause unfairness

In the process by which a sensitive attribute unfairly affects the output of a system, the sensitive attribute may directly affect the result, or there may also be contributions from other attributes that are affected by the sensitive attribute, or even further indirect influence through those attributes. When these series of causal *paths are not understood*, it is often the case that the final fairness requirement cannot be met nor explained, even removing the direct impact.

In addition, when there are attributes correlated both the attributes; that should contribute to the system decision and that should not, a typical pre-processing such as eliminating bias in the dataset regarding the sensitive attribute can have a negative impact on the AI system decision performance, thus requiring careful consideration.

2) Simpson’s Paradox.

Simpson's paradox refers to the phenomenon that when analyzing to find a relationship between two data attributes ignoring the third factor that is actually involved in both of them, a correlation that does not exist in reality or a correlation that is the opposite of reality may be shown [158]. This paradox was proposed over 50 years ago in the field of statistics, but it also applies to machine learning today. In order to prevent such a phenomenon, it is necessary to understand the relevant third attribute in advance and to prepare learning data with *case separation* for the third attribute. We will discuss specific examples later.

3) Necessity of using sensitive attributes

There can be cases where fairness cannot be achieved without daring to use sensitive

attributes. In other words, *fairness by unawareness* may sometimes be insufficient for the objective.

For the support of analysis described above, a graphical method that can represent these interrelationships in a relatively concise manner, called Causal Bayesian Networks (CBNs) [78] will be presented here. CBNs are graphical representations that exhibits the causal path between attributes, helping us to consider the patterns of unfairness underlying the training data. It is also useful for examining the types of fairness metrics that should be used.

(Example.)

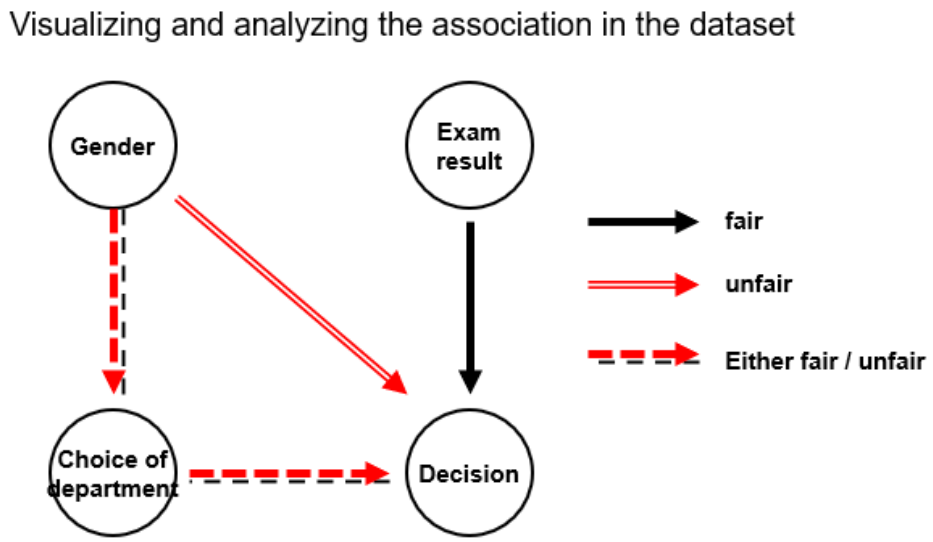

Figure 19: Example of association visualization using CBN (from[78])

Figure 19 shows a college admission scenario in which applicants are admitted based on qualification exam result, choice of department, and gender discussed in [78], etc. CBN visualizes the relationship between attributes by representing a direct influence of α on β is represented by an arrow from α → β (causal path), which is to be considered to see whether their influence is fair or unfair, or "either is possible".

In this example, the causal path from exam results to pass/fail decision is certainly fair. If gender is used directly in the pass/fail decision, i.e., if the two individuals with same exam results can be treated differently just depending on their gender, there exist the causal path gender → decision, which is certainly unfair.

The path from gender to choice of department could be either as fair or as unfair. If the fact that women tend to prefer certain department is due to implicit environmental pressures on women, then it is unfair causal path. Similarly, the path from choice of department to decision could either be considered as fair or as unfair. If the admission rates for departments chosen more often by women would be intentionally lowered, then the path would be unfair.

Thus, by carefully analyzing the paths that may influence pass/fail decisions, we can see that the path gender- choice of department- decision can be a factor that impairs fairness, depending on the actual situation and algorithm. This means that the fairness of the decision result cannot be guaranteed by simply excluding the direct causal path from gender to decision.

Figure 20 is an example of the Simpson's Paradox in CBNs. For the purpose of finding correlation between the attribute exercise days per week and the attribute health (assuming, some health parameter), we may not be able to achieve our goal if we train our model using dataset collected without considering the third factor age, which affects both attributes.

For simplicity's sake, let's assume that age is positively correlated with exercise day per week and negatively correlated with health. In this case, even if exercise and health are positively correlated actually, the opposite result may occur. The third factor, such as age in this case, is referred to as confounding factors and should be carefully handled in training dataset preparation, not only from a fairness perspective. Drawing CBNs encourages the discovery of such confounders, and if an attribute of concern is a confounder, it can be recognized in advance that simply removing it may not be appropriate.

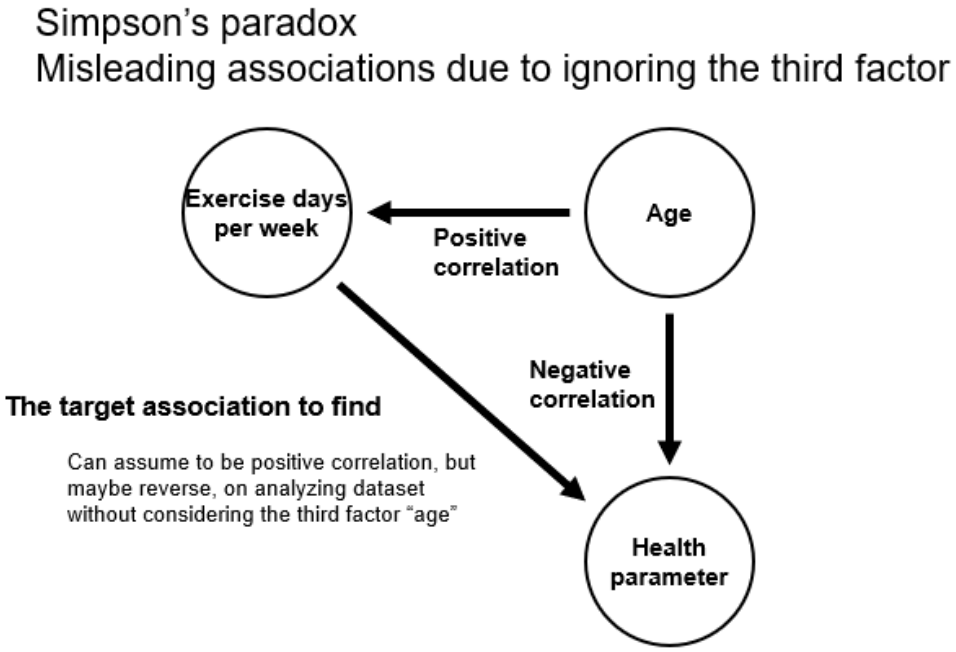

Figure 20: CBN example of Simpson's paradox

8.5.1.3. Different perspective from AI development

The detailed fairness requirements described above as a preliminary preparation may be hardly achieved by only AI engineers, so the participation of domain experts would be desirable. Alternatively, the following approach from a different perspective from AI development would be expected to have a certain effect.

Correlation Analysis (Statistical Method)

As shown in Figure 18: Example of process structure for ensuring fairness quality, the Guideline recommends that the detailed study based on actual data should be conducted in the model "development" stage, and that the requirement analysis stage, should be conducted top-down , without going deeply into the implementation aspect with the detail analysis on actual data.

However, if domain knowledge is insufficient, such a top-down examination might be difficult, thus analysis such as CBN based one, might be impossible at all.

To overcome such difficulties, the following steps might be some realistic options:

1) Draw a network based on the correlations that exist in the actual data;
2) For the drawn network, carefully determine the nature of each correlation (e.g., a fair causal path or an unfair causal path).

However, it should be noted that the fairness requirements should be essentially analyzed from the higher-level requirements and may not necessarily be directly expressed by the attributes in the actual data.

Social and Economic Approach

Fairness and equality in society have long been studied in welfare economics, including the *Social Welfare function (SWF)*. Recently, SWF utilization has been proposed to optimize interventions to ensure fair causality[115].

We expect that approaches from perspectives other than AI engineering, such as those described above, will provide some insight that could resolve the difficulties in analyzing *fairness* requirement.

8.5.2. Fulfillment of fairness requirements

8.5.2.1. Policy outline

This section provides an overview of the actions for meeting the fairness requirements. Depending on the scope of the techniques, they can be broadly categorized as follows.

A) Changing/Adjusting training datasets
B) Techniques on learning algorithm
C) Adjustment to outcome (trained models)

D) Correction in use

These also correspond to the *three categories of fairness techniques by development process* [129][217][150] as follows.

A): Pre-processing approach

B): in-processing approach

C) and D): Post-processing approach

Depending on developers' ability to intervene in the process of the target machine learning module, the choice among measures would be made as follows.

- If the developer can modify training dataset, then pre-processing is worth trying because it allows for relatively straightforward explanations such as *we trained on a dataset with specific metrics and guaranteed fairness* and is generally quite effective.
- If the developer can design model/learning algorithm, then in-processing approach can be used.
- If the developer can only receive a trained model which is basically as black box and can perform some adjustment to it, then only post-processing would be the choice.

Taking into account the target level of fairness, the basic policy of actions is approximately as shown in the following table.

Table 5: Summary of work scope and actions by AIFL

	When the training dataset can be modified	When the design model and learning algorithm can be modified	When only post-processing of trained models is possible
AIFL 2	Define and record fairness metrics *goals* for the training dataset, with multiple pre-processing techniques, if necessary, to improve and ensure they are met as much as possible.	Add fairness metrics for model output to the learning objectives and balance them with other metrics such as AI performance, and to achieve the target using in-processing techniques.	N/A (Consider in advance about the measures that can be taken to adjust the trained model. If the essential conditions are unlikely to be met, the work scope needs to be reviewed.)
AIFL 1	Define and measure fairness metrics *goals* for the training dataset. If there are deviations, pre-processing techniques are used to improve and recorded with the result.	Fairness metrics for model output are added to the learning goals, and if the results of training with AI performance priority deviate from the goals, improve it by in-processing techniques and recorded with the result.	Define fairness metrics for model output and measure them during testing. If there is a deviation from the target, try to improve it by adjusting the trained model, or any other adjustment in the operation.
AIFL 0	Define, measure and record fairness metrics for training datasets.	Define fairness metrics for model outputs and record their measurement during training.	Define fairness metrics for model outputs and record the measurements during testing.

The following sections describe the details of the techniques, mainly in line with the internal quality set forth in Section 1.7 and Chapter 6.

8.5.2.2. A-1: Sufficiency of problem domain analysis (preliminary preparation)

By following the guide presented in the preliminary preparation (Section 8.5.1), the following aspects can be investigated.

- Dependency analysis between attributes
- Clarification of requirements for training dataset distribution
- Metrics to be checked (fairness metrics)

In addition, considering the nature of the overall system to be developed, the requirements of the following perspectives should be sufficiently confirmed.

– Consistency with legal and social requirements

Since the development team may not have sufficient knowledge of legal and social requirements, it is advisable to consult experts in the field.

The required actions for each target AIFL level are as follows in the preliminary preparation phase.

- AIFL 1
  - Define the fairness requirements and record them, including the history.
  - Based on this requirement, requirements for data preparation such as fairness metrics for training datasets shall be defined.
  - Based on this requirement, fairness metrics for model outputs shall be defined.
- AIFL 2
  - In addition to those listed in AIFL 1, the following actions shall be taken
  - To model dependencies and causal relationships between data attributes.
  - The results of modeling should be reflected in the fairness metrics of the training dataset.

The *fairness metrics* for the training dataset are basically from the same perspective as those for the model output. For example, if the fairness requirement is *no difference in pass or fail judgments based on gender* for the output, the training dataset should also be free of bias from that perspective. Examples of metrics are described in Section 8.5.2.4.2.1

8.5.2.3. B-1 to 3: Coverage/uniformity/validity of datasets (data preparation)

In the data preparation stage, *pre-processing* techniques are taken to ensure required fairness, mainly from the internal quality perspective of *coverage*, *uniformity*, and *validity* of datasets. What is commonly referred to as Data Fairness is achieved at this stage, and there are two types to address it; one for data collection process and the other for the data collected.

8.5.2.3.1. Data collection process without introducing bias

From the perspective of preventing disparate treatment, first of all, we must avoid *bias in data collection operation*, such as sample size disparity and tainted examples[11], which makes datasets become biased against the real world.

Actions here are based on the perspective of *coverage* and *uniformity* of datasets in the sense of confirming whether there are enough data for the attribute combinations of interest, or whether the data are close to the distribution in the real world. Techniques such as weighting the collected data by groups (Reweighing [111]) are used for the purpose.

8.5.2.3.2. Data adjustment, enhancement, and synthesis

If we want to prevent disparate impact, even if the dataset corrected is not biased against reality, we may need to take measures against the distortions and discriminatory factors inherent in the dataset from the real world, corresponding to *validity of datasets* point of view.

Examples of major techniques include the following

- Disparate Impact Remover

  To remove the bias caused by the so-called *proxy*, the attribute value that may be correlated to the sensitive attribute is modified to reduce the possibility of indirectly inferring the sensitive attribute from that proxy attribute.

- Optimized Pre-Processing

  A framework for pipelined transformations to pre-process training data. Consideration is given to both the removal of bias and the usefulness of the data which are not too far from the real data [68].

8.5.2.3.3. Required actions for each quality level in data preparation

Required actions for each target AIFL level are as follows in dataset preparation phase.

- AIFL 1
  - Measure and record fairness metrics for datasets.

[11] Sample size disparity: The number of data points varies greatly depending on the value of the attribute requiring consideration.
Tainted Sample: Bias caused by human labeling

- When the dataset preparation is within the work scope, if the metrics deviate from the target, at least *different treatment* measures described in Section 8.5.2.3.1 shall be taken to improve the metrics.

  –

- AIFL 2
  - In addition to those listed in AIFL 1, the following actions shall be taken
  - When the dataset preparation is within the work scope, if the metrics deviate from the target, *different impact* measures described in Section 8.5.2.3.2 shall be taken.
  - When dataset preparation is outside the work scope, consider measures to minimize the impact of the dataset metrics deviation from the target on learned model output later in the process. In some cases, collateral outside of the machine learning component should be considered in advance.

(Note) Some of the pre-processing methods described herein are available on a development platform and tool as described in Section 8.6. However, in almost all the cases there, the input data are assumed to be a series of *attributes and values* (structured data) and *sensitive attributes* are already identified within the data. It is sometimes difficult to use those methods directly with machine learning systems that process non-structured data, for example, natural language processing that takes a large amount of text (corpus) as input. As for fairness in natural language processing, Mehrabi et al. [129] have introduced bias removal methods in language modeling and feature learning methods but further research is still needed to make these methods widely available in the field. At present, a realistic approach is to ensure fairness by testing and adjusting the resulting machine learning models in post-processing.

8.5.2.4. C-1 to 2: Correctness and stability of trained model (training and testing)

In training with the dataset, *in-processing* and *post-processing* techniques are taken to ensure fairness from the internal quality *accuracy of the model* and *stability of the model* perspectives.

8.5.2.4.1. Countermeasures (in-processing) in model development and learning phase

In-processing measures are those that aim to embed fairness in the model and can mainly be implemented by modifying the objective function with additional constraints or adding model

elements, as described below.

8.5.2.4.1.1. Example of additional constraint technique

- Prejudice Remover

  The resultant probability distribution is adjusted by adding a constraint to remove indirect prejudice[12] that remains even if the sensitive attributes are not used directly. For this purpose, the constraints are implemented as a regularizer using the sensitive attributes and added to the objective function.

  This method is highly explanatory, although its applicability is limited to identification problems [112].

8.5.2.4.1.2. Example of adding model elements technique

- Adversarial Debiasing

  This technique adds an adversarial model that uses the inference results from the target model as input to infer the sensitive attribute. The target model should be trained so that the inference of that adversarial model fails as much as possible.

  Adversarial debiasing allows for a wide variety of target model algorithm and may handle the problem of insufficient proxy analysis in data preparation phase. [206].

8.5.2.4.2. Adjustment to trained models

Among the measures classified as post-processing techniques, this section discusses *adjustments to the trained model*, from model accuracy and model stability aspects, while in the next section, *adjustments during operation assuming that discrimination remains* will be discussed.

Depending on the required fairness goal metrics for the trained model, there are adjustments such as the Equalized Odds Framework and calibration [160].

Of course, quantitative evaluation of target metrics is necessary during learning as an internal quality target, and it is essential to define and measure them appropriately.

8.5.2.4.2.1. Example of target metrics

All of the metrics are positioned as indicators to confirm that there is no *different impact*

[12] One source of bias. The statistical dependence of a target variable or a non-sensitive attribute on a sensitive attribute.

discrimination. They are selected according to the concept of discrimination concerned for the system. Some commonly used fairness metrics are listed below. For others, please refer to the literature [129].

Equalized odds	Make the *right decision* without relying on a sensitive attribute. ($TP/(TP+FN)$ and $TN/(FP+TN)$) are equal.
Predictive parity	Regardless of a sensitive attribute, the *precision rate* ($TP/(TP+FP)$) are equal.
Demographic parity	The percentage of *Desired outcome* is equal regardless of a sensitive attribute. $(TP+FP)/(TP+FP+TN+FN)$ Desired outcome (equal percentage)

8.5.2.4.3. Required actions for each quality level in model training

Required actions for each target AIFL level are as follows in model training phase.

- AIFL 1
  - Measure and record the model output metrics.
  - If there is a deviation from the target, at least one in/post-processing technique shall be used to make improvements, and if there is still a deviation, record it.
- AIFL 2
  - In addition to those listed in AIFL 1, the following actions shall be taken
  - If there is a deviation from the target, try multiple in/post methods to make improvement as much as possible.
  - If there is still a discrepancy that cannot be resolved, consider combining additional measures in actual operation, sometimes extending to outside the machine learning module.

8.5.2.5. Adjustment and usage during operation

Even with the measures described so far, the bias may not be fully removed. To deal with such cases, there are measures to adjust the learned model in some way when actually performing inference. The following is one example of such methods that has been proposed. Of course, other adjustments can be made to the system as a whole, including outside the machine learning components, to suit the situation.

- Reject Option-based Classification (ROC)

  Based on the hypothesis that discrimination occurs when the *confidence level* of the model inference is low, ROC rejects or revises the results of inferences that may lead to discrimination. Specifically, if the *favorable outcome* of the *favored group* or the *unfavorable outcome* of the *non-favored group* comes out lower than the specified confidence level, the result is rewritten [110].

8.5.2.6. Maintainability of qualities in operation (E-1)

Products and services for which fairness is important are often used in public spaces and other open environments, and therefore, sufficient consideration must be given to changes in quality in operation, such as concept drift. It is necessary to measure the fairness metrics for model outputs and datasets at appropriate intervals, regardless of the target AIFL level.

8.6. Development infrastructure and tools for fairness

8.6.1. Purpose of using development platforms and tools

The work described in Section 8.5 performed in the machine learning lifecycle, along with other quality goals than fairness. Fairness often requires a trial-and-error approach, including the identification of biases inherent in datasets and the measurement of fairness metrics for trained models. For this reason, the use of development platforms and tools is quite beneficial for the followings:

- Efficiently consider the perspectives described in Section 8.5.2 for the dataset.
- Perform some kind of visualization, metrics evaluation, or other fairness assessment of the trained model.
- Maintain appropriate records of work, the datasets used in the process, and the sequence of work.

8.6.2. Examples for applicable tools

8.6.2.1. Typical functions and modules

The main functional groups that are desired for the platforms and tools are listed below, based on the purpose described in the previous section.

(a) Visualization of datasets for analysis to discover relationships regarding sensitive attributes
(b) Discovering fairness matters other than those clarified with metrics for counterfactual experiments to test the impact of changes in certain attribute values on output decisions
(c) Various fairness metrics measurement
(d) XAI Library for visualization of which attributes contributed to the decision
(e) Infrastructure for building pipelines to support the lifecycle of continuous monitoring and relearning after operation, so-called DevOps infrastructure + infrastructure for datasets and models)

Google and IBM Fairness360 are described below for reference, but new features from various other vendors will continue to be provided, so please check the latest information before making any selection.

8.6.2.2. Example 1: Google tools

Functions (a) through (e) described in the previous section will be covered as follows.

- What-if-tool (a), (b)
  - https://pair-code.github.io/what-if-tool/learn/tutorials/walkthrough/
- Fairness Indicators (c)
- Explainable AI (d)
  - https://cloud.google.com/explainable-ai/
- AI-Platform (e)
  - https://cloud.google.com/ai-platform

The What-If-tool, which provides a variety of functions directly related to fairness, is not just a visualization of the training dataset, but also helps to discover potential bias patterns in the dataset through hypothetical scenarios and analysis functions when sliced by various attributes.

Explainable AI can also contribute to confirming the bias, as it can quantitatively indicate which attributes contributed to the decision during inference. (Note that, as mentioned in 8.2having influence does not necessarily mean unfairness.)

8.6.2.3. Example 2: IBM tools

Functions (a) through (e) described in the previous section will be covered as follows.

- AI Fairness 360/IBM Watson OpenScale … (a), (b), (c)
  - https://aif360.mybluemix.net/

  - https://www.ibm.com/jp-ja/cloud/watson-openscale
- AI Explainability 360/IBM Watson OpenScale … (d)
  - https://aix360.mybluemix.net/
  - https://www.ibm.com/jp-ja/products/cloud-pak-for-data
- IBM Cloud Pak for Data … (e)

AI Fairness 360 and AI Explainability 360 are open source software developed and released by IBM Research Foundation and donated to the Linux Foundation in 2020. AI Fairness 360 provides a rich set of features to address biases in machine learning models and datasets, and generally covers the methodologies described in the Guideline.

- Pre-processing: Supports reweighing of training data, disparate impact remover, optimized preprocessing of training data, etc.
- In-processing: supports prejudice remover, adversarial debiasing, etc.
- Post-Processing: Equalized odds and calibrated equalized odds are supported.
- Metrics: Predictive parity and others are supported; Equalized odds and demographic parity can be calculated using the provided modules.

IBM Cloud Pak for Data is a multi-cloud software that covers the lifecycle of machine learning models in an integrated manner, and IBM Watson OpenScale is one of the software modules that make it up. The ability to monitor input and output data during model operation and detect time-series changes in fairness metrics is useful for improving biases that occur after operation begins.

9. Privacy

This chapter first summarizes social backgrounds in which quality management issues for privacy are considered (Section 9.1), then presents some quality characteristics specific to the privacy in machine learning (Section 9.2), and finally describes issues to be considered in the quality management (Section 9.3).

9.1. Privacy protection

This section summarizes the social context in which the notion of privacy in information systems, the right to control personal information, is considered.

9.1.1. Privacy risk

9.1.1.1. Ethical AI

AI technology, powered with data, enables realization of advanced functions and wide spread use of them. Such application functions may include services that have much impact on the right to privacy, depending on how they are used. For example, remote facial recognition in public places, profiling, and micro-targeting, if used outside the user's control, increase the danger of privacy violation. These observations have led to regulations that restrict the usage of socially unacceptable services and impose legal obligations on service developers and providers so that they protect the right to privacy. The legal issues have been discussed together with the notion of Ethical AI [23][30].

Ethical AI aims to mitigate threats to human safety and includes concerns about threats to privacy of users in general [79]. The discussions in Europe lead to regulations which enumerate specific services that may violate privacy rights and restrict high-risk AI systems or AI systems being used that pose the unacceptable harms[25].

In developing information systems, including systems to use AI technology, some measures on the protection of information and users' personal data are mandatory [31][26]. Personal data should not be identifiable from available information in places where they are outside the user's control. Unrestricted sharing and distribution of personal data bring about threats to human safety. It should be ensured throughout the system lifecycle, from development to operation, that personal data are appropriately protected.

9.1.1.2. Negative externalities

The views on personal data protection have changed over time. In general, information has

positive externality: the greater the amount of information, the higher new values [90]. Until around 1980, restricting data from public use was considered as economical inefficiencies, sometimes resulting in double investment, so that appropriate data sharing mechanisms were pursued. In the mid-1990s, with the development of information technology, it became possible to share large digital data or databases. The benefits of secondary use were emphasized, and digital data distribution was actively promoted.

It is often thought that data protection problems do not arise if personal information that would allow identification of individuals is pre-processed to be removed from public databases that are shared and distributed. However, it has been shown, in a public medical database, that re-identification of personal information is possible from the pre-processed and de-identified medical records [183]. Moreover, if there is public information available externally, the auxiliary information or background knowledge in it can be used in the re-identification attacks: the greater the amount of information, the greater the danger of personal data re-identification, which is called negative information externality[48].

From the viewpoint of service development and provision, the negative externalities directly link to adversarial impacts on business. Indeed, in the 21st century, as Internet services become more active than before, the threat of personal data re-identification has increased, and countermeasures against this have become a major technical challenge[100] [146]. In the future, AI technology can serve as a basis for those advanced services to use personal data that might include sensitive information. It is necessary to understand the dangers posed by the negative externalities, and to establish technical and institutional frameworks to mitigate such business risks.

9.1.1.3. Personal data

Personal data represent information that identifies an individual. As information technology has advanced, the context for discussing protection issues has been broadened, and the types of personal data have been varied. They can be generally classified into registered data and activity data.

9.1.1.3.1. Registered data

Registered data represent information that directly identifies an individual. Items are covered in Census such as name, date of birth, gender, race, address [219], and sensitive attributes, such as SSN or medical records. Some of them are explicitly listed as the subject of laws and regulations; they include specific sensitive personal information found in JIS Q 15001, Personal Information Protection Management Systems, Japan or sensitive personal information defined by Revised Act on the Protection of Personal Information, Japan.

9.1.1.3.2. Activity data

Activity data include not only information resulting from real-world activities, but also

information related to Internet service use, online activities. The latter has become understood as personal data as the Internet services grow. Typical examples of activity data include search history when using search engines, *like* button information in SNS, e-commerce purchase history, on-demand video viewing history, and cookies. Some may also be explicitly subject to legal and regulatory requirements [26]. With the development of information technology, including AI, activity data are considered to become even more diverse in the future.

9.1.2. **Rights of the data subject**

9.1.2.1. Data subject

The right to privacy is understood as one of the fundamental human rights and is given a legal basis as a right of natural persons. On the other hand, it is not an absolute right, but rather follows the principle of proportionality. Regulations refer to a balance between positive and negative aspects. Natural persons associated with the personal data discussed in the previous section are called data subjects.

9.1.2.2. Control by data subject

The rights of data subjects involve a legal basis. Due to different national basis laws, the rights of data subjects may not be the same. This chapter provides an overview of data subject rights, focusing on the European general data protection regulation (GDPR) [26]. Since its entry into force in 2018, legislations similar to the GDPR have been discussed in many nations, and global markets are being formed in line with the GDPR's approach. This chapter focuses on the GDPR as a representative example of legislation.

Consent: The provision of personal data is subject to consent by the data subject. There are two general approaches: opt-in, which requires prior consent for the provision of personal data, and opt-out, which deems the provision of personal data to be consented to unless there is an explicit refusal to provide it. And only free consent is valid [Article 7 of the GDPR].

In relation to the consent that has legally basis, the following points should be noted [Article 5.1 of the GDPR].

Purpose limitation: Personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes. Further processing is considered incompatible.
Data minimization: Personal data shall be adequate, relevant, and limited to what is necessary in relation to the purposes for which they are processed.

Storage limitation: Personal data shall be kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed.

Note that, as will be discussed later, these bring about major challenges in the context of data protection related to AI (Section 9.2.1.1). The importance of control by the data subject is also discussed as below.

Right to rectification and erasure: Data subjects have the right to obtain from the controller without undue delay the rectification of inaccurate personal data concerning him or her [Article 16 GDPR] and the right to obtain from the controller the erasure of personal data concerning him or her without undue delay [Article 17 GDPR]. The right to erasure is referred to as the right to be forgotten.

9.1.3. Protected processed data

Threats to personal data protection arise from re-identification. If personal data were to be released in their unprocessed form, they would clearly not be protected. Whether or not re-identification is possible is dependent on the method of data protection pre-processing [146][161]. The GDPR considers multiple methods of data pre-processing and requires that appropriate institutional measures to achieve required level of protection be put in place, depending on the strength of the data pre-processing method (Section9.1.3.2).

9.1.3.1. Spectrum of data protection pre-processing

The terminology used for protection pre-processing methods differs in divergent research fields. In this chapter, we avoid preconceptions given by the wording of the terms, and divide them into four levels, from the protection processing level 0 to protection processing level 3, according to the protection strength. Below, the subject matter is equally applicable to both registered and activity data.

Protection processing level 0: Unprocessed or unmodified data. If the data concerned are personal data, technical or institutional measures should be taken to achieve the expected data protection level. For example, mechanisms should be used to prevent the data from being leaked to the outside world.

Protection processing level 1: Data obtained by pre-processing unmodified data. They include pseudonymous data. Those data can be reconstructed by inverse transformations. In other words, data can be re-identified once the function responsible for the pre-processing or its inverse function is obtained. It is mandatory to take technical and institutional measures to achieve the expected data protection level. For example, information related to the pre-processing process should be properly managed to avoid the application of the inverse

transformation process, and a mechanism to prevent the leakage of such information to the outside world should be used together.

Protection processing level 2: De-identified data in which unmodified data are pre-processed to make re-identification difficult. While the level 1 allows direct de-identification, the level 2 often requires a trial-and-error method of indirect de-identification while making use of auxiliary information or background knowledge. Differences in the difficulty of indirect de-identification affect the technical or institutional measures to be taken.

Protection processing level 3: Anonymous data are those pre-processed in such a way that the inverse conversion is proven to be mathematically impossible in principle. Anonymous data are defined as those that are irreversible. Note that at the current level of technology, there is no generic transformation method that guarantees the irreversibility.

9.1.3.2. (Supplementary) Views under existing Regulations

The classification of the levels of protection pre-processing is based on technical methods (Section9.1.3.1). Below is an informal overview of the approaches taken under existing regulations on data protection.

9.1.3.2.1. Binary classifications

The binary classification approach was adopted until the mid-1990s, when benefits of secondary use, or publicly sharing, of digital data were primary, in which data are divided into two categories: unmodified data of the protection processing level 0 and the protection processing level 1 or beyond. Unmodified data, if referring to personal information, are subject to regulations, which requires that technical or institutional measures be taken to protect them. On the other hand, pre-processed data, of the level 1 or beyond, are collectively referred to as anonymized data and attempts to re-construct or re-identify these are legally prohibited. However, it is argued that the expected protection cannot be achieved only by imposing penalties on re-construction or re-identification as a violation of the law. [161].

It should be noted that these anonymized data are different from anonymous data characterized by the irreversibility (the protection processing level 3).

9.1.3.2.2. Ease of re-identification

The basic idea of the GDPR is to regulate unmodified data of the protection processing level 0 and pseudonymized data of the protection processing level 1. Anonymous data of the protection processing level 3 is out of its scope and not subject to GDPR, thanks to the irreversibility.

In addition, Article 11 and the preface of the GDPR [26] refer to de-identified data of the protection processing level 2 equivalent, which are not subject to the regulation if it is demonstrated that *they cannot be re-identified by the technology available at the time, at a*

reasonable cost and time. If it is not subject to the GDPR regulation, there is an incentive to adopt the devised methods of protection by de-identification, because technical or institutional measures are no longer needed [101].

On the other hand, with technological developments in re-identification methods, it may change whether data by means of particular de-identification methods are subject to the GDPR regulation. Traditionally, census aggregate information, for example, has been considered as statistically processed information equivalent to the protection processing level 2. In other words, statistically processed aggregate information have been regarded as data that are difficult to re-identify. In addition, census aggregate information must be appropriately protected, which is often stated as legal requirements in many nations. For example, in the United States, the law requires that the difficulty of re-identification of census results be guaranteed. However, it was recently pointed out that micro-data or unprocessed data can be re-identified even if conventional protective processing methods are applied [96]. Thus, the application of the method of protection by differential privacy (Section 9.1.4.2) is being considered for the 2020 census [220].

9.1.4. **Technical privacy metrics**

The data protection method is accompanied with technical privacy metrics [192]. They discuss the strength of data protection using quantitative metrics. Two representative ideas are introduced below. In this section, the target data with multiple attributes are called records, and a collection of many records is called a database, following the terminology used in the literature in this research field.

9.1.4.1. Data similarity

A method that focuses on records. A record is protected if the record is sufficiently similar to other records and cannot be separated from them. When a protected record can be linked to a data subject via particular sensitive attributes, it is sufficient to introduce multiple records such that the quasi-identifiers derived from the sensitive attributes are to have the same value. When the original values are distinct, there is a method to obtain similar records by replacing them with a kind of abstract values through generalization, aggregation, etc., considering the meaning of the attributes. For example, the attribute of birth date is turned to be represented only by the information of year and month, discarding the day. Last, K-anonymity is a systematic approach to obtain K or more similar records [168][184].

9.1.4.2. Outcome indistinguishability

A method that focuses on database query results. Consider two adjacent database D and D’

where D' is constructed as a copy of D but a particular record r removed. A record, r in this case, is protected if the results of queries to both databases are sufficiently indistinguishable; the presence or absence of the record does not affect the results, and thus the record is protected.

The ε-differential privacy (ε-DP) is based on the indistinguishability, and is a theory of provable privacy to guarantee that a given measure ε is an upper bound on the protection strength [85]. The protection is also persistent during a posteriori processing; the same differential privacy strength εis guaranteed when ε-DP protected data are processed afterwards. Furthermore, the composition theorem allows to determine the worst-case protection strength for a combination of multiple queries.

After the proposal of the ε-DP, (ε,δ)-differential privacy was introduced as an approximate differential privacy with relaxed applicability conditions [86]. Subsequently, further research has been conducted on improving the estimated privacy measure required to achieve the expected level of protection strength for multiple queries [133].

Differential privacy protects records as data in a database from data analysts or database users. Local differential privacy (LDP), on the other hand, protects data provided by data subjects from data processors or database creators [195]. A good example of LDP is seen in the process of data transfer from a client to a server via the Internet, in which the client intends to protect his or her own data[89].

9.1.5. **Data protection impact assessment**

Personal data protection is a problem on reducing threats of re-identifying pre-processed data: from the beginning of the development of information systems, including AI, data protection processing techniques are introduced according to the principle of *privacy by design* and the assessment is conducted afterwards. The GDPR requires the implementation of data protection impact assessment (DPIA) [Article 35 of the GDPR], and other guidelines also discuss the need for similar privacy risk assessments [36][218].

9.2. **Machine learning and personal data protection**

This section presents the notion of personal data protection in machine learning systems.

9.2.1. **Life cycle and protected Information**

First gives an overall picture of the lifecycle of machine learning components, from development to operation, as well as how artifacts are delivered or reused, and then introduces personal data protection issues regarding machine learning components.

9.2.1.1. Protection interface

The lifecycle of machine learning components goes through the process of collecting raw data, preparing training datasets, conducting training, building systems incorporating trained models, installing the systems in operation contexts, and keeping on working. It must be noted that what is protected is personal data provided by the data subject, and that threats to data protection may arise in various stages of the machine learning component lifecycle, which is summarized below in order.

Case 1: The data subject provides personal data. These raw data are to be provided with appropriate consents by the data subjects. Ignoring the consents may violate the right to privacy.
Case 2: The training data processor builds training dataset from a group of collected raw data. Since the raw data, referring to personal information, are processed, the training data processor can be a threat. Even if the purpose of use is aligned with the scope of users' consent, care should be taken in relation to the detailed content of the consent, as it is often difficult to determine in advance the minimum amount of data definitively needed for the training dataset, which may conflict with the data minimization requirements (Section 9.1.1.2).
Case 3: The component developer conducts machine learning with the training dataset as input information to derive a trained model. If the training dataset is not properly protected and pre-processed, the machine learning component developers could be a threat because they can potentially refer to the raw data from the training dataset.
Case 4: The machine learning system developer builds a machine learning system that incorporates the trained model. If the trained model is not properly protected or pre-processed, the machine learning system developers may be a threat because they can potentially refer to the training data indirectly and even to the raw data from the trained model as well.
Case 5: The system operator installs a final product of the machine learning system and supports their operation. If the input data at operation time include personal data of a certain data subject, and the system operator can access to the input data information, then the system operator may be a threat. Additionally, if the location information concerning with the system installation is indirectly related to sensitive information, the system operator may again be a threat. For example, the very fact that a remote facial recognition system installed in a certain location recognizes a particular data subject reveals the behavior of that data subject, and thus is a threat.
Case 6: The clients, or primary users, use the machine learning system according to the access right given to them. If training data identification (Section 9.2.2) is possible, there may be leakage of personal data and thus those clients can be a threat.

The training dataset (Case 2) and the trained model (Case 3) can also be provided as artifacts for use in the development of new machine learning components. In such reuse cases, threats to

data protection arise.; first, the reuse development itself can pose threats to the terms of consent approved by the data subject (Case 1), i.e., purpose limitations, data minimization, and storage limitations (Section 9.1.2.2), for example, in the case of the compliance with GDPR.

Second, if the training dataset is not properly protected and pre-processed (Case 2), it is possible to refer to the raw data from the training dataset, which implies that developers who reuse the data are a threat. Additionally, if the trained model is not properly protected (Case 3), it is possible to refer indirectly to the training data or even to the raw data from the trained model, which implies that developers who reuse the data are a threat.

9.2.1.2. Training data identification problem

The basis of the personal data protection problem in machine learning is that identifying information about the training data is possible from the trained model. In other words, the process of deriving the trained model does not achieve sufficient de-identification and the trained model cannot be anonymous data of the protection processing level 3 in a technical sense. The trained model is at the protection processing level 2 against the training data identification attempts. The reason for this is that the trained model "remembers" the label of each individual input training data[69] [207].

The memorization problem can be summarized as follows. We compare the likelihood of the predicted label obtained when the input x is given to the trained models M and M' derived from dataset S containing data points <x, y> consisting of data x and label y, and a training dataset S' with this data point removed (S' = S\{<x, y>}). In case of memorizing label of training data, the result of the former M(x) is more likely, while the latter M'(x) is less likely. It is easily seen that M(x) is far likely than M'(x) in case of over-fitting to the trained dataset. This “memorization” phenomenon depends not only on the training mechanism to result in poor generalization performance, but also on the data distribution of the training dataset. [125].

9.2.2. Training data identification methods

Training data identification problem is sometimes categorized as a membership inference, attribute inference and model inversion, or property inference, etc. [132]. These are methods of re-identification against trained models of the protection processing level 2, and are called “inference” because the problem itself leads to a result as a probability value. For example, a membership question is whether some data belong to the training dataset. Since the prediction will always be accompanied with a certain probability even if answered randomly, it is necessary to decide whether the membership inference is successful or not in terms of a given threshold probability value and to compare it with the probability value in case of random answers.

9.2.2.1. Membership inference

Membership inference is a basis of the training data identification, given a trained model, to find out if specific data were used for training[169][175][204]. Note that knowing this membership relationship is, indeed, a threat to the data subject. For example, suppose that a training dataset is built from a list of heavy debtors and a trained model is derived from the dataset. If membership inference demonstrates that certain data were used as training data, it indirectly reveals that the data subject linked to this data was a heavy debtor, too.

9.2.2.2. Attribute inference and model inversion

Attribute inference is to infer sensitive information about the training data by means of information that are publicly available [147][178]. Model inversion is to infer training data from inference results of the model training and to reconstruct some data of the training dataset[93].

9.2.2.3. Property inference

Property inference is to infer global properties of a training dataset by obtaining information that was not originally intended for the target trained model [54][95]. Global properties are characteristics of the training dataset itself and are not necessarily threats related to personal data. For example, properties may refer to specific conditions under which those data were collected, or to the fact that the training dataset is biased when focusing on a certain attribute. The former is an example of trade secrets, the condition used for the raw data collection in this example, being leaked, and the latter suggests the possibility of problematic issues in fairness.

9.2.3. Quality perspectives overlapped

Privacy is a quality perspective in relation to the protection of privacy of data subjects or data protection and relates to the other quality perspectives that the machine learning component should exhibit. Below presents the relationship with fairness and security of both information security and cybersecurity.

9.2.3.1. Relationship to fairness

Ethical AI brings up privacy and fairness as the two main aspects to support human safety [28]. Fairness concerns whether the outcomes of AI systems are biased depending on the sensitive attributes of data subjects. Bias, however, is divergent, and thus needs further criteria based on social justice, regarding discrimination, to determine whether bias at hand is threat to

fairness. Group fairness addresses bias among groups with different sensitive attributes, while individual fairness addresses whether outcomes for a particular data subject are biased when compared with the group that the data subject belongs to.

Fairness has much in common with privacy in that it involves sensitive information such as personal data of data subjects. On the other hand, fairness depends on the system requirements because the notion of bias is based on the higher-level concept of social justice, while privacy is discussed from the general criterion of whether the sensitive information of the data subject is leaked, which is based on legal regulations.

9.2.3.2. Relationship to information security

Privacy is related to information security because privacy violation is considered a specific type of information leakage. Information security has been discussed in terms of three properties (CIA): Confidentiality, Integrity, and Availability [60]. Basically, CIA protects the information managed inside the system by controlling accesses to the information according to the granted authorization. Confidentiality is concerned with direct information leakage of information against unauthorized accesses.

Privacy focuses on the issue of information leakage about data subjects, even when the access to the information is legitimately allowed, and is concerned with indirect information leakage of information. Privacy is related to confidentiality in that both deals with information leakage, and privacy was once discussed as a part of the information security (CIA+P)[155]. In privacy, however, the target of threats is not information managed inside the system, but the data subject, located outside the system, associated with the protected information; this protected information is inferred in an indirect way, using system output information possibly together with auxiliary information and background knowledge.

Because of the inherent technical challenges, privacy has recently come to be considered an area independent of the information security. Privacy is a quality property that ethical AI should satisfy and is one of the main research challenges of machine learning [65].

9.2.3.3. Relationship to cybersecurity

The causes of privacy issues can be discussed in relation to cybersecurity when considering unexpected access to information. Training data inference (Section 9.2.2) can be divided into two specific methods: black box and white box. Black box methods infer from results obtained through normal execution such as authorized access and do not go into the details of the trained model. In a sense, such uses are malicious such that developers and operators did not consider. On the other hand, the white box method uses the internal information that makes up the trained model such as machine learning model, training parameter values, etc.. If the information is obtained through cybersecurity attacks such as model theft, which is the theft of the trained model itself or the learning parameters, it may introduce further threats of indirect information

leakage with the white box methods. Privacy and cybersecurity may overlap significantly and require multifaceted countermeasures to reduce both threats.

9.2.4. **(Supplementary) Countermeasure technology**

This section introduces research trends in data protection measures, focusing on techniques for applying differential privacy to machine learning technologies.

9.2.4.1. Privacy-preserving machine learning

Privacy-preserving machine learning aims to protect training data from the trained model by incorporating differential privacy methods in the learning algorithms that form the core of the training process. The goal is to protect the training data “embedded” in the trained model [47], and basically to increase the strength of de-identification protection. The methods can be used to protect the entire training data or the data attributed to specific data subjects [66].

9.2.4.1.1. Protection of the entire training data

Training of deep neural networks is formulated as a numerical search method for nonconvex optimization problems. A standard method is the stochastic gradient descent (SGD) algorithm, which searches, in an iterative manner, for appropriate training parameter values.

DP-SGD combines SGD with (ε,δ)-differential privacy. When updating the training parameter values in the search process, Gaussian noise determined from the given protection strength ε is added [47]. The training parameters or trained model obtained by DP-SGD are protected by the strength represented by the accumulated protection strength calculated in the iterations. As a result, the threat of training data identification can be reduced.

9.2.4.1.2. Protection per data subject

Federated learning is a method of applying client-server distributed computing architecture to machine learning. A large training dataset is segmented into multiple datasets, and for each segmented dataset, training is conducted on the client using the segmented dataset as input. The intermediate training results obtained from this client-side computation process are aggregated at the server. The client-server collaborative process is repeated to obtain the final trained model [61][108].

When applied to privacy-preserving machine learning, a client is prepared for each training data related to a particular data subject, and a machine learning method with differential privacy, such as DP-SGD, is adopted for the training mechanism at the client side. Since only the client processes the personal data associated with the data subject, it can be expected that the data subject's information will not be leaked to the outside such as the server.

9.2.4.2. Privacy-preserving synthetic data

Privacy-preserving synthetic data is a method for synthesizing a differential privacy-preserving dataset from an input training dataset [202]. This method uses a generative model trained on the input dataset to synthesize new data that follow the same data distribution as the input dataset[165]. Basically, machine learning architectures such as GAN or VAE are used to obtain the generative model together with privacy-preserving machine learning methods (Section9.2.4.1).

In general, the technical and institutional costs of training dataset maintenance are high, and thus reuse or cross-organizational sharing of training datasets is expected. In addition, it is recommended to release a well-prepared dataset as a standard for a specific application domain. Therefore, ensuring the protection of training datasets becomes important.

9.2.4.3. Limitations of existing measures

The privacy-preserving machine learning method is an application of (ε,δ)-differential privacy (Section 9.2.4.1). In general, the strength of protection depends on the value of ε. Therefore, whether the protection is achieved as expected depends on the ε value chosen. In other words, the use of privacy protection learning methods does not necessarily mean that adequate protection is achieved. In addition, increasing the protection strength by decreasing the ε value will worsen the performance of the trained model and go against the usefulness of machine learning performance exhibited by the machine learning components. There is a trade-off between protection strength and usefulness. Furthermore, establishing a general way to determine the appropriate ε value is an open problem.

Now, if a training dataset is biased in the first place and the trained classifier obtained from it produces prediction results that are unbalanced of large variability, then using DP-SGD for this biased training dataset will further increase the variability [56]; the bias in the prediction results may be amplified. For example, if DP-SGD is used with the importance of privacy protection in mind, the privacy protected trained model leads to a situation where group fairness is adversely affected and more unbalanced. It has also been reported that trained models derived by fairness-conscious learning methods increase the threat of membership inference for unprivileged subgroup[74]. Fairness and privacy are two quality perspectives of ethical AI, but are difficult to reconcile with the existing technologies.

Privacy-preserving machine learning methods can, in principle, guarantee data protection. On the other hand, so long as adopting known training data identification methods (Section 9.2.2), it requires evidence through empirical experiments whether effective data protection is achieved. Furthermore, if the worst-case privacy protection estimation method based on the existing differential privacy theory is used, it is difficult to completely avoid training data identification with a reasonable ε value of protection strength[104]. Although the application of differential privacy has become dominant in privacy-preserving machine learning now, further

advances in differential privacy theory that improve the estimated worst-case value are needed (e.g.,[133]).

Last, machine learning training data are multidimensional and sparse, which makes it difficult to protect data with K-anonymity [49]. Thus, there seems no silver bullet, and problem-specific approaches should be taken according to the characteristics of the data to be protected.

9.2.4.4. Data protection impact assessment tool

As research on training data identification has progressed (Section 9.2.2), technical solutions to DPIA (Section 9.1.5) are now underway.

ML Privacy Meter [136] is being developed under the support of a national project in Singapore called AI Singapore. It is a checking tool for training data identification for trained models and is intended for use in risk analysis. ML Privacy Meter provides a set of tools based on black box and white box methods for membership inference, as well as a support for estimating ε values needed for requested protection strength when machine learning mechanisms based on differential privacy (e.g., DP-SGD) are used.

ML-Doctor [124] is a systematic assessment tool being developed in a project of the German National Institute, providing an extensible framework for checking the degree of data protection from various perspectives. It provides membership inference, model inversion, and attribute inference as features to assess the protection strength of trained models in a systematic manner.

Even if your business is in areas not covered by regulatory laws related to data protection (e.g., GDPR), you can mitigate business risk by reducing the likelihood of data protection failures during operations. Therefore, DPIA will be important in the future.

9.3. Quality management of privacy

This section presents an advice on quality management concerning with privacy in view of the internal quality characteristics during the development of machine learning components. The advice is an implementation of the *privacy by design* principle, and can be discussed in relation to activities of domain analysis and system design (Table 6).

Table 6: Overview of quality management implementation

Domain analysis	Protected Data	Use of Artifacts	
	(Section 9.3.1.1) * Compliance with law * Identification of personal data	(Section 9.3.1.2) *Artifacts reuse plan *Confirmation of consent	
System Design	Pre-Stage	In-Stage	Post-Stage

	(Section 9.3.1.2) *Data quality *Pre-processing *Data distribution (the outlier)	(Section 9.3.2.2) *Generalization performance *PPML	(Section 9.3.2.3) *Safeguard
	Trade-off analysis		
	(Section 9.3.2.4) *Protection strength vs. Fairness *Data protection measures vs. Usefulness (Model Accuracy)		

9.3.1. **Domain analysis phase**

For machine learning components, requirements analysis focuses on the privacy domain in terms of personal data protection, whether those data bring about socially unacceptable privacy risks (Section 9.1.1.1). The primary concerns are (1) protected data themselves, and (2) delivered artifacts for future use.

9.3.1.1. Protected data

Assume that the machine learning system under development is subject to privacy-related quality management. It requires studying whether the training data refers to personal data (Section 9.1.1.3). First, it needs to figure out which information is sensitive or personal to be protected, and to comply with what related regulations specify. Such data are either registered or activity; identifying activity data on the Internet requires thorough analysis of the Internet services that the data are collected from, and thus software requirements methods, focusing on data modeling, are helpful. See *A-1: Sufficiency of problem domain analysis* and *A-2: Sufficiency of data design*.

9.3.1.2. Use of artifacts

Artifacts other than trained models can be delivered outside or reused for further development. It is inevitable to make it clear which type of artifacts is subject to privacy-related threats: either training datasets or trained models, or both (Section 9.2.1.1).

First is the case of training datasets. Adaptive maintenance in machine learning software technology is implemented with re-learning, which requires introducing some modifications to the training dataset used for obtaining the current trained model. In such reuse of the training dataset, it required to faithfully follow the terms of consent. (Section 9.1.2.2). Furthermore, training datasets are sometimes delivered outside the organization that developed the original dataset, which is referred to as the third-party reuse. The party must follow the terms of consent

given to the organization, not violating right to privacy of the data subjects.

Second is the case of trained models, namely of pre-trained models. Transfer leaning and knowledge distillation are machine learning methods to reuse pre-trained models; the resultant trained models become components incorporated in new machine learning systems. Transfer learning may be employed as a method to conduct adaptive maintenance, including enhancement of functionalities. Such enhancements should not conflict with the consent of the data subject of purpose and storage period. For example, consider a case that we have a training dataset consisting of facial image data provided with the consent of the data subject to develop age prediction components. Then, conducting transfer learning based on the pre-trained model to build a new machine learning component for race prediction would be re-purposing and thus, violate the terms of consent. Knowledge distillation will comply with the rights of the data subjects in a manner similar to the case of transfer learning. In addition, re-using pre-trained models may make the personal data embedded in the training dataset it available outside. It needs to ensure that the reuse is consistent with the terms of consent by the data subject.

9.3.2. **System design phase**

System design is a group of activities to decide how the machine learning components at hand are developed. The activities at this system design phase are divided into three steps, each looking at different aspect concerning with machine learning component developments, together with analysis of trade-off between varying quality characteristics (Section 9.2.2). The three steps are a Pre-stage regarding to the preparation of the training dataset, an In-stage to devise the machine learning algorithm or the training mechanism adopted in the training process, and a Post-stage to concern with the output of the trained model when input data are entered. These measures adopted in the stages are not always independent, but may be combined appropriately by considering the trade-off relationship. After the development, it is recommended to conduct the data protection impact assessment preferably with the tool assistance (Section9.2.4.4).

9.3.2.1. Pre-stage

Studying data in relation to the privacy protection is conducted in view of both the quality model of individual training data and the data distribution of the training dataset as a whole.

9.3.2.1.1. Quality model for training data

In the first place, it is mandatory to comply with the basis clauses in regulations. Each regulation may specify different quality characteristics that the training data should satisfy for the compliance. For example, the GDPR specifies accuracy and currentness [Article 5]. Mapping the requested data quality characteristic to SQuaRE data quality model may help studying the

quality model of individual training data[143] [162]. See “B-3: Adequacy of data”

9.3.2.1.2. Data protection pre-processing

In the process of preparing raw data into training data, appropriate pre-processing should be applied to prevent leakage of personal information (Section 9.1.3.1). Some measures are known empirically effective such as revising attributes (components of multidimensional vectors), converting to quasi-identifiers, and adding noises, as well as measures such as K-anonymity to introduce data similarity (Section9.1.4.1). It, however, needs to take into account the requirements on the system; there may be possibility that the requirements are not fulfilled due to the protection pre-processing. See *A-1: Sufficiency of problem domain analysis* and *A-2: Sufficiency of data design*.

9.3.2.1.3. Control of data distribution

To lower threats of the training data identification, the data distribution of the training dataset should be controlled so that the trained model does not introduce *memorization* unexpectedly (Section 9.2.1.2). Specifically, the measures start with outlier detection using problem-oriented heuristics [50][216]. Then, the data distribution is manipulated by those methods of removing outliers or adding training data near outliers, depending on the requirements of the target system. See *B-1: Coverage of datasets* and *B-2: Uniformity of datasets*.

9.3.2.2. In-stage

The threat of training data identification can be lowered by using machine learning methods that avoid *memorization* as much as possible (Section 9.2.1.2); the measures to enhance generalization performance make contribution. For this purpose, methods such as regulation and drop-out have been devised so far in standard learning algorithms for deep neural networks. These appropriate state-of-the-art techniques are to be used.

In some cases, privacy-preserving machine learning (PPML) might be used (Section 9.2.4.1). However, it is often difficult to obtain expected data protection performance with the current PPML technology (Section9.2.4.3). Using PPML carelessly may bring about the trade-off issues with usefulness or fairness (Section9.3.2.4).

9.3.2.3. Post-stage

The post-stage measures reduce threats of the training data identification by means safeguards introduced at the exit of trained models. The idea is based on the observation that the training data identification is basically reduced to the membership inference problem, which makes use of the information on prediction probabilities output from the target trained models (Section 9.2.2).

To give a concrete explanation, let us consider a machine learning task to classify the data into C categories. The trained model output result for input x is a C-dimensional vector P^x , whose j-component P^x [j] represents the probability that the input is classified as the category j. When the component j*, which is the maximum of P^x [j]s, matches the tag y, the classification result is interpreted to be correct. The membership inference uses the information in this C-dimensional vector. Then, the idea here is to filter out some of the output information by means of safeguards placed at the end of the trained model.

Specific methods are that the output is to be limited (1) to the predictive label information (the index j* of the component with the largest probability value) only, (2) to the top 1 (<j*, P^x [j*]>) of the predictive probability, (3) to the top 2, or (4) to reduce numerical precision for P^x , etc. [146]. In addition, some known membership inference methods rely on meta-classifiers, which are themselves machine learning classifiers, and the safeguards generate P^x data added with adversarial perturbation so that the meta-classifiers exhibit mis-classification, which may be able to reduce the threats of the membership inference[106].

It must be noted, however, the safeguards introduce a change in the external interface of the trained model, which makes impacts on the machine learning components that use the safeguarded trained models. Therefore, it requires careful consideration whether such safeguards are appropriate in view of the system requirements.

9.3.2.4. Trade-off analysis

Pre-stage measures such as control of empirical distribution of training data and post-stage measures such as reduction of output information are techniques aimed at increasing privacy protection. Although trial and error iterations are involved, those techniques can be practiced as a kind of know-how in the development. Removing or adding data to change the distribution of data in view of working on outliers affects the data distribution, which implies that privacy protection measures may affect the usefulness and functional behavior of the machine learning system to be developed.

If a training mechanism with the differential privacy (Section 9.2.4.1) is employed as an in-stage measure, it is necessary to consider the relationship between the privacy strength ε and its effect on the training data identification method. Since there is no known general method for determining the appropriate ε value, such choice is dependent on the machine learning problem at hand, and might be determined in a trial-and-error manner. It is also necessary to consider how the results of differential privacy protection affect the usefulness and fairness (Section9.2.4.3).

In summary, system design issues in relation to the privacy protection require thorough discussions on the trade-off with the other quality characteristics such as machine learning performance, or fairness.

9.3.3. **(Supplementary) Neural language model**

In neural natural language processing, the privacy issues appear in a different way than in the previous discussions. A neural language model (NLM) is a pre-trained model derived using a large corpus of natural language sentences as training data, and thus an NLM can be regarded as encoding of the input corpus. Typically, natural language processing application systems, such as Q&A systems, are built based on the pre-trained NLM by, for example, transfer learning methods.

When the original corpus contains contents that are contrary to social justice or personal information of data subjects, it is known that the output of the Q&A system may raise fairness issues [124] and privacy issues[137], because the NLM "remembers" the training data [151].

NLM can be used in application systems such as Q&A systems because they faithfully reflect the information embedded in the corpus, which implies that a certain form of *memorization* is inevitable. Therefore, what is problematic is the phenomena of *unintentional memorization* [70]. Eliminating the threat of training data identification or corpus contents identification is, in general, difficult unless the corpus is free from any threat to data subjects[64]. It can be said that the level of threat to privacy issues is determined already at the stage of selecting corpus, i.e., selecting the training data to be used.

9.3.4. **Privacy quality levels**

This section introduces privacy quality levels for different types of artifacts: machine learning systems to be delivered as final products and the other types of artifacts to be delivered for reuse. Since privacy is intended to protect the rights of data subjects, the underlying regulatory law must be complied with (Section 9.1.2). The required level of protection is achieved through a combination of technical and institutional measures. Technical measures include not only those directly related to data protection, but also information technologies that guarantee technical robustness, such as safety, reliability, and countermeasures to cybersecurity.

In the following discussion, the measures are divided into technical measures related to data protection and the others such as the information technology and institutional measures, and the presentation below focuses on the data protection technology. It is because increasing the level of data protection pre-processing reduces the cost of institutional measures and provides an incentive to spend efforts on data protection technologies (Section 9.1.3.2). Last, if a preliminary analysis shows that no personal data are involved, there is no need to consider the privacy quality level at all (Section9.3.1).

9.3.4.1. Machine learning systems

First considers the case for machine learning systems to incorporate machine learning

components. The privacy quality levels refer to the discussions on the cases of protection targets (Section 9.2.1.1). Countermeasures against the threat to training data identification depend on data protection measures at the time of the development (Section9.3.2). Installation and operational threats are addressed by the other technical and institutional measures. Achieving required quality is divided into the following levels.

Quality Level 0 (AIPrLc0): No data protection measures are taken. Use state-of-the-art training methods that focus on improving the generalization performance.
Quality Level 1 (AIPrLc1): Reduce threats to training data identification by improving the training data distribution and introducing safeguards, in addition to the latest training methods. Moreover, tool-assisted data protection impact assessments will be conducted, if necessary.
Quality Level 2 (AIPrLc2): Data protection is given the highest priority and privacy-preserving machine learning methods are used. In addition, the threat of training data identification will be reduced by improving the training data distribution and introducing safeguards. Moreover, tool-assisted data protection impact assessments will be conducted.

9.3.4.2. Artifacts for reuse in further development

When providing artifacts outside for reuse in further development, the provider should be responsible for the quality level regarding data protection for the delivered artifacts, depending on the approved scope of the reuse (Section 9.3.1.2).

9.3.4.2.1. Training datasets

When personal data are included, the training dataset itself is a group of unprocessed data of the protection processing level 0 and needs some protection pre-processing. In general, the required quality level is determined by extents that the training dataset is delivered to. Take as a basis the organization that prepare the training dataset. The required quality level to be guaranteed is moderate within the base organization, because the artifacts are under the organization's control. The level increases if the artifacts are delivered outside, because the control is conducted by third parties. The quality level must be determined by considering the cost of the institutional countermeasure spent to achieve the strength of the privacy protection as specified by underlying regulations.

Quality Level 0 (AIPrLd0): Protection pre-processing is applied to the training dataset as a countermeasure against the re-identification. The other technical and Institutional measures should be implemented, as a precondition, to prevent the re-identification threats, which refers to the appropriate management of information that are related to the protection pre-processing and thus potentially employed in re-identification methods.
Quality Level 1 (AIPrLd1): Protection pre-processing is applied to the training dataset as a

countermeasure against the re-identification. The protection measures must be such that the re-identification requires the external auxiliary information or background knowledge of the protection processing level 2. The measure may refer to privacy-preserving synthetic data to achieve the desired level of protection.

Quality Level 2 (AIPrLd2): Privacy-preserving synthetic data methods are used as a countermeasure against the re-identification. In addition, as tool-assisted data protection impact assessments, training data identification for the generative model used in privacy-preserving data synthesis is performed to confirm that threats are kept below the required level.

9.3.4.2.2. Pre-trained models

When the training data refer to personal data, the pre-trained model itself is a kind of de-identified data of the protection processing level 2, but may be subject to known methods of the training data identification. Therefore, it is necessary to apply protection pre-processing so as to make the re-identification difficult. In general, the required quality level is determined by extents that the pre-trained model is delivered to. Take as a basis the organization that prepare the training dataset. The required quality level to be guaranteed is moderate within the base organization, because the artifacts are under the organization's control. The level increases if the artifacts are delivered outside, because the control is conducted by third-parties. The quality level must be determined by taking into account the cost of the institutional countermeasure spent to achieve the strength of the privacy protection as specified by underlying regulations.

Quality Level 0 (AIPrLm0): No data protection measures are taken. Use state-of-the-art training methods focusing on improving the generalization performance. Strict institutional measures are taken by the organization.

Quality Level 1 (AIPrLm1): Use state-of-the-art training methods that focus on improving the generalization performance, using training datasets after protection pre-processing against the re-identification. If necessary, a tool-assisted data protection impact assessment will be conducted.

Quality Level 2 (AIPrLm2): Use privacy protection learning methods to counter re-identification threats. In addition, the use of training datasets pre-processed with privacy-preserving synthetic data methods will be considered. Moreover, tool-assisted data protection impact assessments will be conducted.

10. AI security

This chapter explains the security of machine learning based systems.

We first describe the overview (Section 10.1), the attacks and damage to machine learning based systems (10.2), the machine learning specific attacks and their countermeasures (10.3), and the reconnaissance and pre-attacks for conducting machine learning specific attacks (Section10.4). We then describe the quality management of the security for machine learning based systems (Section 10.5). We outline risk assessments and present technical controls for improving AI security in a comprehensive and systematic manner for each phase of system design, system development, and system operation.

This chapter can be used not only for improving and assessing the AI security alone, but also for conducting the management of other external qualities shown in other chapters of the Guideline at the same time. As reference information, we remark on the security perspectives for each chapter of the Guideline (Section10.6).

10.1. Overview

10.1.1. The importance of AI security

As with conventional information systems, machine learning based systems are vulnerable to attacks from outside the system. One of the best-known attacks is an evasion attack, which causes a trained model to malfunction with malicious input data.

Such malfunctions of trained models may degrade the quality in use of the system and cause damage to the system, operators, users, and third parties. Thus, security controls to prevent or mitigate attacks are important, especially when the malfunction of trained models results in a high level of human and economic risk, such as in autonomous car and pathological diagnosis systems.

10.1.2. Machine learning specific attacks and their countermeasures

Machine learning specific attacks (hereafter referred to as ML-specific attacks) and their countermeasures have the following characteristics:

- Attacks against trained models used in systems are often hard to be detected technically. For example, backdoors embedded in models may not be detectable. Input data that malfunction the model may not be detected in advance.
- Vulnerabilities in trained models can be mitigated by improving training methods, but are hard to be fixed completely. Therefore, it is essential to take technical controls to mitigate attacks and damage in the overall system design, such as restricting the input to trained

models (Section 10.5.2.4).

– Attacks during data collection, data processing, or system development may cause damage during system operation (Section 10.3.2). Therefore, security controls for machine learning based systems need to cover the entire process of the data collection and pre-processing, and the system development and operation.

10.1.3. Risk assessment and security controls for machine learning based systems

The Guideline presents a quality management method for the external quality *AI security* so that the stakeholders can take *ML-specific security* into account from the design phase of machine learning based systems. In quality management, the developers should conduct risk assessments for the entire system lifecycle to implement security controls, as in the case of conventional information systems.

Risk assessment and security controls are closely linked to the quality management presented in other chapters of the Guideline. While the damage by attacks is characterized as the degradation of external qualities and qualities in use, security controls consist of (a) the assessment and improvement of internal qualities of machine learning components, (b) the security controls at the system level, and (c) the security controls in development and operation.

In the Guideline, we present the tasks during (1) risk assessment (Section 10.5.1), (2) system design and development (Section10.5.2), and (3) system operation (Section10.5.3) along the system life cycle.

We mainly focus on the technical security controls against attacks that exploit vulnerabilities specific to machine learning. We refer to existing standards and frameworks for the security controls common to conventional information systems (Section 10.5.4).

We also remark that this chapter focuses on the technical controls for supervised learning. In future editions, we plan to address the security controls for unsupervised, semi-supervised, reinforcement, and distributed learning.

10.1.4. Related documents

Public organizations have published technical documents on machine learning security. Draft NIST IR 8269 by the U.S. NIST [36] gives definitions of terms and classifications for the security of machine learning based systems in terms of *attack*, *defense*, and *impact*. The ENISA's report [44] gives a glossary of threats to AI systems, classifying assets in the AI ecosystem in six perspectives: *data*, *models*, *actors*, *processes*, *environment and tools*, and *artifacts*. The ENISA's 2021 report[45] classifies threats, vulnerabilities, and controls for machine learning based systems. China's National Information Security Standardization Technical Committee's Draft National Standard[46] gives the security requirements, assessment methods, and evaluation metrics for machine learning algorithms.

In later sections, we present literature that may be helpful for risk assessments and security

controls.

10.2. Attacks and damage to machine learning based systems

We present an overview of the damage caused by attacks on machine learning based systems.

10.2.1. Classification of damage

In the Guideline, we classify attacks against machine learning based systems (hereafter referred to as *systems*) as follows.

a) System malfunction (Section 10.2.2)
   a1) Due to a malfunction of a trained model (Section 10.2.2.1)
   a2) Due to a malfunction of an interpretation functionality of a trained model (Section 10.2.2.2)
   a3) Due to an unintended functionality of a trained model (Section 10.2.2.3)
   a4) Due to other factors (Section 10.2.2.4)
b) Leakage of information on a trained model (Section 10.2.3)
c) Leakage of sensitive information on training data (Section 10.2.4)

We summarize the attack methods that cause these damages in Table 7.

Table 7: Attacks that cause damage to machine learning based systems

<table>
<tr><th colspan="2" rowspan="2">Damage</th><th colspan="2">Attacks causing damage</th></tr>
<tr><th>ML-specific attacks</th><th>Other attacks</th></tr>
<tr><td rowspan="3">System malfunction</td><td>Due to a malfunction of a trained model (Section 10.2.2.1)
Due to a malfunction of an interpretation functionality of a trained model (Section 10.2.2.2)</td><td>Data poisoning attack (Section 10.3.2)
Manipulation of validation/test data (Section 10.3.3)
Model poisoning attack (Section 10.3.4)
Evasion attack (Section 10.3.5)</td><td rowspan="2">Conventional attacks against the software/hardware that implements a trained model and its interpretation functionality
(Not covered in this chapter)</td></tr>
<tr><td>Due to an unintended functionality of a trained model (Section 10.2.2.3)</td><td>Data poisoning attack (Functionality change attack; Section 10.3.2)
Manipulation of validation/test data (Section 10.3.3)
Model poisoning attack (Functionality change attack; Section 10.3.4)</td></tr>
<tr><td>Due to other factors (Section 10.2.2.4)</td><td></td><td>Conventional attacks against the system
(Not covered in this chapter)</td></tr>
</table>

Leakage of information on a trained model (Section 10.2.3)	Model extraction attack (Section 10.3.6)	Conventional attacks for model thefts (Not covered in this chapter)
Leakage of sensitive information on training data (Section 10.2.4)	Information leakage attack of training data (Section 10.3.7) Data poisoning attack (pre-attack for information leakage; Section 10.3.2) Model poisoning attack (information embedding attack; Section 10.3.4)	Conventional attacks for data thefts (Not covered in this chapter)

10.2.2. System malfunction

10.2.2.1. System malfunction due to a malfunction of a trained model

Attacks for a malfunction of a trained model can prevent achieving the functional requirements of a machine learning based system that uses that model. Here are examples of possible incidents caused by such attacks.

- Reduction of safety or increase in risk:
  - Failure to detect objects in automated driving; missing anomalies of drivers in driver assistance;
  - Bypassing malware detection in information security controls;
  - Failure of intrusion detection and abnormal behavior detection in security systems;
  - Failure of facial recognition or other biometrics;
  - Increase in false positives and false negatives in pathological diagnosis systems;
- Decrease in AI performance:
  - Inefficient allocation of vehicles in transportation and logistics; increase in traffic congestion and logistics costs;
  - Decrease in the accuracy of product recommendations, and demand prediction in the retail sector;
  - Inadequate enrollment, hiring, and staffing;
- Decrease in fairness:
  - Discriminatory lending through credit screening systems;
  - Unfair admissions, hiring, and staffing through personnel rating systems;
  - Discriminatory criminal risk assessments by crime prevention systems.

Furthermore, a malfunction of a trained model may cause the system to malfunction, resulting in the loss of privacy and other qualities in use.

ML-specific attacks that can cause a malfunction of a trained model are as follows: *data poisoning attacks* (Section 10.3.2), *model poisoning attacks* (10.3.4), and *evasion attacks* (Section 10.3.5). Detecting a malfunction of a trained model can be failed due to the *manipulation of*

validation/test data (Section 10.3.3).

We remark that a malfunction of a trained model may be caused by conventional attacks against the software/hardware implementing the trained model, such as buffer overrun and fault injection attacks. Countermeasures against such conventional attacks are analogous to the security controls for conventional information systems, hence not addressed in the Guideline. (The same for Sections 10.2.2.2 and 10.2.2.3).

10.2.2.2. System malfunction due to a malfunction of an interpretation functionality of a trained model

If there is a functionality that gives an interpretation of the behavior of the trained model, the attacks that malfunction this functionality can degrade the system's quality in use and worsen the system's transparency and accountability. For example, there are known attacks that reduce the value of the explanations provided by the interpretation functionality or generate incorrect explanations [83][176].

Attacks that cause a malfunction of an interpretation functionality are as follows: *data poisoning attacks* (Section 10.3.2), *model poisoning attacks* (Section10.3.4), and *evasion attacks* (Section 10.3.5).

10.2.2.3. System malfunction due to an unintended functionality of a trained model

The behavior of a trained model depends on the training dataset used to train the model. If a model is trained using an unexpected different dataset, it may learn a functionality that the developer does not expect. In this case, even if the trained model works appropriately, it does not meet the system's requirements, hence resulting in a system malfunction.

Attacks that cause a system malfunction by using a model with an unintended functionality are as follows: *data poisoning attacks for functionality changes* (Section 10.3.2) and *model poisoning attacks for functionality changes* (10.3.4).

10.2.2.4. System malfunction due to other factors

System malfunctions may be caused by the malfunction of components other than machine learning components. Countermeasures against such malfunctions are security controls for conventional information systems, which are not specific to machine learning and hence not addressed in the Guideline.

10.2.3. Leakage of information on a trained model

Attacks that leak private information, such as the parameters and functionality of a trained

model, may result in the leakage of trade secrets and other secrets related to the trained model functionality. It may also be used for the following different attacks against the trained model.

- Information about the trained model may be used to generate adversarial input data that malfunction the model or its interpretation functionality.
- Information about the trained model may leak the privacy information in the training data used to train the model.

An ML-specific attack that leaks information about trained models is known as a *model extraction attack* (Section 10.3.6).

There are also conventional attacks that are *not* specific to machine learning: those against vulnerabilities in the development software and the development environment (Section 10.4.2.1), in the system, the computing environment, and the operation organization during system operation (Section 10.4.2.2). Countermeasures against these attacks are the same as those for conventional information systems and are not covered in detail in the Guideline.

10.2.4. Leakage of sensitive information on training data

When sensitive information is included in training data, an attack to leak training data information may result in the breach of privacy, the leakage of trade secrets, and the violation of laws, regulations, and contracts. An attack to compromise training data may cause damage to third parties, for example, if the training dataset includes sensitive information on the third parties, such as medical personal data, customer sales information, or photos of military facilities.

Notable examples of ML-specific attacks that leak information on training data are *membership inference attacks* and *model inversion attacks* (Section 10.3.7). In these attacks, information on training data used to train a model is compromised by observing the input/output behavior of the trained model. In this chapter, these attacks are collectively referred to as *information leakage attacks of training data*. The details of privacy protection are summarized in Chapter9.

On the other hand, there are conventional attacks for direct data theft. Countermeasures against this type of attack are the same as conventional information security controls and are not covered in detail in the Guideline. As a measure to prevent conventional data theft, secure multi-party computation is a technology for computing encrypted data without decrypting them, and has been actively studied for its applications to machine learning[55][134][157].

10.3. ML-specific attacks and their countermeasures

This section briefly describes the ML-specific attack methods (Table 8) and their countermeasures. First, we describe the classification of attacks based on prior knowledge of

the trained model under attack (Section10.3.1). Then, we explain each ML-specific attack and its countermeasure (10.3.2 through10.3.7).

Note that ML-specific attacks and their countermeasures are currently being actively researched[73] [135] [154] [196]. Thus, the readers of the Guideline are recommended to check the latest information on the technical details of those attacks and their countermeasures.

10.3.1. Access to the models under attacks

Attacks against a trained model are classified in terms of the attacker's knowledge as follows:

- White-box attacks that use parameters of the trained model;
- Black-box attacks that do not use the parameters of the trained model;
- Gray-box attacks between the white-box and the black-box attacks.

A white-box attack assumes a situation where an attacker has access to the parameters and other information of the trained model, for example:

case (1): a publicly available trained model is used in the system;
case (2): the attacker has stolen some parameters of the trained model in advance.

In the case (2), to steal parameters and other information of the trained model, attackers conduct pre-attacks that exploit vulnerabilities in the development software and the development environment (Section10.4.2.1) and in the system, the computing environment, and the operation organization during system operation (Section10.4.2.2).

A black-box attack assumes a situation where the attacker does not need to access the parameters of the trained model but can input data into the trained model. Typically, the attacker provides input to the model or system under attack and observes the output to obtain the information used for the attack. Even if attackers cannot fully observe the output of the model or system, they may generate adversarial data in advance and inputs them into the system.

Table 8: Examples of attack execution phases, attackers, and means of ML-specific attacks

Attack methods	Attack execution phases	Typical attackers	Typical means of attacks
Data poisoning attack (Section 10.3.2)	During the collection and pre-processing of training data	Data providers External attackers	Adding malicious data to the training dataset
	During system development	External attackers	Manipulation of the training dataset
Manipulation of validation data and test data (Section 10.3.3)	During system operation	System users	Inputs during system operation that cause a malfunction of the model (e.g., in case of backdoor exploits)

Model poisoning attack (Section 10.3.4)		During the training and provision of the pre-trained model	Model providers External attackers	Backdoor to pre-trained models
		During system development	External attackers	Malicious training programs; Manipulation of trained models
		During system operation	System users	Inputs during system operation that malfunction the model (e.g., in case of backdoor exploits); Observation of output information, etc., during system operation (for attacks that leak information)
Evasion attack (Section 10.3.5)	white-box	After obtaining the trained model	Providers of input data for system operation	Observation of inputs, outputs, and internal information of the obtained trained model; Generation and manipulation of input data for system operation
	white-box	During system operation	System users	Input of malicious data to the system during operation
	gray-box black-box	During system operation	Providers of input data for operation System users	Manipulation of input data for operation; Input of malicious data to the system during operation; Observation of output information, etc. during system operation
Model extraction attack (Section 10.3.6)	gray-box black-box	During system operation	System users	Input of data into the system during operation; Observation of output information, etc., during system operation
Information leakage attack of training data (Section 10.3.7)	white-box	After obtaining the trained model	External attackers	Observation of inputs, outputs, and internal information of the obtained trained model
	gray-box black-box	During system operation	System users	Input of data into the system during operation; Observation of output information, etc., during system operation

10.3.2. Data poisoning attacks and their countermeasures

10.3.2.1. Overview of the attacks

A *data poisoning attack* is an attack that manipulates training datasets to malfunction the trained models or their interpretation functionalities, to construct models with unintended functionalities, or to induce the leakage of sensitive information. This kind of attack is conducted

during the data collection and pre-processing phase or during the model learning phase. Its damage occurs during the system operation phase.

A data poisoning attack malfunctions the system and decreases safety, AI performance, and fairness[130][177]. It can also induce the leakage of sensitive information[72] and may also cause privacy breaches.

Manipulations of training datasets in data poisoning attacks are classified into (i) data injection, (ii) data modification, and (iii) label manipulation [196].

Typical attackers and their situations are:

(1) a third party manipulates the source or environment of the data (the population);

(2) the data provider manipulates the dataset;

(3) the data curator (who collects/pre-processes data and constructs the dataset) manipulates the dataset;

(4) a third party manipulates the dataset in the process of providing or using the dataset.

We list the following five major types of data poisoning attacks:

(a) Targeted attacks: Attacks that cause a malfunction of a trained model or its interpretation functionality for specific inputs to the model;

(b) Backdoor attacks: Attacks that cause a malfunction of a trained model or its interpretation functionality for unspecified inputs that contain triggering information;

(c) Non-targeted attacks: Attacks that cause a malfunction of a trained model or its interpretation functionality for unspecified inputs to the model;

(d) Functionality change attacks: Attacks that cause a model to learn an unintended functionality;

(e) Pre-attacks for information leakage: attacks that train models to leak sensitive information during operation.

(a) Targeted attacks do not cause a malfunction of the trained model except for certain input data to the model. (b) Backdoor attacks [75] [120] do not cause a malfunction for inputs that do not contain trigger information. For example, they cause the model to malfunction for the images that contain a specific symbol but not for other images. For this reason, it is more difficult to detect the executions of targeted attacks and backdoor attacks.

(d) In a functionality change attack, a model with functionality not intended by the developer is trained by replacing the training dataset with another. For example, if an attacker adds a large amount of data containing discriminatory information to the training dataset, then the trained model will output discriminatory information, contrary to the developer's expectations [170].

In data poisoning attacks other than functionality change attacks, the characteristics of the machine learning algorithm are often exploited to reduce the scale of the manipulation of the

dataset. However, in a functionality change attack, since the goal is to train a model that has an unintended functionality, it tends to require larger-scale manipulations of the dataset, and the characteristics of the machine learning algorithm may not be used to mount the attack.

(e) Pre-attacks for information leakage are classified into (e1) attacks that add sensitive information to training data and (e2) attacks that manipulate the dataset to change the model's behavior and induce information leakage attacks of training data during system operation [72].

10.3.2.2. Technical controls

Technical controls to prevent or mitigate data poisoning attacks are as follows:

(1) Checking the adequacy of the dataset;
(2) Applying data poisoning detection techniques to the dataset;
(3) Using techniques to assess and improve the robustness of the dataset against data poisoning;
(4) Training with a robust learning method against data poisoning;
(5) Conventional security measures against vulnerabilities in the development software and the development environment.

To prevent data poisoning attacks, it is essential to check the adequacy of the dataset, which is classified into the *authenticity* of the dataset, the *credibility* of the provider of the dataset, and the *adequacy* of the process of the data collection/pre-processing (Section 10.5.2.1). However, when input data from external sources are directly used to train the model, e.g., in online learning, it is often impossible to exclude or reduce malicious data from the dataset.

As for data poisoning detection techniques, there are methods for identifying and removing outlier data from training datasets [182]. These techniques are effective for a small number of poison data, since they rely on the fact that poison data tend to have characteristics of outlier data. In contrast, functionality change attacks can be more easily detected by checking the contents of the dataset manually or automatically, since they require a larger scale of the dataset manipulation.

As a technique to improve the robustness of a dataset against data poisoning attacks, data augmentation[63] is effective. Using a sufficient number of training data can also mitigate the impact of poisoning without decreasing the accuracy of the trained model.

Certain robust training methods such as randomized smoothing[166] and ensemble learning (e.g., Bootstrap Aggregating)[105] can be used for countermeasures against data poisoning.

To prevent dataset manipulations, it is important to conduct conventional security measures against vulnerabilities in the development software and the development environment (Section 10.5.2.6).

10.3.3. Manipulation of validation data and test data and their countermeasures

10.3.3.1. Overview of the attacks

If validation or test data are manipulated, the developers cannot evaluate a trained model correctly and may miss attacks. Then, the trained model or system may malfunction during operation.

Manipulation of validation or *test data* is an attack that manipulates validation or test data to keep the developers unaware of the incorrect or unintended functioning of a trained model or its interpretation functionality. This kind of attack is conducted during the dataset collection and pre-processing phase or during the model training phase. Its damage occurs during the system operation phase.

10.3.3.2. Technical controls

Technical controls to prevent or mitigate the manipulation of validation or test data are as follows:

(1) Checking the adequacy of the dataset;
(2) Conventional security measures against vulnerabilities in the development software and the development environment.

For more information, see countermeasures against data poisoning attacks (Section 10.3.2.2).

10.3.4. Model poisoning attacks and their countermeasures

10.3.4.1. Overview of the attacks

A *model poisoning attack* is an attack that either (1) manipulates a pre-trained model, (2) manipulates a trained model, (3) provides a malicious training program, or (4) manipulates a training program to malfunction a trained model or its interpretation functionality or to leak sensitive information. This kind of attack is conducted during the training and the provision of a pre-trained model or during the system development. Its damage occurs during system operation.

The malfunction of a trained model and its interpretation functionality caused by model poisoning attacks can affect the system's safety, AI performance, and fairness.

Model poisoning attacks are classified into (a) targeted attacks, (b) backdoor attacks, (c) non-targeted attacks, (d) functionality change attacks, and (e) information embedding attacks.

As with data poisoning attacks (Section 10.3.2), (a) targeted attacks and (b) backdoor attacks malfunction trained models only for specific inputs that trigger them, while (c) non-targeted

attacks degrade the performance of trained models for unspecified inputs. (d) A functionality change attack may produce a trained model with an unintended functionality, e.g., by replacing it with a different model.

In (e) model poisoning attacks for information embedding [179], sensitive information of training data is embedded in model parameters and hyperparameters before the system operation phase, and is leaked during system operation. This may result in the leakage of privacy and trade secrets.

(1) Manipulation of a pre-trained model:

In model poisoning for pre-trained models, backdoors are embedded in the model in advance and provided to the system developer. When this pre-trained model is used for derivational development or transfer learning, the backdoors in the pre-trained model may be effective even after training, and cause a malfunction of the model or system.

(2) Manipulation of a trained model:

Model poisoning of a trained model can occur (i) when the model training is outsourced to an external developer, (ii) when Machine Learning as a Service (MLaaS) is used as a development platform for model training, or (iii) when a vulnerability in the development software is exploited to mount the attack.

(3) Providing a malicious training program, (4) manipulation of a training program:

A model poisoning attack via a malicious/manipulated training program (Section 2.3.1) is conducted in the cases of (i)-(iii) in (2) above or (iv) when the model training uses a training program developed by third parties.

In federated learning, where multiple clients cooperate in model training, there is a possibility that a malicious client conducts a model poisoning attack. Now countermeasures against this kind of attack are being studied actively. We leave their details for a future edition of the Guideline.

10.3.4.2. Technical controls

Technical controls to prevent or mitigate model poisoning attacks are as follows:

(1) Checking the adequacy of the process of model training and provision;
(2) Using model poisoning detection techniques;
(3) Removing and reducing poisoning in a pre-trained and trained model;
(4) Conventional security measures against vulnerabilities in the development software and the development environment;

(5) Conventional security measures against vulnerabilities in the systems, the computing environment, and the operation organizations during system operation.

To prevent the use of a manipulated pre-trained model, the developers should check the process of model training and provision, which is classified into *the authenticity of the pre-trained model*, *the credibility of the provider of the pre-trained model*, and *the credibility of the training process of the pre-trained model* (Section 10.5.2.1).

To identify the poisoning of a pre-trained model, the developers can use model poisoning detection techniques. However, these techniques cannot always detect poisoning, and there is a possibility of attacks that evade detection techniques. For example, there is an attack method that uses cryptographic techniques to install undetectable backdoors.

To remove or reduce poisoning in a pre-trained and trained model, the developers may pre-process the model. For example, a method to remove backdoors in deep learning models is to prune some nodes of the neural network after fine-tuning (i.e., updating model parameters through additional training) [122]. Poisoning may not be removed by pre-processing the pre-trained model, e.g., when only removing nodes. Furthermore, a model pre-processing may degrade the performance of the model if the fine-tuning is not performed with a sufficient amount of data.

To prevent the manipulation of a training program and a trained model, the developers should conduct conventional security measures against vulnerabilities in the development software and the development environment (Section 10.5.2.6) and against those in the system, the computing environment, and the operation organization (10.5.3.3).

10.3.5. Evasion attacks and their countermeasures

10.3.5.1. Overview of the attack

An *evasion attack* is an attack that causes a malfunction of a trained model by providing the model with specific malicious input called an *adversarial example* during system operation. For example, in an evasion attack against an image classifier, an adversarial example is a carefully perturbed image input that appears natural to human eyes but causes the image classifier to misclassify given image data.

In an evasion attack, a system user is assumed as the attacker. Namely, we assume a situation where a system user inputs adversarial examples to the system during operation. If the system operator uses the system, the attacker is assumed to be the provider of the input data for the system operation.

There are two types of evasion attacks: *white-box attacks* and *black-box attacks*. These attacks differ in the process and method of generating adversarial examples.

In white-box attacks[98][185], adversarial examples for a particular trained model are prepared in advance using internal information about the model (e.g., parameters of the model).

Then the adversarial examples are input into the system during operation.

In contrast, black-box attacks [154] generate adversarial examples without using any internal information about the trained model. Typically, to generate adversarial examples, the attacker inputs multiple data to the system during operation and observes the system's output. Thus, this kind of attack may be prevented or mitigated by restricting the input data to the system or the observation of the output.

However, some black-box attacks may not require many input and output pairs of the system. Such attacks are realized by exploiting the *transferability* of adversarial examples (i.e., the tendency for adversarial examples against other trained models to also be adversarial examples against the trained model under attack). For example, there are attacks that generate adversarial examples by using (1) an approximate model that mimics the input/output behavior of the attacked trained model or (2) another model trained using another dataset that resembles the training dataset.

Evasion attacks are classified into two types in terms of the malfunction or error specificity of the trained model[59].

(a) Error-generic: Attacks that cause some unspecified malfunction of the trained model;

(b) Error-specific: Attacks that cause a specific malfunction of the trained model.

Conventionally, (a) is often referred to as non-targeted and (b) as targeted.

There is another classification of evasion attacks regarding the range of adversarial inputs during system operation or attack specificity.

(i) Indiscriminate: Attacks that cause malfunctions of trained models for unspecified inputs during system operation;

(ii) Targeted: Attacks that cause malfunctions of trained models for specific inputs during system operation.

10.3.5.2. Technical controls

Technical controls to prevent or mitigate evasion attacks are as follows:

(1) Using methods for improving and evaluating the robustness of trained models against adversarial examples;
(2) Restricting inputs to the trained model (restrictions on access rights and on the number/frequency of accesses);
(3) Using techniques to detect adversarial examples;
(4) Using multiple different models and systems together;
(5) Technical controls to prevent or mitigate model extraction attacks (Section10.3.6).

Countermeasures against evasion attacks have been studied primarily in terms of the robustness of trained models against adversarial examples[51][98][117]. Methods for improving and evaluating the robustness of models are summarized in Section10.3.6 of the Guideline.

- Well-known methods for improving robustness are *adversarial training*, *robust training*, and *smoothing*.
- Examples of methods for evaluating robustness are *evaluation by generating adversarial examples* and *approximate calculation methods of maximum safe radius*.
- There are tools to evaluate the robustness of models against adversarial examples: the Adversarial Robustness Toolbox [148], RobustBench[81], CleverHans [156], and Foolbox [163].

Note that the application of methods to improve the robustness of trained models may produce models that may leak more information on the training data. For example, adversarial training (Section 7.6.2.6) may increase the risk of membership inference attacks (Section10.3.7)[180].

In general, it is difficult to remove the vulnerability of trained models to adversarial examples. Thus, if there are no restrictions on input to trained models, it is hard to prevent evasion attacks. Therefore, to mitigate evasion attacks, it is important to restrict the input and output of trained models in the system (Section 10.5.2.4). Moreover, to restrict the input of adversarial examples to the system, there are adversarial example detection techniques[52][127][203]. However, these techniques often fail to detect adversarial examples and should be used only as a secondary countermeasure.

Furthermore, evasion attacks may be mitigated by using multiple models trained using different learning algorithms and different hyperparameters (Section 10.5.2.4). However, we remark that due to the *transferability* of adversarial examples, attackers may be able to mount evasion attacks against different models effectively.

Evasion attacks that malfunction the interpretation functionality of the model (Section 10.2.2.2) tend to be specific to the interpretation method, and thus may be prevented or mitigated by using multiple interpretation methods together.

10.3.6. Model extraction attacks and their countermeasures

10.3.6.1. Overview of the attack

A *model extraction attack* is an attack that aims to steal (1) information about the attributes of a trained model (architecture [151], hyperparameters[193], parameters[189], decision boundaries[107], etc.) or (2) information about functionality of a trained model [174][152] during system operation.

Model extraction attacks may cause the disclosure of trade secrets or other information related to the functionality of the trained model itself. Furthermore, the compromised information of the trained model can be used in evasion attacks (Section 10.3.5). The information of the leaked trained model may be used for evasion attacks or for information leakage attacks of training data (Section10.3.7).

In a model extraction attack, the attacker inputs data to the system in operation, and

observes the input-output relation of the trained model. In other words, an attacker (e.g., a system user) is assumed to have black-box access to the trained model under attack.

10.3.6.2. Technical controls

Technical controls to prevent or mitigate model extraction attacks are as follows:

(1) Using detection techniques for model extraction attacks;
(2) Modifying the output information of the trained model;
(3) Ensemble learning.

To detect model extraction attacks, the developers can use a technique that observes the distribution of a set of input data to a trained model and detects inputs for model extraction (PRADA [107], etc.).

Countermeasures for model output information are: (i) not outputting confidence values, (ii) rounding confidence values, and (iii) adding perturbations to confidence values. While such output modification may help mitigate some model extraction attacks, it often does not mitigate the attacks very well.

Ensemble learning can mitigate model extraction attacks to some extent by using multiple models together.

10.3.7. Information leakage attacks of training data and their countermeasures

10.3.7.1. Overview of the attack

There are many known attacks that leak information on training data used to train the model by observing the model's behavior during system operation. In this chapter, such attacks are collectively referred to as *information leakage attacks of training data*. These attacks can lead to the leakage of sensitive information on training data, possibly resulting in the leakage of privacy or trade secrets.

There are two types of information leakage attacks of training data: *white-box attacks* and *black-box attacks*. In a *white-box attack*, an external attacker obtains a trained model and then observes this model's behavior. In this case, the attacker does not need to observe the system's input and output during system operation. In contrast, a *black-box attack* assumes a system user as an attacker that inputs data into the system during operation and observes the system's output (and possibly other internal information).

Information leakage attacks of training data include the following (Section 9.2.2 for details).

- Membership inference attack (Section 9.2.2.1)[175]: Attacks that infer whether specific data belong to the training dataset used to train the model (membership information);
- Model inversion attack (Section 9.2.2.2)[93]: Attacks that (approximately) recover

partial data from the training dataset;
- Attribute inference attack (Section 9.2.2.2)[178]: Attacks that infer sensitive information related to the training data;
- Property inference attack (Section 9.2.2.3)[54]: Attacks that infer global properties about the training dataset.

Note that there are data poisoning attacks that can induce information leakage attacks of training data (Section10.3.2)[72]. There are also model poisoning attacks that embed sensitive information in advance to leak it during system operation (Section 10.3.4).

10.3.7.2. Technical controls

We describe technical controls to prevent or mitigate information leakage attacks of training data in Sections 9.2 and9.3.

10.4. Preliminary stages of ML-specific attacks and their controls

An attacker against a machine learning based system may combine ML-specific attacks with other attacks to perform a multi-stage attack (Section 10.3). They may collect information and make a pre-attack to conduct an ML-specific attack. Therefore, the system developers and operators may be able to deter, prevent, or mitigate the ML-specific attack by implementing technical controls against such information collection and pre-attacks.

Therefore, this section describes information collection and pre-attacks to conduct ML-specific attacks.

10.4.1. Information used for ML-specific attacks

We remark on information used for ML-specific attacks.

10.4.1.1. Misuse of specification information

Information about the specifications of the system and the trained models can be used for ML-specific attacks. For example, information about the learning algorithm and hyperparameters for training the model can be used in a data poisoning attack (Section 10.3.2).

10.4.1.2. Misuse of models and datasets

Information on trained models or on resembling models can be used to generate adversarial

examples for evasion attacks (Section 10.3.5), and to generate inputs to the models in membership inference attacks (Section10.3.7). Thus, it is important to prevent attackers from obtaining information on the trained models or resembling models. We remark that the risk of evasion attacks is greater when the system is built using publicly available trained models without modifying them.

The training dataset or resembling one can be used to build a model that approximates the trained model under attack, and can be used to mount evasion attacks or membership inference attacks. Therefore, preventing attackers from obtaining the training dataset or resembling one may help mitigate those attacks. It should be noted that the risk of evasion attacks is greater when the model is trained using only a single publicly available dataset.

10.4.2. Pre-attacks for ML-specific attacks

This section describes pre-attacks to conduct ML-specific attacks. Conventional information security controls are used to counter those pre-attacks (Sections 10.5.2.6,10.5.3.3, and 10.5.3.4).

10.4.2.1. Pre-attacks exploiting vulnerabilities in the development software and the development environment

Manipulations of training datasets, training programs, and trained models may cause data poisoning attacks (Section 10.3.2) and model poisoning attacks (Section 10.3.4). These manipulations can be conducted through *pre-attacks that exploit vulnerabilities in the development software and the development environments*.

For example, a model poisoning attack can be conducted through a pre-attack that exploits vulnerabilities in software libraries for machine learning, such as TensorFlow and PyTorch, to install backdoors in the development software or in the trained models.

10.4.2.2. Pre-attacks exploiting vulnerabilities in the system, the computing environment, and the operation organization during system operation

Manipulation of the trained models in model poisoning attacks (Section 10.3.4) and the generation of input during system operation in evasion attacks (Section 10.3.5) and information leakage attacks of training data (Section10.3.7) can be conducted through *pre-attacks that exploit vulnerabilities in the system, the computing environment, and the operation organization during system operation*.

For example, it may be possible to steal trained models and execute white-box evasion attacks or information leakage attacks of training data through pre-attacks exploiting vulnerabilities, e.g., *unauthorized access to the system*, *reverse engineering of the trained model*, or s*ide-channel attacks*. In some cases, an attacker may be able to steal trained models and

execute a *white-box evasion attack* or *information leakage attacks of training data*. If an attacker has physical access to the hardware, a *fault injection attack* or *hardware trojan attack* may cause a malfunction of the trained model [205].

10.5. Quality management for AI security

This section outlines the quality management for the security of machine learning based systems so that the stakeholders can take *the ML-specific security* into account from the system design phase. We overview the security risk assessment (Section 10.5.1) and present security controls in the system design and development phase (Section 10.5.2) and in the system operation phase (Section 10.5.3) in a comprehensive and systematic manner.

10.5.1. Security risk assessment

For the security of a machine learning based system, a security risk assessment should be performed for the entire system lifecycle, analogously to the assessment of a conventional information system. In a security risk assessment, we first identify the stakeholders and assets appearing in the system lifecycle. Then we analyze possible threats and vulnerabilities by investigating attack surfaces and evaluating the impact of attacks on the external qualities and the qualities in use of the Guideline, e.g., by using attack trees. These analyses should be performed not only in the design/development of the system, but also in periodic security risk assessments.

A risk assessment for machine learning based systems needs to identify not only the threats and vulnerabilities specific to machine learning, but also those for conventional information systems in the entire system lifecycle. However, it may be hard to identify all threats and vulnerabilities, and to implement sufficient security controls for all possible threats and vulnerabilities. For this reason, we need to prioritize the threats and vulnerabilities to be handled in the security risk assessment, and should implement security controls in order of priority.

We remark that some security controls to protect an external quality or a quality in use may degrade another external quality or quality in use. For example, a security control for strengthening the robustness against adversarial examples may increase the risks of privacy attacks (Section 10.3.5.2). For another example, a countermeasure against privacy attacks may be incompatible with fairness (Section9.3.2.4). Therefore, we may need to analyze the trade-offs among different external qualities and qualities in use.

At this moment, both security risk assessments and security controls for machine learning based systems still need more research and development. The attack methods in Sections 10.3 and10.4 and the security controls in Section10.5 may not be sufficient. Therefore, the system developers and operators should investigate the latest information on the security of machine learning technologies used in the specific system they deal with. Then, they need to take a risk-

based approach to identify the threats and vulnerabilities in the system, the development environment, and the operation organization.

10.5.1.1. Reference information and case studies of risk assessments

We present information possibly helpful to conduct a security risk assessment of a machine learning based system. Microsoft and MITRE et al. present ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) [39] to systematize the threats to machine learning based systems and the tactics and techniques at each stage of actual attacks. As for case studies, a Japanese domestic study group on machine learning engineering works on a risk analysis based on the second edition of the Guideline (released on July 5, 2021) and other documents, and proposes a simple analysis tool for developers [222].

In the risk assessment for conventional information systems, the developers should refer to ISO 27000 series [10], ISO/IEC 15408 (Common Criteria)[3], and NIST SP 800-30[37]. As for the security controls for control systems in factories and critical infrastructure, the developers should refer to IEC 62443[16] [17] and the Cyber/Physical Security Framework (CPSF) [35] issued by the Ministry of Economy, Trade, and Industry (METI).

10.5.2. Security controls in the system design and development phase

The developers should design and implement *security controls* based on a security risk assessment (Section 10.5.1) for the specific system, the development environment, and the operation organization. In Table 9, we summarize the security controls for machine learning based systems that the developers may need to design and implement during the system design and development. It should be noted that the security controls discussed in this section may not cover all security controls necessary for the specific system and the environment.

In the actual system design and development, it is not always necessary to implement all the security controls listed in this section. For example, if the training data do not contain sensitive information, the developers do not need to implement countermeasures against the information leakage attacks of training data.

Hereafter, we present security controls against only attacks from the outside of the system. To prevent or mitigate the damage caused by malicious or negligent developers, we should implement the security controls for conventional information systems, not addressed in the Guideline.

Table 9: Examples of security controls in the system design and development phase

Security controls	Goals	Attacks	Examples of details of security controls
Controls in obtaining datasets and pre-trained models	Suppressing or preventing	Data poisoning attack	Checking the authenticity of the datasets and the pre-trained models

(Section 10.5.2.1)		manipulations	Manipulations of test data etc. Model Poisoning Attack	Checking the credibility of the providers of the dataset and the pre-trained models	
				Checking the process of data collection and pre-processing	
				Checking the process of the training of the pre-trained models	
Controls in the evaluation and processing of datasets (Section 10.5.2.2)		Avoiding using datasets that lead to unintended functionality	Data poisoning attack	Checking the internal qualities "A-1: Sufficiency of problem domain analysis", "A-2: Sufficiency of data design", "B-1: Coverage of datasets", "B-2: Uniformity of datasets", "B-3: Adequacy of data	
		Detecting attacks	Data poisoning attack	When the dataset is not adequate	Using data poisoning detection techniques
		Preventing or mitigating attacks			Processing the datasets (data augmentation, addition, synthesis, removal, etc.)
			Evasion attack	Using data augmentation techniques (adding adversarial examples, etc.)	
			Information leakage attack of training data	Removing sensitive information	
				Using privacy-preserving data synthesis techniques (Section 9.2.4.2)	
Controls in training the model (Section 10.5.2.3)		Preventing or mitigating attacks	Data poisoning attack	When the dataset is not adequate	Using robust training methods against data poisoning
		Detecting attack	Model poisoning attack	When the pre-trained model is not adequate	Using model poisoning detection techniques
		Preventing or mitigating attacks			Model pre-processing (parameter reduction, re-training, etc.)
			Evasion attack	Using techniques to assess and improve the internal qualities "C-1: Correctness of trained model" and "C-2: Stability of trained model" (e.g., adversarial training, use of assessment tools)	
			Information leakage attack of training data	Mitigation measures for the information leakage attacks of training data (Section 9.2.4.1)	
				Using tools to assess information leakage of training data (Section 9.2.4.4)	
Controls in the design	(1) Pre-processing	Detecting attacks	Evasion attack	Using techniques to detect malicious input data	

and development of the system (Section 10.5.2.4)	program	Preventing or mitigating attacks	Model extraction attack Information leakage attack of training data	Restricting access rights and the number/frequency of accesses to the system to limit or stop malicious input data
	(2) Post-processing programs and interpretation functionalities	Detecting attacks		Using techniques to detect attacks in the post-processing programs and in the interpretation functionality
		Preventing or mitigating attacks		Restricting the disclosure of the model's output and internal information
				Adding perturbations to the model's output and internal information
	(3) System configuration	Preventing or mitigating attacks	Attacks in general	Preventing or restricting the observation of the system's behavior during system operation
		Preventing or mitigating damage	Attacks in general	Using techniques for detecting malfunctions of the model and the system
				Using multiple different models and systems in parallel
	(4) Conventional vulnerability	Removing or mitigating vulnerability	Conventional attacks against the system	Security controls for conventional systems (including the use of techniques for the internal quality "D-1: program reliability")
Controls against the misuse of specification information, models, datasets, and related information (Section 10.5.2.5)		Preventing or mitigating attacks	Attacks in general	Restricting the disclosure of specification information on the model and the system
				Restricting the disclosure of models used in the system, datasets used for model training, and related information
Controls for vulnerabilities in the development software and the development environment (Section 10.5.2.6)		Removing or mitigating vulnerability	Conventional attacks against the system	Security controls for conventional systems

10.5.2.1. Security controls in obtaining datasets and pre-trained models

The developers should take measures to suppress or prevent the manipulation of the datasets and the pre-trained models used to train the model.

- Check *the authenticity of the datasets and the pre-trained models*. For example, by using digital signatures or frameworks for providing trusted datasets and pre-trained models, check that the dataset or the pre-trained model has not been manipulated.
- Check *the credibility of the provider of the datasets and the pre-trained models*. For example, check information to determine the social credibility of the provider.
- Check *the process of data collection and p*re-*processing*. For example, check with the data providers the process of data collection and pre-processing and the security controls to prevent or reduce the manipulation of data that may cause poisoning attacks or contain

sensitive information. (See Section 7.5.1 for the data collection policy).
- Check *the process of the training of the pre-trained models*. For example, check with the model provider regarding the process of model training and the development environment.

In certain situations, sign a contract for the provision of datasets and pre-trained models, and the process management. See Sections 6.5.2.4 and 6.5.2.6 concerning the process management for data validity.

10.5.2.2. Security controls in the evaluation and pre-processing of datasets

The developers should evaluate the adequacy of the dataset used to train the model.

- To avoid using a dataset that results in a model with functionality not intended by the developers (Sections 10.2.2.3 and 10.3.2), check the internal qualities *A-1: Sufficiency of problem domain analysis*, *A-2: Sufficiency of data design*, *B-1: Coverage of datasets*, *B-2: Uniformity of datasets*, and *B-3: Adequacy of data*.

If the developers cannot confirm the process of data collection/pre-processing, they should assume a possibility of an attack, use attack detection techniques on the dataset, and pre-process the dataset to prevent or mitigate the possible attack.

- To prevent or mitigate data poisoning attacks, apply data poisoning detection techniques to identify and remove data that degrade the model's performance (Section 10.3.2.2).
- To mitigate data poisoning attacks, increase the amount of data in the dataset sufficiently, e.g., by data augmentation (Section 10.3.2.2).
- To prevent or mitigate evasion attacks, increase the training/validation/test datasets by adding adversarial examples and other data.
- If the dataset may include privacy information, trade secrets, or other sensitive data violating laws, regulations and contracts, then identify and remove these data or information and add other data to reduce the impact of these sensitive data.
- Use the training data generated by privacy-preserving data synthesis techniques (Section 9.2.4.2).

10.5.2.3. Security controls in training the model

The developers should take measures to prevent or mitigate poisoning attacks when they assume a possibility of model poisoning, e.g., when they cannot confirm the process of the model training/provision.

- To prevent model poisoning attacks, use model poisoning detection techniques (Section 10.3.4.2) to detect poisoning in the pre-trained and the trained models.

- To mitigate model poisoning attacks, pre-process the pre-trained and the trained models (Section 10.3.4.2).

Furthermore, the developers should take measures during the model training process to prevent or mitigate the attacks during system operation.

- In the model training process, use techniques to evaluate and improve the internal quality *C-1: Correctness of trained model* (Section 7.6).
- In the model training process, use techniques to evaluate and improve the internal quality *C-2: Stability of trained model* (Section 7.6). In particular, use techniques to evaluate and improve the robustness of models against adversarial examples (Section 7.6.2). Well-known tools for evaluating model robustness against adversarial examples are Adversarial Robustness Toolbox [148], RobustBench[81], CleverHans[156], and Foolbox[163].
- In the model training process, takes measures to reduce information leakage of training data from the trained model (Section 9.2.4.1).
- In the model training process, use tools to evaluate information leakage of training data from the trained model. Well-known tools are ML Privacy Meter [136] and ML-Doctor[124]. See section9.2.4.4 for details.

We remark that more and more new libraries and benchmarks for evaluating trained models have been developed and released. Thus, the developers should find and use libraries and benchmarks that cover the latest attack methods.

10.5.2.4. Security controls in the system design and development

In designing and developing the system, the developers should implement security controls for (1) pre-processing programs, (2) post-processing programs and interpretation functionalities, (3) the overall configuration of the system, and (4) conventional vulnerabilities.

(1) Pre-processing programs

To prevent or mitigate attacks, the developers should implement measures to detect, suppress, or prevent malicious input in pre-processing programs, i.e., programs that process input to the trained model during system operation.

- Use a program for detecting suspicious inputs during system operation that attempt to learn the behavior of trained models (Section 10.3.6.2) to prevent or mitigate evasion attacks, model extraction attacks, and information leakage attacks of training data.
- Restrict access rights to the system and the number or frequency of accesses to the system to prevent or reduce malicious input to the model during system operation.
  - Restrict or suspend access rights and services and impose other penalties on users

who query suspicious inputs into the system.

- When input data during system operation is obtained from an external environment (e.g., image data obtained by a camera), limit the number/frequency of data input to the model.

- Use a program for detecting adversarial example inputs during system operation to suppress or prevent evasion attacks.
- Use a program for detecting suspicious input data during system operation to suppress or prevent malicious inputs other than adversarial examples (e.g., input to trigger backdoor attacks) during system operation.
- Pre-process (e.g., cleanse or transform) the input data during system operation to prevent or reduce the impact of malicious input on the model.

We remark that detection programs are not always able to detect attacks, and attackers may query inputs that evade the detection programs. Therefore, the developers should implement other measures to prevent or mitigate attacks, and use detection programs only as a secondary measure.

(2) Post-processing programs and interpretation functionalities

To prevent or mitigate attacks, the developers should use a *program for detecting attacks during system operation* in post-processing programs that processes the output of the trained model and interpretation functionalities.

To prevent or mitigate attacks, the developers should also implement measures in post-processing programs and interpretation functionalities to restrict the observation of output and internal information of the trained models such as confidence scores and interpretations of the models during system operation.

- To prevent or mitigate attacks, restrict the system from outputting information unnecessary for the system's operation.
- To prevent or mitigate attacks, limit the quantity/quality of the output/internal information by adding perturbations (Sections 9.3.2.3 and10.3.6.2).

Although these measures may reduce the possibility/probability of successful attacks, they do not guarantee the prevention of attacks. Furthermore, restricting the observation of the model's output and internal information may reduce the transparency and accountability of the system.

(3) Overall configuration of the system

To prevent or mitigate attacks, the developers should implement security controls to prevent or limit the observation of the system's behavior during system operation.

Furthermore, the developers should implement technical measures regarding the system configuration to mitigate or prevent damage caused by the malfunction of the trained model

during system operation or by information leakage from the trained model.

- To prevent system malfunctions, use techniques for *detecting malfunctions of the model or system*.
- To reduce the degree and frequency of system malfunctions and information leakage in the system, use multiple different models and systems in combination if necessary. For example, *ensemble learning* is used to mitigate evasion attacks and other attacks by using the outputs of multiple different models. Note, however, that it may be possible to configure successful attacks against models of multiple architectures. Furthermore, using multiple models and systems increases the cost of training and system operation.

(4) Conventional vulnerabilities

The developers should implement security controls for conventional vulnerabilities in the system. We overview conventional security controls in Section 10.5.4. See also *D-1: Program reliability* in Sections 6.8 and7.7 for the details of the following:

- Reliability of *training programs* used for training;
- Reliability of *prediction or inference programs* used in system operation;
- Authenticity and reliability of programs provided by third parties.

10.5.2.5. Security controls against the misuse of specifications, models, datasets, and related information

To deter or mitigate attacks, the developers should restrict the disclosure of (1) specification information on the models and system (e.g., hyperparameters and architecture), (2) models used in the system, (3) datasets used to train the models, and (4) information related to these models, as much as possible.

These security controls prevent information collection during the initial reconnaissance phase of attacks. Although they are useful in limiting the attacker's prior knowledge and in deterring attacks, they do not guarantee the prevention of attacks.

In actual development, models are often trained on publicly available datasets, and thus it may not be possible to keep the dataset information confidential. In such cases, the developers should implement other security controls.

10.5.2.6. Security controls for vulnerabilities in the development software and development environment

The developers should implement security controls for conventional information systems to prevent or mitigate attacks exploiting vulnerabilities in the development software and the development environment (Section 10.4.2.1). They should collect the latest information on

vulnerabilities in the development software and the development environment, such as machine learning frameworks.

10.5.3. Security controls in the system operation phase

Operators of machine learning based systems should enforce security controls for system operation. In Table 10, we summarize the security controls during system operation. It should be noted that the security controls shown in this section may not cover all necessary controls. The developers should conduct a security risk assessment for the specific system and the operation environment, design security controls in the system operation phase, and present them to the system operators.

We remark that this section deals only with technical controls against external attacks. To prevent or mitigate the damage caused by (1) malicious or negligent operators and (2) non-human factors such as natural disasters and infrastructure failures, the system developers and operators should enforce security controls for conventional information systems, not addressed in the Guideline.

Table 10: Examples of security controls in the system operation phase

Security controls	Goals	Attacks	Examples of details of security controls
Monitoring of inputs, models, and system during system operation (Section 10.5.3.1)	Detecting attacks	Evasion attack Model extraction attack Information leakage attack of training data	Using techniques to detect malicious input data Manually checking input data for system operation
	Suppressing or preventing attacks		Checking the authenticity of input data for system operation Checking the credibility of the providers of input data for system operation Checking the process of the collection and pre-processing of input data for system operation
	Preventing or mitigating damage		Using techniques for the internal quality "E-1: maintainability of qualities in system operation" Detecting the malfunction of the model and the system during system operation Manually checking malfunctions
	Detecting, preventing, or mitigating attacks	Data poisoning attack	Controls in obtaining, evaluating, and pre-processing data for re-training

Additional controls for the system and the operation environment (Section 10.5.3.2)	Responding to the deteriorated quality of the model in operation Recovering from the damage by attacks	Attacks in general	Re-training the model during system operation, and security controls for the re-training Rewinding the model to a previous version Changing the system location or external environment
Controls for vulnerabilities of the system, the computing environment, and the operation organization during system operation (Section 10.5.3.3)	Removing or mitigating vulnerability	Conventional attacks	Security controls for conventional information systems

10.5.3.1. Monitoring of inputs, models, and system during system operation

The system operator monitors and checks the inputs to the system during system operation.

- When input data during system operation are obtained from users or an external environment (e.g., when image data are obtained by a camera), the system operator should check *the input data during system operation* by using detection techniques for (i) *suspicious input data attempting to learn the model's behavior* or (ii) *malicious input data during system operation*, or by manually checking the input during system operation.
- When input data during system operation are obtained from a third party or a public database site, the system operator should check (i) *the authenticity of the input data for system operation*, (ii) *the credibility of the provider of the input data for system operation*, and (iii) *the process of the collection/pre-processing of the input data for system operation*. The checking procedure is similar to the case of the training datasets for the model (Section 10.5.2.1).
- When acquiring data or datasets for additional training during system operation, the system developers and operators should enforce the same security controls as in the design and development phase (Sections 10.5.2.1 and10.5.2.2).
- The system operators should use techniques to evaluate and improve the internal quality *E-1: maintainability of qualities in system operation*.
- The system operators should manually or automatically detect the malfunctions of the models and the system during system operation.

We remark that detection techniques may not be able to detect attacks, and attackers may query inputs that evade the detection. Therefore, the system operator should enforce other security controls to prevent or mitigate attacks and use detection techniques only as secondary measures.

10.5.3.2. Additional security controls for the system and the operation environment

The system operators should evaluate the quality of the model in operation and the attack/damage situation. To respond to the deteriorated quality of the model in operation or to recover from the damage by attacks, the system operators should re-train the model during system operation or rewind the model to a previous version if necessary. In the case of re-training of models during system operation, the system developers and operators should enforce the same security controls as those for model training during the design and development phase (Section 10.5.2.3).

The system operator may also change the system location and the external environment to reduce the opportunity for malicious input to the system during system operation.

10.5.3.3. Security controls against the vulnerabilities of the system, the computing environment, and the operation organization during system operation

The system operator should enforce security controls for conventional information systems to prevent or mitigate the attacks that exploit vulnerabilities in the system, the computer environment, and the operation organization during system operation (Section 10.4.2.2).

10.5.4. Security controls not specific to machine learning

We briefly describe security controls that are not specific to machine learning. To enforce security controls for conventional information systems, the developers should refer to standards and frameworks, such as ISO 27000 series [10], ISO/IEC 15408 (Common Criteria)[3], NIST SP800 series [37], NIST Cyber Security Framework[38], and guidelines issued by IPA (Information-technology Promotion Agency, Japan).

The developers should also refer to sector-specific guidelines and references provided by ISACs (Information Sharing and Analysis Centers). For example, the following ISACs have been established in Japan:

- Japan automotive ISAC (https://j-auto-isac.or.jp)
- ICT ISAC Japan (https://www.ict-isac.jp/)
- Japan Electricity ISAC (https://www.je-isac.jp/)
- Financials ISAC Japan (http://www.f-isac.jp)

For business areas that deal with life and property risks but do not have ISACs, the developers can refer to security standards such as Payment Card Industry Data Security Standard (PCIDSS)[42].

As for the control systems in factories and critical infrastructures (IACS, Industrial

Automation Control System), the developers should refer to IEC 62443 [16][17] and the Cyber Physical Security Framework (CPSF) by the Ministry of Economy, Trade, and Industry (METI) [35].

IEC 62443 covers (1) common concepts, reference models, and roles of relevant stakeholders in general, (2) requirements and guidelines for policies and procedures for the management and operation of organizations involved in control systems, (3) security functional requirements and design and technology of security functions required for IACS, and (4) security of the components that make up the system.

CPSF provides a framework for security controls in industrial society in Society 5.0. In order to properly capture the risk sources, industrial society is viewed as a three-layered structure by connections between organizations, mutual connections between physical space and cyberspace, and connections in cyberspace and six components including organization, people, components, data, procedure, and system. Then the functions at each layer are defined as objects to be protected. CPSF presents examples of security measures from the viewpoint of possible security incidents, risk sources of incidents for each component, and requirements for security measures.

10.6. (informative) Remarks on security perspectives in each chapter

In Table 11, we list the examples to do for each section of the Guideline under the consideration of security. It should be noted that the developers need to check and add requirements according to the specific system they design and develop. As for fairness and privacy, we focus on cases where fairness and privacy are subject to cyber-attack damage.

Table 11: Security considerations

Chapter and section	Guideline description	Security response plan	Remarks
1.3.2	Lifetime-long requirements for risk assessment	At the time of security risk assessment, the results of the functional safety risk assessment and the items to be treated with priority are included in the targets of assumed attacks.	Include items such as privacy and fairness that related to the target business for which the system will be used.
1.3.3	when a machine leaning based system is developed by sharing works	When referring to terms and guidelines related to AI and machine learning, include compliance with referenced standards and guidelines for the entire supply chain in the contract.	Machine Learning Quality Management Guidelines, ISO/IEC regulations, etc. GDPR, China Cyber Security Law, etc. depending on the business to be designed and developed. Also refer to laws related to outsourcing, such as the Dispatched Worker Law and the Subcontracting Law.
1.3.3	Security risk of contamination of learning results due to intentional inclusion of improper data	● Collect information on attacks specializing in AI and machine learning, such as evasion attacks, data poisoning attacks, and model poisoning attacks exemplified in Chapter 9 of the Machine Learning Quality Management Guidelines, analyze attack scenarios, and identify threats and vulnerabilities. implement measures for ● Collect information on attack and defense methods and update defense measures periodically.	The cycle of review of overall defense measures should be within one year. Defense measures should be prioritized and implemented from the most feasible items.

1.5.1	Safety	To confirm and examine threats and vulnerabilities with respect to risk targets. From the perspective of safety, human life, and economic efficiency, items with significant damage spillover are identified, risks are classified according to the magnitude of damage, assumed as attack targets, and used to analyze threats and vulnerabilities. (e.g., fairness, privacy)	
1.5.2	AI performance	Assuming an attack scenario in which performance is the attack target, identify threats and vulnerabilities and take countermeasures.	
1.5.3	Fairness	● Assume attack scenarios for information that may be subject to attacks (e.g., gender, skin color, etc.) using Chapter 8, identify threats and vulnerabilities, and implement countermeasures. ● Collect information on the targeted business to periodically review the items subject to fairness attacks and update the content of monitoring and countermeasures.	Chapter 9 will be used as a reference for analysis.
1.5.4	Privacy	● In particular, when the change the nature of the information contained in information assets is assumed (for example, strong correlations generation or lost between multiple pieces of information), monitor and implement regular countermeasures periodically. ● If the new properties acquired above are sensitive, the damage will be greater if attacked, so be careful. ● Confirm whether the new characteristics observed by monitoring violate laws and regulations such as domestic personal information protection law, GDPR, China cyber security law, and US CCPA.	● Pay particular attention to GDPR and China Cyber Security Law (because the regulations and sanctions are very strict) ● Collect information from NIST and ISO/IEC on a regular basis.
1.6.2	Social aspects such as ethics	Examine that the data to be used meets the legal requirements for regulations and contracts regarding the acquisition and use of data for legal rights and public interest policy reasons. Furthermore, examine the need for rights/contractual adjustments.	For reference information.[32][223], etc. Keep as up-to-date information as possible, since it is necessary to follow updates in the legal system.

1.7	Internal quality characteristics	Organizing the status of internal quality characteristics in the system to be developed and assuming data to be attacked and attack scenarios that will cause serious damage, take countermeasure threats and vulnerabilities.	Refer to the reference in the Guideline to check the internal quality status of the development target system.
1.7	Property B-1: Coverage of datasets	Assuming an attack scenario in which integrity is compromised among security requirements, analyze threats and vulnerabilities by enumerating the causes that triggered the attack, and consider countermeasures.	
1.7.1	Sufficiency of requirements analysis	Consideration should be given to whether the inference results and learning models output by the system to be developed encompass sensitive information. When conducting a BIA (Business Impact Analysis)/PIA (Privacy Impact Analysis), a list of system operation flows should be made to verify the combination of situations (use case review).	Countermeasures are required if the sensitivity of the information changes in time series.
1.7.1	analysis from the request side such as risk analysis/failure mode analysis and bottom-up analysis	Security risk analysis during the development phase should be conducted in parallel with other risk analyses, such as functional safety.	Conduct both traditional information security and security with a focus on AI/machine learning.
1.7.2	Combination of situations	Identify threats and vulnerabilities to input/output data for each combination and consider countermeasures.	
1.7.4	"what kind of fairness" is required	Assume damage to the target system with reference to Chapter 9 and consider countermeasures.	Consider attacks that make it impossible to ensure fairness by artificially processing data.
1.7.5	The data must not have been inappropriately altered (authenticity), and the data must be sufficiently new.	Among the perspectives of security risk assessment, integrity verification is applied to the calculation of the impact of attack scenarios.	Perform a security risk assessment of the environment where the data used for learning is stored.
1.7.5	data selection adequacy Adequacy of labeling	When identifying the damage caused by an attack, include attacks that violate the adequacy of data selection and labelling.	Identify the causes of damage and apply them to the listing of threats and vulnerabilities.

1.7.5	Reorganizing data to meet new requirements and policies	If the rearrangement of data changes the damage assumptions from an attack, the list of threats and vulnerabilities should also be updated according to the increase or decrease in damage.	Check the scope of system affected by the data rearrangement.
1.7.6	overfitting	Include attacks that force the system to overfit the target when assuming damage.	For example, addition and processing of training data by worms
1.8.1	Quality Inspection	Include items related to security risks in quality inspection items and establish a process work path that returns to upstream processes for reconsideration if risk issues are discovered during inspection (any inspection item).	Include non-security risk items as necessary.
1.9.1	Social Principles on Human-centric AI	Implement risk countermeasures based on periodic information collection and damage assumptions regarding new attack target perspectives.	Include in the PDCA cycle of security.
2.2.1	IEC 15408	IEC15408 is a typical framework for security risk analysis and countermeasure planning but select one that fits your requirement analysis from IEC27000 series, IPA guidelines, etc. Assumption of AI-specific attack scenarios, identification of threats and vulnerabilities, and study of countermeasures should be conducted.	
2.3.1, 2.3.2	machine leaning based system structure stakeholders of development and their roles	For AI-specific attacks, identify machine learning-specific threats and vulnerabilities through attack tree, use case analysis, and information asset accounting, and study countermeasures.	● For example, assume an attack tree for the attack targeting AI exemplified in Chapter 9, and determine countermeasures and priorities based on the results. ● In the use case analysis, players other than those involved in development (e.g., system users and maintenance staff) should be added.
No. 3 item of 2.3.3	Fairness	List damages that infringe on fairness as targets of security damages. ● List the items of fairness that are subject to damage ● Create an attack tree that targets the listed fairness items.	

NO.4 item of 2.3.3	Attack resistance	The following are the viewpoints of security attack resistance. ● Defend against attacks in the first place. ● Suprress the effect to the actual business by minimizing the damage according to the tampering of the information even if attacked. ● Even if attacked, the calculation status before the attack remains, all the resources are instantly discarded and replaced with the remaining resources without any impact on the actual business.	Consider the scale of implementation of each theme.
No. 5 item of 2.3.3	Ethicalness	Confirm items of ethics and examine scenarios in which damage could occur (attack tree study).	
No. 6 item of 2.3.3	Robustness	Assume an attack tree that causes damage to robustness.	
No. 1 item of 2.3.4	System life cycle process	Provide risk assessments (gates) at each stage of the system life cycle.	Regarding the content and scale of implementation, items will be implemented in accordance with the order of priority, with regular implementation (within one year) in mind. (If it does not fit on a feasible scale, there is a high possibility of disfigurement.)
2.3.5	Terms related to use environment	In system components specializing in AI and machine learning, such as learning and inference, and the environment where learning models are stored, AI-specific requirements are required in addition to conventional information security.	
2.3.6	Terms related to data used for building machine learning	Same as above	

No. 4 item of 2.3.7	A form of operation to collect data and carry out additional training for machine learning during operation and to update trained machine learning models when necessary	General information security does not assume that the data used by the system will acquire new characteristics (such as fairness or privacy) during operation. The system will include periodic monitoring to ensure that the data stored by the system has acquired new characteristics through additional learning, and to consider process and procedure for the new characteristics if necessary.	
3.1	Table 1: Estimation of AI safety levels for human-related risks Table 2: Estimation of AI safety levels for economic risks	For each cell in the table, an attack tree of cases with serious damage from the expected impact is assumed and prioritized.	While it is possible that safety may not be considered when business solutions are built solely on IT systems, safety considerations are mandatory for safety-related embedded systems, such as collision mitigation brakes.
3.3	Fairness	Create an attack tree assuming a case where personal rights and assets are damaged and add it to the threat and vulnerability consideration items.	Particular attention should be paid to scenarios associated with additional learning
3.4	Provisions for personal data protection processing referred to by national and regional laws and regulations	● Use personal data protection processing regulations as a reference for attack trees and damage assumptions in security risk assessments. ● While planning countermeasures, it is advisable to consider a mechanism for monitoring violations of personal data protection processing regulations. ● Confirm the trend of damage cases and regulations related to personal data protection processing.	Sanctions for confirmed violations could affect management.
4.1.1 Figure 11	Handling of development processes with several operational stages	● Risk factors (e.g., spurious correlation) should be verified each time a PoC is conducted to verify whether the requirements meet the specifications. ● Implement the mechanism to monitor risk factors and consider countermeasures should also be included in quality inspection and operational performance monitoring.	

4.2.1 Figure 12	Process model in the stage of machine learning building	● Confirm the handling of sensitive information and generation of new sensitive information at the design stage of datasets, models, and tests, analyze threats and vulnerabilities, and incorporate countermeasures. ● Confirm the handling of sensitive information and the generation of new sensitive information based on the inference results output from repeated training and quality confirmation/verification work, analyze threats and vulnerabilities, and incorporate countermeasures. responses.	
4.2.1 Figure 13	Model of preprocessing for training	Handle sensitive information during each of the preprocessing steps, check for the generation of new sensitive information, analyze threats and vulnerabilities, and incorporate countermeasures.	
4.2.1.1	ML Requirements Analysis Phase	● Regarding learning data as information assets, assuming the relationships among data (e.g., correlations) that are sensitive, which are expressed by reorganizing the nature of input/output data and reorganizing them as a requirement for concrete construction as the target of damage, identify threats and vulnerabilities, and consider countermeasures. ● For quality requirements to be determined based on the nature of the data and the specific data set itself, threats and vulnerabilities are identified and countermeasures are discussed, assuming damage from attacks.	
4.2.1.2	Training data composition phase	● Accounting for datasets as information assets, scrutinizing the relationships (e.g., correlations) between data types stored by the dataset and the nature of the data (e.g., fairness and privacy), identify threats and vulnerabilities and consider how to respond to them, based on the damage that could occur in the event of an attack. ● Investigate new relationships (e.g., fairness or privacy) have been created between data after preprocessing, identify threats and vulnerabilities based on the assumption of damage in the event of an attack, and consider countermeasures.	

4.2.1.3	Iterative training phase	● Extract machine learning models, hyperparameters, and learning models for implementation as information assets, and scrutinize relationships among data stored in datasets (e.g. correlation) and characteristics of data (e.g., fairness and privacy) Then, identify threats and vulnerabilities based on the assumption of damage in the event of an attack, and consider countermeasures. ● Investigate new relationships (e.g., fairness or privacy) have been created between data after preprocessing, identify threats and vulnerabilities based on the assumption of damage in the event of an attack, and consider countermeasures.	
4.2.1.4	Quality check/assurance phase	Treat test datasets, test results (including test confirmation), and data causing false inferences as information assets, assume damage due to falsification or theft of information that can be used as a reference for attacks, identify threats and vulnerabilities and consider countermeasures.	
4.2.2	System building/ integration test phase	Conduct a security risk assessment of the parts of the system other than machine learning components in accordance with the ISO/IEC 27000 series, ISO/IEC 15408 Common Criterial, and guidelines, frameworks, and tools provided by the IPA, and take action based on the results of the assessment and consider countermeasures based on the results.	● Design the assessment methodology with reference to the various ISACs for business sector-specific requirements, or the PCI DSS requirements if no ISAC has been established. ● Pay attention to the selection of guidelines and frameworks when life, safety and property are involved.
4.3	Quality monitoring/operation phase	● Treat relationships (e.g., correlation), properties (e.g., fairness and privacy), and hyperparameters among data included in learning data as information assets, perform periodic monitoring to assume damage and take appropriate measures. ● Periodically review the relationship and nature of data.	

5.1.1.1	In cases where safety functions are required,	Events assumed to cause serious damage in the course of safety considerations will result in greater damage if attacked, so such events should be handled with higher priority.	Examine safety related to systems using AI and machine learning.
5.1.1.3	Examination on risk scenarios related to system use	Prioritize high-risk events because the damage will be greater if they become the attack target.	Identify risks deeply related to AI and machine learning and how to take countermeasure to them.
5.1.1.4 5.1.1.5	Qualities in use	If damage is assumed from the perspective of the adopted quality metrics, analyze threats and vulnerabilities and take countermeasures.	When using external quality characteristics, follow the damage assumption,of threat/vulnerability analysis and countermeasure regarding external quality characteristics.
5.1.2	risks of physical or human damages	Prioritize high-risk events because the damage will be greater if they become the attack target.	Identify risks deeply related to AI and machine learning and how to take countermeasure to them.
5.2.1	entruster and the development entrustee should build a consensus	Select a policy from consideration and implementation of risk countermeasures, risk transfer, and risk approval by listing security risks by extracting those relevant to the business in question from attack cases deeply related to AI and machine learning at the time of consensus building.	● Risk transfer: Entrusting security requirements to others by implementing systems and functions responsible for security (example: solution development and security measures on AWS) ● Risk acceptance: A management decision that accepts the relevant security risk and does not implement countermeasures or transfer.

5.2.2	Role clarification	Confirm that security risk assessments are carried out at the following stages in the progress of the work process, and decisions are made regarding the implementation, transfer, and approval of countermeasures. ● Approval of requirements analysis results ● When the external specification of the software is approved ● When software operation verification specifications are approved ● When software verification and validation results are approved	In particular, check KPI items that are deeply related to AI and machine learning.
5.2.3 3)	Quality management methods during operation	● Treat relationships (e.g., correlation), properties (e.g., fairness and privacy), and hyperparameters among data included in training data as information assets and perform periodic monitoring to assume damage and take appropriate measures. ● Periodicaly review the relationship and nature of data.	
5.3	Delta development	When reusing existing software components, it is necessary to conduct a security risk assessment in the context in which the new system is used. This also applies to risk countermeasures implemented as a result of assessment.	
6	Internal quality	Based on the quality level setting and design decided by the developer, for Sufficiency of problem domain analysis, Coverage of distinguished problem cases, Coverage of datasets, Uniformity of datasets, Adequacy of data, Correctness of trained models, Stability of trained models, Maintainability of qualities in operaion, Reliablity of underlying software systems, assume that the most serious damage will occur, identify threats and vulnerabilities based on attack scenarios, and implement countermeasures.	From each perspective, identify and take countermeasure to threats and vulnerabilities based on the level (LV) set by the developer and its basis.

6.1.2.1	attributes and their viewpoints	● Check whether the attribute is sensitive information (eg skin color, race, etc.), and if it is sensitive, consider whether to stop using the attribute or take measures such as anonymization. ● Scrutinize the relationship between attributes (eg correlation) and the characteristics of attributes (eg fairness and privacy), identify threats and vulnerabilities based on the assumption of damage in the event of an attack, and respond accordingly.	
6.1.3	Safety levels	Factors that increase risk are the cause of greater damage when an attacker attacks, so identify threats and vulnerabilities and respond based on the results of analysis in accordance with requirements.	
6.2	Coverage for distinguished problem cases	Identify and respond to threats and vulnerabilities assuming attacks that violate designed completeness.	
6.3	Coverage of datasets	Analyze and respond to threats and vulnerabilities by enumerating the causes triggered the attack, assuming an attack scenario compromising the integrity among security requirements.	
6.5.2.1	Unification and scrutiny of labeling policies	Assuming an attack scenario in which an attack causes data to go against policy, identify and respond to the threats and vulnerabilities involved.	
6.5.2.2	Consistency checking and rechecking of datasets	Assessments to be conducted in response to changes in functional requirements and use environments, and the process for outsourcing work, should be included in the security risk assessment.	It is recommended to include the perspectives of information security and supply chain security.
6.5.2.3	Handling the long tail and determining mismeasurement and outliers	Assuming an attack scenario in which stored data falls into a state violating the handling policies for long-tail, mismeasurement and outlier, identify and respond to the threats and vulnerabilities involved.	

6.5.2.4	Addressing Data poisoning	● Respond to relevant threats and vulnerabilities assuming attack scenarios in which data are poisoned. ● Consider how to detect poisoning of data. ● Respond to system configurations and usage scenes of development target systems for security requirements other than cyber security.	It is recommended to use guidelines (provided by ISAC and ministries and agencies) according to the business model in which the development system is used to respond to the usage scene. If guidelines are not provided, refer to the guidelines for finance and critical infrastructure.
6.5.2.5	Freshness	Identify and respond to the threats and vulnerabilities involved in a hypothetical attack scenario in which the currentness of training data is compromised.	
6.5.2.6	Establishment of a system and mechanism for process	Establish security risk assessment rules that is applied to the entire supply chain and include in the contract regular audits and storage of audit trails in accordance with the rules.	
6.5.3	Requirements for quality levels	Reflect on each of the points listed in 6.5.2 of this table.	
6.6.2	Relative behavior of indicators	● Prepare and monitor means to check for damage (tampering, etc.) to the trained model regarding the behavior of the input in the training dataset. ● Prepare countermeasures assuming possible causes when trained models are tampered with or otherwise damaged.	It is recommended to leave a trail (a record of the evaluation index) about the behavior for the input included in the training dataset.
6.7.2	Evaluate and improve stability	● Prepare and monitor means to check for damage (tampering, etc.) to the trained model regarding the response to the input in the training dataset. ● Prepare countermeasures assuming possible causes when trained models are tampered with or otherwise damaged.	It is recommended to keep a trail of responses to inputs not included in the dataset for each of the iterative training phase, quality verification evaluation phase, and quality monitoring operation phase.
6.8.1	Open-source implementation	Respond to publicly available threat and vulnerability information.	You may refer to CWE and CVE.
6.9.2	quality degradation	Prepare a response process and system infrastructure if the cause of quality degradation during operation is by attack.	

7.1.2	Estimation of risk factors	Categorizing risk items according to the degree of damage, and assuming those that cause serious damage as the target of damage by attackers, extracting attack scenarios, identify threats and vulnerabilities, and consider countermeasures.	
7.1.3	final implementation form to some extent in order to conduct such an analysis	Apply the key points of the forecast to operational monitoring, and consider methods to distinguish from the initial business assumptions (concept drift) and attacks.	
7.2	Coverage of distinctive problem cases	See proposed security measures in Chapter 6, " quality management characteristics".	
7.3	Coverage of datasets	Refer to security measures in Chapter 6 " quality management characteristics" ● Scrutinize and reflect properties such as hidden correlations between adopted feature values and overlooked feature values through additional test at the data preparation stage and additional test at the test stage ● Scrutinize the characteristics of the feature periodically by monitoring processing.	
7.3	Coverage of datasets	Assuming an attack scenario that compromises the integrity of the security requirements, extract the causes of the attack, analyze the threat and vulnerability, and consider countermeasures.	
7.4	Uniformity of data sets	See proposed security measure in Chapter 6, "quality management characteristics".	
7.5.1	Data collection policy	See proposed security measures in section 6.5.2.1	

7.5.1	Assessment of whether the requirements definition has been updated in line with the data collection policy, and whether it is necessary to reconfirm the contents confirmed in the previous steps (e.g., internal quality A-1 to B-2) in line with the update.	If policies and requirement definitions have been updated, the assumption of security damage may have also changed, and if changes are necessary, they should be reflected in the threat and vulnerability analysis and countermeasures.	
7.6.1.3	fuzz testing	If the input data are related to threats/vulnerabilities, include the test items to input fuzzing data if necessary.	
7.6.1.5	Automatic generation of test inputs	If necessary, include verification items targeting model evasion, model extraction, and adversarial example, which are attacks deeply related to AI and machine learning.	
7.6.2.2	regularization	In dropout, the percentage of inactive neurons is a hyperparameter. Such hyperparameters that determine the design of the neural network, optimization, etc., should be listed as information assets.	Other items in Section 7.5.2 should be considered in the same manner as on the left.
7.6.2.3	adversarial data generation	Assume threats and vulnerabilities and consider countermeasures from attack scenarios that assume attacks by adversarial example generation, if necessary.	
7.6.2.6	adversarial training	If necessary, assume threats and vulnerabilities from attack scenarios that assume attacks using adversarial training and consider countermeasures.	
7.7.3	Common Vulnerability Enumeration (CVE)	CVEs and CWEs are very useful as threats and vulnerability information for general information security, but they do not include perspectives deeply related to AI and machine learning. Therefore, security perspectives on AI and machine learning should be set according to the system to be developed and used to evaluate the system.	See Sections 10.1.4 and 10.5.1.1 for information on security.

7.7.5	Software updates	One of the recent advanced attack methods is related to software updates. Since it is possible that attacks focusing on machine learning components may be carried out through malware contamination and transaction hijacking, attack scenarios should be envisioned and verified.	
7.8.1	monitoring	● If there are security requirements for monitoring during operation, a monitoring mechanism should be placed according to the requirements. ● Relationships between features that may cause serious damage if attacked (e.g., hidden correlations), especially in sensitive cases, will also be monitored from a security perspective.	
7.8.2	Concept drift	Same as monitoring	
7.8.3	updating trained machine learning models	Same as monitoring	
8.1.1	Social demands	As for laws and regulations, social principles, guidelines, and international standards, assume damage in the event that a guarantee is requested or have a large impact on business, and incorporate them into attack scenario assumptions. (See Section 8.4 for assumed attack scenarios.)	
8.1.1.1	ethics fairness	Assuming damage and attack scenarios in the event of a breach of ethics and fairness, identify threats and vulnerabilities leading to the cause and establish countermeasures.	
8.1.2.1	Discrimination by race, religion, etc.	Assume damage in order of greatest impact on business in the event of damage, such as race and religion.	
8.3.3	fairness requirements social demands	Assume an attack scenario in which the content of the fairness requirements or social demands is the final attack target.	
8.3.5	Variables	Assume an attack scenario in which variables are attacked.	
8.4.2	bias	Assume an attack scenario that violates system requirements for bias.	

8.4.3	data with various sensitive attribute sensitive attributes	Assume an attack scenario that infringes on sensitive attributes requiring special consideration.	
8.5.1	AIFL	Since the impact of the output of the relevant product/service on the individual rights and the social acceptability of the system is evaluated, it is advisable to assume the areas that would be damaged and the damage level if the attacks were successful in parallel.	Requirement definition stage in the system development process
8.5.2.18.	Pre-processing approach, in-processing approach, post-processing approach	Consider whether the means to realize the pre/in/post processing approaches can be included in the security risk assessment and conduct the assessment if necessary.	
9.1.1.2	negative externalities	Include in the attack scenario that assumes personal data re-identification as the damage of the attack.	
9.1.1.3	registered data and activity data	Assume registered data and activity data as the damage target in the attack scenario.	
9.1.2.2	the right to be forgotten Purpose limitation Data minimization Storage limitation	Scrutinize whether consent, right to be forgotten, purpose limitation, data minimization, and storage limitation are subject to attack, include them in the assumed attack scenario in the case that they are the target of attacks.	
9.1.3	processed data	Scrutinize whether the protected processed data is the target of an attack, and if so, include it in the assumed attack scenario.	

9.1.4	privacy metrics	If an attack method that violates the metrics is found, it will be included in the scope of assumed damage. Example Data similarity: a method to uniquely identify a record in multiple records. Outcome indistinguishability: methods by which desired records can be distinguished between adjacent databases.	
9.1.5	re-identifying pre-processed data d	Include the damage of re-identification of protected processed data in the attack scenario assumed in the security risk assessment.	
9.2.1.1	Case 1 -Case 6	Assuming attack scenarios for which the provided personal data, training datasets, trained learning models, machine learning systems, and operational systems are targeted, include them into the target of security risk assessment.	
9.2.1.2	identifying information about the training data is possible from the trained model.	Include " identifying information about the training data is possible from the trained model " in the attack scenario envisioned by the security risk assessment.	
9.2.2.1	Membership inference	Include attack scenarios with membership inference in the security risk assessment.	
9.2.2.2	Inferring sensitive information from public information	Include threats and vulnerabilities related to sensitive information in security risk assessment considerations.	
9.2.2.2	Attribute inference is to infer sensitive information about the training data by means of information that are publicly available	Include predictive inference results in the information assets handled by the security risk assessment, if necessary.	

9.2.2.3	Property inference	The following information not included in the design content will be scrutinized i) Information other than that covered by the learned learning model ii) Information that is not assumed as a result of predictive inference. Attack scenarios that infer global properties by identifying the above information should be included in the security risk assessment. (If necessary, third-party identification of information should also be considered.)	
9.2.3.3	Relationship to cyber security	As an effort toward privacy in cybersecurity, PIA (Privacy Impact Analysis) will be conducted alongside BIA (Business Impact Analysis) in the requirements definition stage of system development. The content of PIA is defined in ISO/IEC 29134 (JIS X 9251). In particular, ANNEX B provides examples of threats to privacy and can be used as a reference when creating a threat list for security risk assessment. In addition, it is recommended to refer to the "Promotion of PIA Efforts" provided by the Personal Information Protection Commission as a practical guideline for conducting PIAs.	
9.3.1	Domain analysis phase	As mentioned above, PIA is implemented as a way of responding to privacy in cyber security.	
9.3.2	Pre-stage, In-stage, Post-stage	If there are stage-specific development items, include them in the security risk assessment.	
9.3.2.1.1	accuracy, currentness	The security risk assessment in accordance with the ISO/IEC 27000 series requires that correct information is maintained as integrity, so it is recommended to verify this as well.	

9.3.2.1.2	appropriate pre-processing should be applied to prevent leakage of personal information	Since the security risk assessment along the ISO/IEC 27000 series verifies confidentiality, it is advisable to check the status of information leakage prevention measures when considering this item.	
9.3.4.1 9.3.4.2.1	quality level regarding data protection	It is advisable to proceed with the PIA in parallel with the study of the quality level of the design object.	

11. (informative) Information on related documents

The content of this chapter is informative.

11.1. Relation with other guidelines

11.1.1. Contract guidelines for AI of the Ministry of Economy, Trade and Industry

The Contract Guidelines on Utilization *of AI and Data* [32] published by the Ministry of Economy, Trade and Industry summarizes considerations concerning contracts with business partners when they cooperatively develop systems containing AI (e.g. machine learning) in accordance with contracts such as sales order or quasi-mandate.

The Contract Guidelines clarify roles and responsibilities in contracts between business partners engaged in development, while the Contract Guidelines summarize quality of service which providers have to provide to users of developed system. From the standpoint of the Guideline, qualities in use of products and services defined in the Guideline are elaborated with the cooperation of all stakeholders listed in the Contract Guidelines for development, and then they are provided to users. It is out of the scope in the Guideline how to concretely realize the quality with the cooperation of business partners and how to conclude a contract and share responsibilities. Stakeholders should agree on these matters based on the Contract Guidelines. On the other hand, the Guideline can serve as a base of examining technical matters related to quality when stakeholders share responsibilities and exchange information. In Section 5.2, we briefly analyzed a possibility of sharing roles as one example.

The Contract Guidelines state that a non-waterfall model is appropriate for sales orders in the development process, while the Contract Guidelines (p.46) focus not only on (3) *Development phase* but also phases such as (4) *Additional learning phase* before and after (3). Therefore, the Guideline defines a system lifecycle process by a wide range from system planning to disposal after operation. A model mixed with waterfall model is basically used and summarized as *Figure 6: Conceptual diagram of mixed machine learning lifecycle process* in page 26. With regard to the relation with *gradually exploratory* development processes recommended in the Contract Guidelines, the developmental phase in the center of Figure 6 in the Guideline corresponds basically to *the development phase* in the Contract Guidelines.

11.1.2. Relations with Guidelines for Quality Assurance of Machine leaning based Artificial Intelligence (QA4AI)

The Consortium of Quality Assurance for Artificial Intelligence-based products and service (QA4AI) in Japan has published *Guidelines for Quality Assurance of Machine leaning based Artificial Intelligence 2020.08* [43] in August 2020. The QA4AI Guidelines propose the following five axes to be considered in quality assurance of AI products with the aim of giving *common*

guidelines for quality assurance of AI products.

A) Data integrity
B) Model robustness
C) System quality
D) Process agility
E) Customer expectation

Moreover, the QA4AI Guidelines present check lists of these axes and a list of specific quality management technologies.

As regards the relation between the Guideline with the QA4AI Guidelines, the three quality-assurance axes above (A, B and C) are considered to correspond to the internal qualities listed in Section 1.7 (page 17) of the Guideline. Therefore, the technologies specifically listed in the QA4AI Guidelines correspond to the quality management method for each internal quality characteristic presented in Chapter 7 and give some insight to engineers who are responsible for realizing the quality with the complementary help of the two guidelines. Table 12 explains the specific relation between the items listed in the checklists for those three axes of the QA4AI Guidelines and the internal qualities listed in Section 1.7 of the Guideline.

Although no clear quality management axis has been established in the Guideline for the quality assurance axes (D and E above), they can be realized through the application process envisioned in the Guideline (Section 5.1) and a business process with customers included therein.

Consequently, the QA4AI Guidelines published by the Consortium are beneficial as a reference for engineers who actually create machine leaning based AI to find feasibility of technologies to improve and elaborate internal qualities of machine learning components. On the other hand, the Guideline intends to comprehensively analyze matters necessary for businesses that plan and develop overall systems containing machine leaning based AI to ensure qualities in use throughout the lifecycle process and list them as much as possible. Therefore, it is thought that these two guidelines complement each other.

Table 12: Analysis of the relation with the QA4AI Guidelines (Version of August 2020)

Internal quality characteristics in the Guideline	Checklist in the QA4AI Guidelines
A-1: Sufficiency of problem domain analysis	2.2.1 Data integrity (b) Adequacy of training data (b.i) 2.2.2 Model robustness (h) Diversity of validating data (h.i)
A-2: Sufficiency of data design	2.2.1 Data integrity (a) Sufficiency of the amount of training data (a.i) (a.ii) (a.iii) (b) Adequacy of training data (b.i)

B-1: Coverage of datasets	2.2.1 Data integrity (d) Impartiality of training data (f) Consideration of characteristics in data
B-2: Uniformity of datasets	2.2.1 Data integrity (d) Impartiality of training data
B-3: Adequacy of data	2.2.1 Data integrity (b) Adequacy of training data (b.ii) (b.iii) (c) Suitability of training data for requirements (c.ii) (e) Complexity of training data (g) Adequacy of value ranges in training data
C-1: Correctness of trained models	2.2.1 Data integrity (i) Adequacy of test data 2.2.2 Model robustness (a) Sufficiency of model accuracy (b) Sufficiency of model generalization performance (c) Sufficiency of model evaluation (d) Adequacy of learning process (e) Adequacy of model structure 2.2.3 System quality (a) Suitability of value provision by the system (a.ii)
C-2: Stability of trained models	2.2.2 Model Robustness (b) Sufficiency of model generalization performance (d) Adequacy of learning process (e) Adequacy of model structure (f) Adequacy of model testing (g) Model robustness
D-1: Reliability of underlying software systems	2.2.1 Data integrity (k) Adequacy of data processing programs 2.2.2 Model robustness (k) Adequacy of model implementation in code
E-1: Maintainability of qualities in operation	2.2.1 Data integrity (j) Consideration of effects of online learning 2.2.2 Model robustness (i) Sufficiency of testing after model updates (j) Consideration of model obsolescence 2.2.3 System quality (i) Consideration of system quality degradation
Items corresponding to external qualities	2.2.3 System quality (a) Suitability of value provision by the system (c) Adequacy of units of system evaluation

	(d) Mitigation of impacts due to accidents (e) Prevention of accident occurrence (f) Mitigation of AI influence

11.2. Relations with international initiatives for quality of AI

Currently, diverse international initiatives for quality of AI have been taken including those mentioned below. The Guideline adopts adaptable parts from these initiatives. Moreover, we will actively present outstanding knowledge acquired from the Guideline toward international standardization.

11.2.1. Quality and safety

An examination on quality and safety has started at ISO/IEC JTC 1/SC 42/WG 3. Gaps with existing standards for quality characteristics and quality assurance technologies of AI (ISO/IEC 25000 (SQuaRE)[7][8], ISO/IEC/IEEE 29119-4 (testing technology)[11] and functional safety (IEC 61508[12], ISO 26262[9])) have been analyzed. In addition, discussions on topics such as data quality have started at SC42.

In Europe, the European Commission published the European AI Policy Guidelines [27] in 2018, including ensuring an appropriate ethical and legal framework, and established the AI-HLEG (AI Expert Group) [29]; in 2019, the AI-HLEG published the European AI Ethics[30]. In 2020, the European Commission published the European AI White Paper [28], which requires six requirements for AI systems in high-risk industrial domains and high-risk applications, including training data and accuracy. In April 2021, following a public consultation on the European AI White Paper, the European Commission published the European AI Bill [25], which requires eight requirements for data and accuracy. The content of the European AI legislation may change in the process of discussion in the European Parliament following the Commission's initiative.

It is possible that the European AI Bill, the world's first legally binding AI hard law proposal, will ultimately establish the objectives or goals to be protected, where the Guideline can provide the technical means to achieve the goals in the future. The bill requires, in its data and data governance requirements, that high quality training, validation, and test data be used for development, and that the characteristics, properties, and elements of the geographic, behavioral, and functional use environments be considered. It suggests the importance of B-1: Coverage of datasets and B-2: Uniformity of datasets, corresponding in the Guideline. The bill also calls for accuracy, robustness and cybersecurity requirements to ensure consistent performance throughout the lifecycle, declared accuracy levels and standards, and cybersecurity against adversarial data. It also indicates the importance of C-1: Correctness of trained models and C-2: Stability of trained models in the Guideline.

11.2.2. Transparency

In Europe, the European AI Ethics Guidelines and the European AI Bill have outlined requirements for transparency, which are expected to have an international impact, especially in EU member states. In April 2019, the AI-HLEG presented a checklist for ensuring transparency, which will be verified through actual demonstrations with companies, and in 2020 published a self-assessment list of trusted AI [31], including seven requirements for transparency and accountability. The European AI Bill requires transparency and provision of information to users, which means transparency of operation by design and development of transparent operation so that the user can understand and control the processing process.

On the other hand, IEEE is currently examining IEEE P7001 (transparency of autonomous systems)[19] and it may have a certain level of influence over future standardization in terms of the definitions of terms and concept. The six transparency levels from 0 to 5 have been defined for the five types of stakeholders such as users, accident investigation committee, etc.. A higher number does not always mean that it is stricter. The certification of its compatibility is demonstrated through the pilot validation project called ECPAIS.

ISO includes terms and concepts related to transparency in the document TR24028 [6] at WG3 (trustworthiness) of ISO/IEC JTC 1/SC 42.

11.2.3. Fairness (bias)

EU AI high-level expert group (AI-HLEG) compiles high-level principles for problems of ELSI of AI including bias. The European AI Bill calls for data and data governance requirements that datasets be relevant, representative, error-free, and have appropriate statistical properties, as well as monitoring, detection, and correction of bias.

The above IEEE P7003 (algorithmic bias) [21] specifies a method to identify *negative bias* of both legally-prohibited discrimination based on race or gender and non-legal discrimination in the development of algorithms and keep bias within the acceptable range in the lifecycle from system planning to operation. A demonstration experiment of this method is under way in the pilot project (ECPAIS (Ethics Certification Program for Autonomous and Intelligent Systems)) to validate its conformity.

ISO/IEC JTC 1/SC 42/WG 3 (trustworthiness) is also preparing documents such as TR 24028 (Overview of trustworthiness in Artificial Intelligence)[6], TR 24027 (Bias in AI systems and AI aided decision making)[5].

11.2.4. Other quality aspects

Privacy [20], nudging, and other issues are being considered in the IEEE P7000 series, and governance and other issues are being considered in ISO/IEC JTC1/SC42. The European AI bill requires a risk management system, technical documentation, record-keeping, and human oversight, in addition to the requirements already mentioned. In addition to the existing

requirements, the report also calls for a risk management system, technical documentation, record-keeping, and human oversight.

12. (informative) Analytical information

This Chapter sorts out a process of analysis to draw out the characteristic axes of internal qualities listed mainly in Section 1.7 and Chapter 6 as reference information.

The content of this Chapter is informative.

12.1. Analysis of characteristic axes of internal qualities with respect to safety

First, a quality deterioration mode was identified with regard to external quality axes of safety. This quality deterioration occurs when *individual inference results of machine learning components to describe it in an extreme manner, vector values of neural-network results is not correct, favorable or desirable*. Thus, we analyzed a possibility that a machine learning component gives *undesirable* answers based on a concept similar to Fault Tree Analysis (FTA) using an abstract and binary tree failure mode on the assumption of abstract machine learning components. The analysis results are shown in Figure 21.

This analysis aims to comprehensively decompose the causes of failure modes. It should be noted that, when any misjudgment was made, it might be impossible to identify the cause without so-called oracle perspective. On the other hand, it is possible to reduce a possibility of all failure modes by taking measures for all causes of failure without oracle perspective. Of course, it is impossible to perfectly realize each paragraph. Moreover, errors included in training input data and various gaps such as mathematical issues in super-multidimensional space data are ignored intentionally in this analysis. However, if these items are focused on as a direction of the overall quality improvement process, it can make a sufficient contribution to the improvement of quality despite gaps.

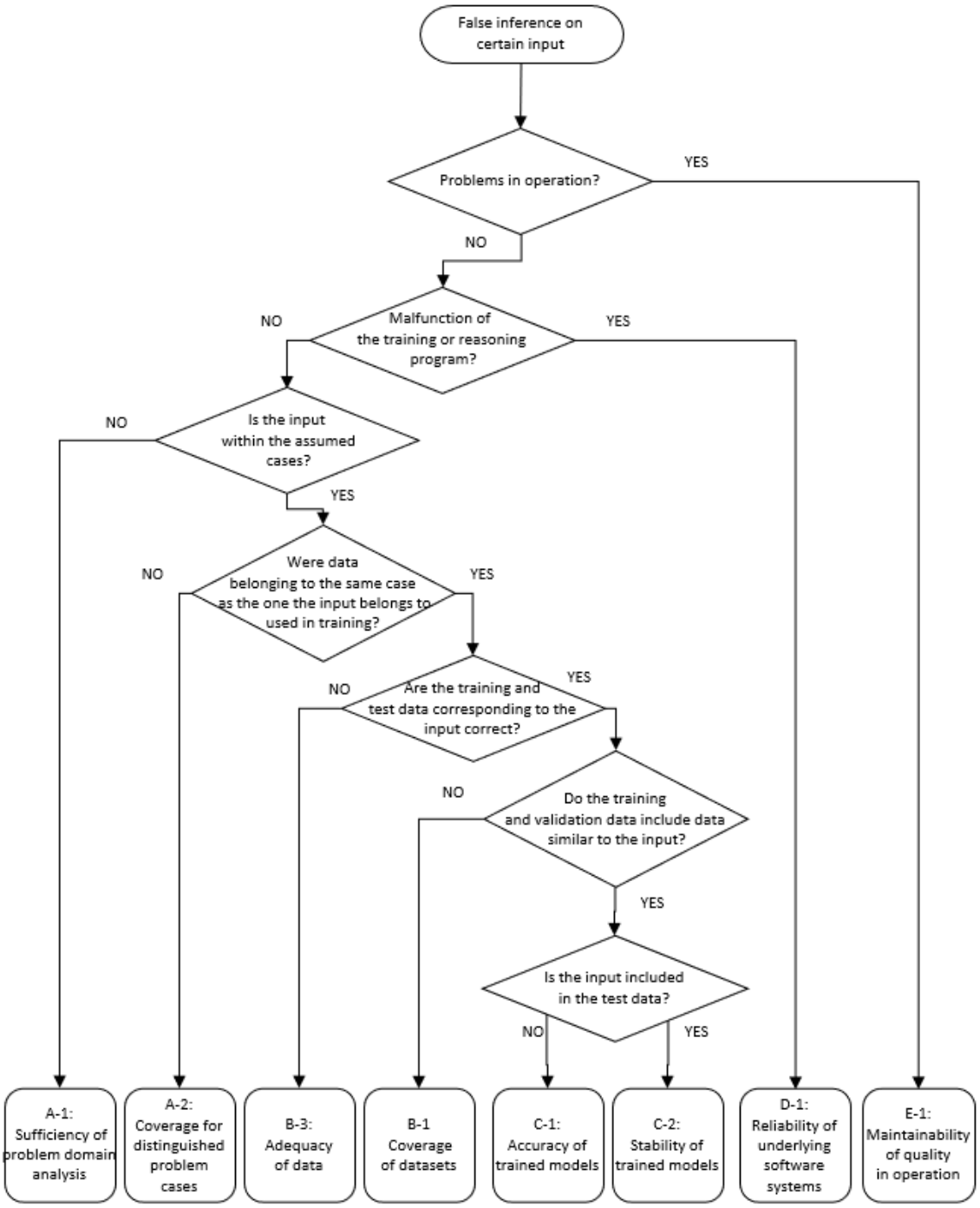

Figure 21: Example of analysis of failure mode with respect to machine learning

12.2. Analysis of quality management axes with respect to AI performance

AI performance was analyzed in a similar way based on the internal quality management axes with respect to *safety* mentioned in the previous section.

Strictly speaking, AI performance also deteriorates, when *individual inference results of*

machine learning components (to describe in an extreme manner, vector values of neural-network results) is not correct, favorable or desirable. A difference from risk avoidance is a difference in overall evaluation functions caused by different weighting of individual cases. Therefore, the same internal quality characteristic axes can be basically used as they are.

However, risk avoidance focuses on sufficient assignment of training data as measures for each anticipated risk case and does not take into account the balance of overall training data intentionally. This is because sufficient learning cannot be achieved by uniformly sampling training data especially from serious risk cases, although their frequency of occurrence is very low. From the viewpoint of AI performance, however, it is widely known that this type of *intentionally-biased reinforced learning data* may cause deterioration of overall performance. That is why we added *Uniformity of training datasets* listed in 6.4 as an internal quality axis which mainly envisions AI performance. The results are nine internal characteristics listed in Chapter 6.

13. (informative) Tables and figures

The content of this Chapter is informative.

13.1. Tables of relations between external quality characteristics and internal quality characteristics

The numbers and symbols in the tables refer to the levels of check items listed in *Requirements for quality levels* of applicable sections. The symbol + means that additional examinations will be made in the future. The bolded items mean that they require special attention.

AISL	0	0.1	0.2	1	2	3	4
A-1 Sufficiency of problem domain analysis		1	2	3	+	+	+
A-2 Sufficiency of data design		1	2	3	+	+	+
B-1 Coverage of datasets		1	2	3	+	+	+
B-2 Uniformity of datasets		S1	S2	S2	+	+	+
B-3 Adequacy of data		1	2	3	+	+	+
C-1 Correctness of trained models		1	2	3	+	+	+
C-2 Stability of trained models		1	2	3	+	+	+
D-1 Reliability of underlying software system		1	2	3	+	+	+
E-1 Maintainability of qualities in operation		1	2	3	+	+	+

AIPL/AIFL	PL 0	PL 1	PL 2	FL 0	FL 1	FL 2
A-1 Sufficiency of problem domain analysis		1	2		1	2
A-2 Sufficiency of data design		1	2		1	2
B-1 Coverage of datasets		1	2		1	2
B-2 Uniformity of datasets		E1	E2		***E2***	E2
B-3 Adequacy of data		1	2		1	2
C-1 Correctness of trained models		1	2		1	2
C-2 Stability of trained models		1	2		1	2
D-1 Reliability of underlying software system		1	2		1	2
E-1 Maintainability of qualities in operation		1	2		2	3

14. Bibliography

14.1. International standards

[1] ISO 13849-1:2015: Safety of machinery — Safety-related parts of control systems — Part 1: General principles for design.

[2] ISO/IEC/IEEE 15288:2015: Systems and software engineering — System life cycle processes.

[3] ISO/IEC 15408-1:2009: Information technology — Security techniques — Evaluation criteria for IT security — Part 1: Introduction and general model.

[4] ISO/IEC IS 22989:2022: Information technology — Artificial intelligence — Artificial intelligence concepts and terminology.

[5] ISO/IEC TR 24027:2021: Information technology — Artificial Intelligence (AI) — Bias in AI systems and AI aided decision making.

[6] ISO/IEC TR 24028:2020: Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence.

[7] ISO/IEC 25010:2011: Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models.

[8] ISO/IEC 25012:2008: Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model.

[9] ISO 26262-1:2018: Road vehicles — Functional safety — Part 1: Vocabulary.

[10] ISO/IEC 27000:2018: Information technology — Security techniques — Information security management systems — Overview and vocabulary.

[11] ISO/IEC/IEEE 29119-4:2015: Software and systems engineering — Software testing — Part 4: Test techniques.

[12] IEC 61508-1:2010: Functional safety of electrical/electronic/programmable electronic safety-related systems - Part 1: General requirements.

[13] IEC 61508-3:2010: Functional safety of electrical/electronic/programmable electronic safety-related systems - Part 3: Software requirements.

[14] IEC 61508-4:2010: Functional safety of electrical/electronic/programmable electronic safety-related systems - Part 4: Definitions and abbreviations.

[15] IEC 62278:2002: Railway applications - Specification and demonstration of reliability, availability, maintainability and safety (RAMS).

[16] IEC TS 62443-1-1:2009: Industrial communication networks - Network and system

security - Part 1-1: Terminology, concepts and models.

[17] IEC 62443-2-1:2010: Industrial communication networks - Network and system security - Part 2-1: Establishing an industrial automation and control system security program.

[18] IEC TS 62998-1:2019: Safety of machinery - Safety-related sensors used for the protection of persons.

[19] IEEE P7001 - IEEE Standard for Transparency of Autonomous Systems.

[20] IEEE P7002 - IEEE Standard for Data Privacy Process.

[21] IEEE P7003 - Algorithmic Bias Considerations. An active project.

14.2. Documents/regulations from countries and international organizations

[22] Council for Science, Technology and Innovation, Cabinet Office of Japan. Social Principles of Human-Centric AI. March 2019.
https://www8.cao.go.jp/cstp/aigensoku.pdf (In Japanese),
https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf (tentative translation).

[23] United Nations Educational, Scientific and Cultural Organization (UNESCO), *Draft Text of the Recommendation on the Ethics of Artificial Intelligence*, 41 C/73 Annex, November 2021.

[24] Organisation for Economic Co-operation and Development (OECD). *OECD Principles on Artificial Intelligence*. May 2019. https://www.oecd.org/going-digital/ai/principles/

[25] European Commission. *Regulation of the European parliament and of the council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts*. April 2021.
https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence

[26] European Commission. *Regulation* (EU) 2016/679 *of The European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*.

[27] European Commission. *Communication Artificial Intelligence for Europe*. 2018.
https://digital-strategy.ec.europa.eu/en/library/communication-artificial-intelligence-europe.

[28] European Commission. *The White Paper on Artificial Intelligence – A European approach to excellence and trust*. February 2020.
https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-

approach-excellence-and-trust_en

[29] European Commission. High-Level Expert Group on Artificial Intelligence. 2018. https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

[30] The High-Level Expert Group on Artificial Intelligence, European Commission. *Ethics guidelines for trustworthy AI*. April 2019. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[31] The High-Level Expert Group on Artificial Intelligence, European Commission. *The Assessment List for Trustworthy Artificial Intelligence (ALTAI)*. July 2020. https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

[32] Ministry of Economy, Trade and Industry, Japan. Contract Guidelines on Utilization of AI and Data. June 2018. In Japanese https://www.meti.go.jp/press/2018/06/20180615001/20180615001-3.pdf

[33] Ministry of Economy, Trade and Industry, Study Group for Architecture of AI Social Implementation. AI Governance of our country - version 1.0. January 2021. In Japanese https://www.meti.go.jp/press/2020/01/20210115003/20210115003-1.pdf

[34] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition*. IEEE, 2019. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html

[35] Ministry of Economy, Trade and Industry . Cyber Physical Security Measures Framework Ver1.0. April 2019. In Japanese. https://www.meti.go.jp/press/2019/04/20190418002/20190418002-2.pdf

14.3. Official standards and forum standards

[36] National Institute of Standards and Technology (United States of America). Draft NIST IR 8269: *A Taxonomy and Terminology of Adversarial Machine Learning*. October 2019. https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8269-draft.pdf

[37] National Institute of Standards and Technology (United States of America). NIST SP 800-30: *Guide for Conducting Risk Assessments*. September 2012. https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-30r1.pdf

[38] National Institute of Standards and Technology (United States of America). *Cyber Security Framework Version 1.1*. April 2018. https://doi.org/10.6028/NIST.CSWP.04162018

[39] MITRE. *Adversarial ML Threat Matrix*. https://github.com/mitre/advmlthreatmatrix

[40] MITRE, *Common Platform Enumerations*. http://cpe.mitre.org/

[41] MITRE, *Common Vulnerability Enumerations*. https://cve.mitre.org/

[42] Payment Card Industry Security Standards Council. Payment Card Industry Data Security Standard (PCI DSS). https://www.pcisecuritystandards.org/.

[43] Consortium of Quality Assurance for Artificial-Intelligence-based products and services (QA4AI). Guidelines for Quality Assurance of Machine leaning based Artificial Intelligence – 2020.08 version. August 2020. In Japanese. http://qa4ai.jp/download/.

[44] European Union Agency for Cybersecurity (ENISA). Artificial Intelligence Cybersecurity Challenges. December 2020. https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges

[45] European Union Agency for Cybersecurity (ENISA). Securing Machine Learning Algorithms. December 2021. https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms

[46] National Information Security Standardization Technical Committee of the People's Republic of China. Information security technology: Assessment specification for machine learning algorithms (draft). August 2021.

14.4. Research articles

[47] Martín Abadi, Andy Chu, Ian J. Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep Learning with Differential Privacy. In *Proc. of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS'16)*, pp. 308–318, 2016.

[48] Alessandro Acquisti, Curtis Taylor, and Lian Wagman, The Economics of Privacy, Journal of Economic Literature 54(2), pp.442-492, 2016.

[49] Charu C. Aggarwal, On k-Anonymity and the Curse of Dimensionality, Proc. *31st VLDB*, pp.901-909, 2005.

[50] Charu C. Aggarwal, *Outlier Analysis (2ed.)*, Springer 2017.

[51] Naveed Akhtar, and Ajmal Mian. Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. *IEEE Access* 6, pp. 14410–14430, 2018. https://arxiv.org/pdf/1801.00553.pdf

[52] Ahmed Aldahdooh, Wassim Hamidouche, Sid Ahmed Fezza, and Olivier Deforges. Adversarial example detection for DNN models: a review and experimental comparison. Artif Intell Rev (2022). https://arxiv.org/abs/2105.00203

[53] P. Ammann, and J. Offutt. *Introduction to Software Testing*, Cambridge University Press, 2008.

[54] Giuseppe Ateniese, Luigi V. Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *International Journal of Security and Networks*, 10(3):137–150, 2015.

[55] Nuttapong Attrapadung, Koki Hamada, Dai Ikarashi, Ryo Kikuchi, Takahiro Matsuda, Ibuki Mishina, Hiraku Morita, and Jacob C. N. Schuldt. Adam in Private: Secure and Fast Training of Deep Neural Networks with Adaptive Moment Estimation. Proceedings on Privacy Enhancing Technologies, 2022 (4), pp.747–768, 2022.

[56] Eugene Bangdasaryan and Vitaly Shmatikov, Differential Privacy Has Disparate Impact on Model Accuracy, arXiv:1905.12101v2, 2019.

[57] Guy Barash, Eitan Farchi, Ilan Jayaraman, Orna Raz, Rachel Tzoref-Brill, and Marcel Zalmanovici. Bridging the gap between ML solutions and their business requirements using feature interactions. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2019)*, pp. 1048–1058, August 2019. https://dl.acm.org/citation.cfm?id=3340442

[58] E. T. Barr, M. Harman, P. McMinn, M. Shahbaz, and S. Yoo. The Oracle Problem in Software Testing: A Survey. In *IEEE Transactions on Software Engineering*, 41(5):507–525, 2015.

[59] Battista Biggio, and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, Vol. 84, pp. 317-331, 2018.

[60] Matt Bishop, *Computer Security – Art and Science*, Addison Wesley 2003.

[61] Kallista Bonawitz, Peter Kairouz, Brendan McMahan, and Daniel Ramage, Federated Learning and Privacy, Comm. ACM, 65(4), pp.90-97. 2022.

[62] Akhilan Boopathy, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. CNN-Cert: An Efficient Framework for Certifying Robustness of Convolutional Neural Networks. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI 2019)*, pp.3240–3247, 2019.

[63] Eitan Borgnia, Valeriia Cherepanova, Liam Fowl, Amin Ghiasi, Jonas Geiping, Micah Goldblum, Tom Goldstein, and Arjun Gupta. Strong Data Augmentation Sanitizes Poisoning and Backdoor Attacks Without an Accuracy Tradeoff. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'21)*, pp.3855-3859, 2021. https://arxiv.org/abs/2011.09527

[64] Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallash, Reza Shokri, and Florian Tramer, What Does it Means for a Language Model to Preserve Privacy?, arXiv:2202.05520v2, 2022.

[65] Miles Brundage, et al., Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims, arXiv:2004.07213v2, 2020.

[66] H. Brendan McMahan, Galen Andrew, Ulfar Erlingsson, Steve Chien, Ilya Mironov, Nocolas Papernot, and Peter Kairouz, A General Approach to Adding Differential Privacy to Iterative Training Procedures, arXiv:1812.06210v2, 2019.

[67] T. Byun, V. Sharma, A. Vijayakumar, S. Rayadurgam, and D. Cofer, Input Prioritization for Testing Neural Networks, Proc. AITest, pp. 63-70, and arXiv:1901.03768

[68] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, December 2017. https://papers.nips.cc/paper/6988-optimized-pre-processing-for-discrimination-prevention.pdf.

[69] Nicolas Carlini, Chang Liu, Ulfar Erlingsson, Jernej Kos, and Dawn Song, The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks, arXiv:1802.08232v3, 2019.

[70] Nicolas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingson, Alina Oprea, and Coloin Raffel, Extracting Training Data from Large Language Models, arXiv:2012.07805v2, 2021.

[71] Nicholas Carlini, and David Wagner. Towards Evaluating the Robustness of Neural Networks. In *Proceedings of the 38th IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, 2017.

[72] Saeed Mahloujifar, Esha Ghosh, Melissa Chase. Property inference from poisoning. In Proceedings of the 2022 IEEE Symposium on Security and Privacy (SP), pp. 1569-1569, 2022. https://arxiv.org/abs/2101.11073

[73] Chakraborty, Anirban, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. A Survey on Adversarial Attacks and Defences. *CAAI Transactions on Intelligence Technology* , 6 (1), pp.25–45. https://doi.org/10.1049/cit2.12028

[74] Hongyan Chang and Reza Shokri, On the Privacy Risks of Algorithmic Fairness, arXiv:2011.03731, 2021.

[75] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy and Biplav Srivastava. Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering. In *Proceedings of the Workshop on Artificial Intelligence Safety 2019*, 2019. http://ceur-ws.org/Vol-2301/paper_18.pdf

[76] T. Y. Chen, S. C. Chung, and S. M. You. Metamorphic Testing – A New Approach for Generating Next Test Cases, Technical Report HKUST-CS98-01, Department of Computer Science, The Hong Kong University of Science and Technology, 1998.

[77] T. Y. Chen, F.-C. Kuo, H. Liu, P.-L. Poon, D. Towey, Y. H. Tse, and Z. Q. Zhou. Metamorphic Testing: A Review of Challenges and Opportunities. *ACM Computing Surveys*, 51(1):1–27, 2018.

[78] S. Chiappa, and W. S. Isaac. A Causal Bayesian Networks: Viewpoint on Fairness. In *Privacy and Identity Management: Fairness, Accountability, and Transparency in the Age of Big Data*, IFIP Advances in Information and Communication Technology, vol. 547, 2019.

[79] Mark Coeckelbergh, *AI Ethics*, The MIT Press Essential Knowledge Series.

[80] Jeremy M Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified Adversarial Robustness

via Randomized Smoothing. *The 36th International Conference on Machine Learning (ICML 2019)*, 2019.

[81] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. RobustBench: A Standardized Adversarial Robustness Benchmark. 2021. http://arxiv.org/abs/2010.09670, https://robustbench.github.io

[82] George Deckert, NASA Hazard Analysis Process. Johnson Space Center, National Aeronautics and Space Administration, 2010. https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20100040678.pdf.

[83] Ann-Kathrin Dombrowski, Maximilian Alber, Christopher J. Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In *Proceedings of the Annual Conference on Neural Information Processing Systems 2019 (NeurIPS'19)*, pp. 13567–13578, 2019. https://arxiv.org/pdf/1906.07983.pdf

[84] Yizhen Dong, Peixin Zhang, Jingyi Wang, Shuang Liu, Jun Sun, Jianye Hao, Xinyu Wang, Li Wang, Jin Song Dong, and Dai Ting. There is Limited Correlation between Coverage and Robustness for Deep Neural Networks. arXiv:1911.05904 [cs.LG]. https://arxiv.org/abs/1911.05904

[85] Cynthia Dwork, Differential Privacy, In *Proc. ICALP'06*, pp.1-12, 2006.

[86] Cynthia Dwork and Aaron Roth, The Algorithmic Foundations of Differential Privacy, *Foundations and Trends in Theoretical Computer Science*, 9(3-4), pp.211-407, 2014.

[87] S. Elbaum and D. S. Rosenblum. Known Unknowns – Testing in the Presence of Uncertainty, In *Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering (FSE 2014)*, pp. 833–836, 2014.

[88] Arisa Ema (ed.). Small Featured Articles "From AI principles to implementation: introduction of international activities". Artificial Intelligence 36(2), The Japanese Society for Artificial Intelligence, March 2021..In Japanese

[89] Ulfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova, RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response, arXiv:1407.6981v2, 2014.

[90] Itay P. Fainmesser, Andrea Galeotti, and Ruslan Momot, Digital Privacy, HEC Paris Research Paper, 2021.

[91] E. R. Faria, J. Gama, and A. C. Carvalho. Novelty detection algorithm for data streams multi class problems, In *Procedings of the 28th annual ACM symposium on applied computing*, pp. 795–800, 2013.

[92] Y. Feng, Q. Shi, X. Gao, J. Wan, C. Fang, and Z. Chen, DeepGini: Prioritizing Massive Tests to Enhance the Robustness of Deep Neural Networks, Proc. 29th ISSTA, pp. 177-188, and arXiv:1903.00661v2, 2020.

[93] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model Inversion Attacks that

Exploit Confidence Information and Basic Countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS '15),* pp. 1322–1333, 2015. https://dl.acm.org/doi/10.1145/2810103.2813677

[94] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, A survey on concept drift adaptation. In *ACM computing surveys*, 46(4):1–37, April 2014.

[95] Karan Ganju, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov, Property Inference Attacks on Fully Connected Neural Networks using Permutation Invariant Representations, In Proc. 25th ACM CCS, pp. 619-633, 2018.

[96] Simson Garfinkel, John M. Abowd, and Christian Martindale, Understanding Database Reconstruction Attacks on Public Data, ACM Queue, September-October, pp.1-26, 2018.

[97] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, pp. 2672–2680, 2014.

[98] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. In *Proc. of the 3rd International Conference on Learning Representations (ICLR'15)*, 2015. https://arxiv.org/pdf/1412.6572.pdf

[99] F. Harel-Canada, L. Wang, M. A. Gulzar, Q. Gu, and M. Kim. Is Neuron Coverage a Meaningful Measure for Testing Deep Neural Networks? In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2020)*, pp. 851–862, 2020.

[100] Ori Heffetz and Katrina Ligett, Privacy and Data-Based Research, Journal of Economic Perspectives 28(2), 75-98, 2014.

[101] Mike Hintze, Viewing the GDPR through a De-Identification Lens: A Tool for Clarification, Compliance, and Consistency, International Data Protection and Law, 8(1), pp.86-101, 2018.

[102] Christina Ilvento. Metric Learning for Individual Fairness. arXiv:1906.00250 [cs.LG]. https://arxiv.org/abs/1906.00250

[103] L. Inozemtseva, and R. Holmes. Coverage is not Strongly Correlated with Test Suite Effectiveness, In *Proceedings of the 36th International Conference on Software Engineering (ICSE 2014)*, pp. 435–455, 2014.

[104] Bargav Jayaraman and David Evans, Evaluating Differentially Private Machine Learning in Practice, In Proc. 28th USENIX Security Symposium, and also arXiv:1902.08874v4, 2019.

[105] Jinyuan Jia, Xiaoyu Cao, and Neil Zhenqiang Gong. Intrinsic Certified Robustness of Bagging against Data Poisoning Attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'21)*, 35(9), pp.7961-7969, 2021. http://arxiv.org/abs/2008.04495

[106] J. Jia, A. Salem, M. Backes, Y. Zhang, and N. Z. Gong. MemGuard: Defending against

Black-Box Membership Inference Attacks via Adversarial Examples. In *Proc. of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS'19)*, pp. 259–274, 2019.

[107] Mika Juuti, Sebastian Szyller, Samuel Marchal, and N. Asokan. PRADA: Protecting Against DNN Model Stealing Attacks. In *Proc. of the IEEE European Symposium on Security and Privacy (Euro S&P'19)*, pp.512–527, 2019. https://arxiv.org/pdf/1805.02628.pdf

[108] Peter Kairouz, M. Brendan McMahan et al., Advances and Open Problems in Federated Learning, arXiv: 1912.04977v3, 2021.

[109] Guy Katz, Clark Barrett, David Dill, Kyle Julian, and Mykel Kochenderfer. Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks. In *International Conference on Computer-Aided Verification (CAV)*, 2017.

[110] F. Kamiran, A. Karim, and X. Zhang. Decision theory for discrimination-aware classification. In *Proceedings of the IEEE International Conference on Data Mining (ICDM 2012)*, 2012.

[111] Faisal Kamiran, and Toon Calders. Data preprocessing techniques for classification without discrimination. In *Knowledge and Information Systems*, 33(1):1–33, 2012.

[112] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, 7524:35–50, 2012. https://link.springer.com/content/pdf/10.1007%2F978-3-642-33486-3_3.pdf

[113] J. Kim, R. Feldt, and S. Yoo. Guiding Deep Learning System Testing Using Surprise Adequacy, In *Proceedings of the 41st International Conference on Software Engineering (ICSE '19)*, pp. 1039–1049, 2019.

[114] Rob Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *The 14th International Joint Conference on Artificial Intelligence*, 2(12):1137–1143, 1995.

[115] Keita Kinjo. Fair causal effect using Social Welfare Function. *The 35th Annual Conference of the Japanese Society for Artificial Intelligence*, 2021

[116] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet Classification with Deep Convolutional Neural Networks. In *Procedings of the Advances in Neural Information Processing Systems 25 (NIPS 2012*), pp.1097–1105, 2012.

[117] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial Examples in the Physical World. In *Proc. of the 5th International Conference on Learning Representations (ICLR) Workshop*, arXiv:1607.02533 [cs.CV], July 2016, https://arxiv.org/pdf/1607.02533.pdf

[118] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified Robustness to Adversarial Examples with Differential Privacy. *The IEEE Symposium on Security and Privacy* (SP), 2019.

[119] S. Lee, J. Kim, J. Jun, J. Ha, and B. Zhang. Overcoming catastrophic forgetting by incremental moment matching. In *Proc. of the Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pp. 4652–4662, 2017.

[120] Yiming Li, Baoyuan Wu, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor Learning: A Survey. arXiv:2007.08745 [cs.CR], July 2020, https://arxiv.org/pdf/2007.08745

[121] P. Lindstrom, B. M. Namee, and S. J. Delany. Drift detection using uncertainty distribution divergence. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 604–608, 2011.

[122] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Proceedings of the International Symposium on Research in Attacks, Intrusions, and Defenses (RAID'18)*, pp. 273–294, 2018. https://arxiv.org/abs/1805.12185

[123] X. Liu, M. Cheng, H. Zhang, and C. Hsieh. Towards robust neural networks via random self-ensemble. In *Proceedings of the European Conference on Computer Vision (ECCV 2018)*, 2018.

[124] Yugeng Liu, Rui Wen, Xinlei He, Ahmed Salem, Zhikun Zhang, Michael Backes, Emiliano De Cristofaro, Mario Fritz, and Yang Zhang, ML-Doctor: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models, USENIX Security Symposium 2022.

[125] Yunhui Long, Vincent Bindschaedler, Lei Wang, Diyue Bu, Xiaofeng Wang, Haixu Tang, Carl A. Gunter, and Kai Chen, Understanding Membership Inferences on Well-Generalized Learning Models, arXiv:1802.04489, 2018.

[126] Lei Ma, Felix Juefei-Xu, Fuyuan Zhang, Jiyuan Sun, Minhui Xue, Bo Li, Chunyang Chen, Ting Su, Li Li, Yang Liu, Jianjun Zhao, and Yadong Wang. DeepGauge: Multi-Granularity Testing Criteria for Deep Learning Systems, In *Proceedings of the 33rd IEEE/ACM International Conference on Automated Software Engineering (ASE 2018)*, pp. 120–131, 2018.

[127] Shiqing Ma, Yingqi Liu, Guanhong Tao, Wen-Chuan Lee, and Xiangyu Zhang. NIC: Detecting Adversarial Samples with Neural Network Invariant Checking. In *Network and Distributed Systems Security Symposium (NDSS)*, 2019.

[128] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *the Sixth International Conference on Learning Representations (ICLR 2018)*, 2018.

[129] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. arXiv:1908.09635 [cs.LG], 2019.

[130] Ninareh Mehrabi, Muhammad Naveed, Fred Morstatter, and Aram Galstyan. Exacerbating Algorithmic Bias through Fairness Attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'21)*, pp. 8890-8938, 2021.

[131] B.P. Miller, L. Fredricksen, and B. So. An Empirical Study of the Reliability of UNIX Utilities, *Communications of the ACM*, 33(12):32–44, 1990.

[132] Fatemehsadat Mireshghallah, Mohammadkazem Taram, Praneeth Vepakomma, Abhishek Singh, Ramesh Raskar, and Hadi Esmaeilzadeh, Privacy in Deep Learning: A Survey, arXiv:2004.12254v5, 2020.

[133] Ilya Mironov, Renyi Differential Privacy, arXiv:1702.07476v3, 2017.

[134] Payman Mohassel and Yupeng Zhang. SecureML: A System for Scalable Privacy-Preserving Machine Learning. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (S&P'17), pp.19-38, 2017.

[135] I. Morikawa Frontiers of Machine Learning Security Research. IEICE Fundamentals and Boundary Society Fundamentals Review, 2021, vol. 15, no. 1, pp. 37-46, 2021.In Japanese

[136] Sasi Kumar Murakonda and Reza Shokri, ML Privacy Meter: Aiding Regulatory Compliance by Quantifying the Privacy Risks of Machine Learning, arXiv:2007.09339, 2020.

[137] Moin Nadeem, Anna Bethke, and Siva Reddy, StereoSet: Measuring Stereotypical Bias in Pretrained Language Models. arXiv:2004.09456, 2020.

[138] Shin Nakajima, Software Testing under Dataset Diversity, Computer Software, 35(2):26-32, 2018 In Japanese

[139] Shin Nakajima. Distortion and Faults in Machine Learning Software, In *Proc. 9th International Workshop on SOFL + MSVL for Reliability and Security*, pp. 29–41, 2020, and also arXiv:1911.11596 [cs.LG], 2019.

[140] Shin Nakajima, “Quality Issues in Machine Learning from Software Engineering Viewpoints”, Maruzen Publisher, ISBN 978-4-612-30573-7, November 2020 In Japanese

[141] Shin Nakajima, Distortion Metrics for Trained Machine Learning Models, IEICE Technical Report, SS2019-44, March 2020. In Japanese

[142] Shin Nakajima. Statistical Partial Oracle for Machine Learning Software Testing, In *Proceedings of the 10th International Workshop on SOFL + MSVL for Reliability and Security*, 2021.

[143] Shin Nakajima and Takako Nakatani, AI Extension of SQuaRE Data Quaity Model, In Proc. 1st IEEE Workshop on FPDRE, 2022.

[144] Shin Nakajima and Tsong Yueh Chen. Generating Biased Dataset for Metamorphic Testing of Machine Learning Programs, In *Proceedings of The 31st IFIP International Conference On Testing Software And Systems (IFIP–ICTSS 2019)*, pp. 56–64, 2019.

[145] Takako Nakatani, Shin Nakajima “Software Engineering”, Foundation for the Promotion of the Open University of Japan, ISBN 978-4-595-14119-5, March 2019. In Japanese

[146] Arvind Narayanan and Vitaly Shmatikov, Robust De-anonymization of Large Sparce Datasets, In *Proc. SSP*, pp.111-125, 2008.

[147] Milad Nasr, Reza Shokri, and Amir Houmansadr, Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning, arXiv:1812.00910v2, 2020.

[148] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, et al. Adversarial Robustness Toolbox v1.0.0. 2019. http://arxiv.org/abs/1807.01069, https://adversarial-robustness-toolbox.readthedocs.io

[149] Steven Nowlan, and Geoffrey Hinton. Simplifying neural networks by soft weight-sharing. *Neural Computation*, 4(4), 1992.

[150] L. Oneto, N. Navarin, A. Sperduti, and D. Anguita. Recent Trends in Learning From Data. *Tutorials from the INNS Big Data and Deep Learning Conference (INNSBDDL2019)*, January 2020.

[151] Seong Joon Oh, Bernt Schiele, Mario Fritz. Towards Reverse-Engineering Black-Box Neural Networks. In Proceedings of the 6th International Conference on Learning Representations (ICLR'18), 2018.

[152] Tribhuvanesh Orekondy, Bernt Schiele, Mario Fritz. Knockoff Nets: Stealing Functionality of Black-Box Models. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'19), pp.4954-4963, 2019.

[153] Tinghui Ouyang, Vicent Sanz Marco, Yoshinao Isobe, Hideki Asoh, Yutaka Oiwa, and Yoshiki Seo. Corner Case Data Description and Detection. arXiv:2101.02494v2 [cs.LG]. https://arxiv.org/pdf/2101.02494.pdf

[154] Nicolas Papernot, Patrick D. McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security (AsiaCCS 2017)*, pp.506-519, 2017.

[155] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael P. Wellman, SoK: Security and Privacy in Machine Learning, In *Proc. IEEE Euro S&P*, pp.399-414, 2018.

[156] Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow et al. Technical Report on the CleverHans v2.1.0 Adversarial Examples Library. http://arxiv.org/abs/1610.00768 https://github.com/cleverhans-lab/cleverhans

[157] Arpita Patra and Ajith Suresh. BLAZE: Blazing Fast Privacy-Preserving Machine Learning. In Proceedings of the 27th Annual Network and Distributed System Security Symposium (NDSS'20), 2020

[158] Judea Pearl. Understanding Simpson's Paradox. *The American Statistician*, 68(1):8–13, February 2014. Also UCLA Cognitive Systems Laboratory Technical Report R-414, University of California, Los Angeles, December 2013.

[159] Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana. DeepXplore: Automated Whitebox Testing of Deep Learning Systems, *The 26th Symposium on Operating Systems Principles (SOSP 2017)*, pp. 1–18, 2017.

[160] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. On fairness and calibration. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17)*, December 2017.

[161] Jules Polonetsky, Omer Tene, and Kelsey Finch, Shades of Gray: Seeing the Full Spectrum of Practical Data De-identification, Santa Clara Law Review, 56(3), pp.593-629, 2016.

[162] Maria C. Psaoletti and Alessandro Simonetta, Data Quality and GDPR, UINOFO, 2019.

[163] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A Python toolbox to benchmark the robustness of machine learning models. https://arxiv.org/abs/1707.04131 https://github.com/bethgelab/foolbox

[164] D. M. dos Reis, P. Flach, S. Matwin, and G. Batista. Fast unsupervised online drift detection using incremental Kolmogorov Smirnov test. In *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1545–1554, 2016.

[165] Lucas Rosenblatt, Xiaoyan Liu, Samira Pouyanfar, Eduardo de Leon, Anuj Desai, and Joshua Allen, Differentially Private Synthetic Data: Applied Evaluations and Enhancements, arXiv:2011.05537, 2020.

[166] Elan Rosenfeld, Ezra Winston, Pradeep Ravikumar, and J. Zico Kolter. Certified Robustness to Label-Flipping Attacks via Randomized Smoothing. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*, pp.8230–8241, 2020. https://arxiv.org/abs/2002.03018

[167] G. Rothermel, R.H. Untch, and M.J. Harrold, Prioritizing Test Cases for Regression Testing, IEEE TSE 27(10), pp. 929-948, 2001.

[168] Pierangela Samarati and Latanya Sweeney, Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression, 1998.

[169] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes, ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models, arXiv:1806.01246, 2018.

[170] Oscar Schwartz. 2019. In 2016 Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation. IEEE Spectrum 11 (2019).

[171] T. S. Sethi, and M. Kantardzic. On the reliable detection of concept drift from streaming unlabeled data. *Expert Systems with Applications*, 82:77–99, 2017.

[172] B. Settles. Active Learning Literature Survey. Technical Report #1648, Computer Science Department, University of Wisconsin-Madison, 2009.

[173] S. Shalev-Shwartz. Online learning and online convex optimization, *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

[174] Jacson Rodrigues Correia da Silva, Rodrigo Ferreira Berriel, Claudine Badue, Alberto Ferreira de Souza, Thiago Oliveira-Santos. Copycat CNN: Stealing Knowledge by Persuading Confession with Random Non-Labeled Data. In Proceedings of the International Joint Conference on Neural Networks (IJCNN'18), pages 1–8, 2018.

[175] Reza Shokri, Marco Stronati, Congzheng Song and Vitaly Shmatikov. Membership Inference Attacks Against Machine Learning Models. In *Proceedings of the 2017 IEEE Symposium on Security and Privacy (S&P'17)*, pp. 3–18, 2017. https://arxiv.org/pdf/1610.05820.pdf

[176] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In *Proc. of the AAAI/ACM Conference on AI, Ethics, and Society 2020 (AIES'20)*, pp.180–186, 2020. https://arxiv.org/pdf/1911.02508.pdf

[177] David Solans, Battista Biggio, Carlos Castillo. Poisoning Attacks on Algorithmic Fairness. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD' 20), Part I*, p.162-177, 2020.

[178] Congzheng Song and Vitaly Shmatikov, Overlearning Reveals Sensitive Attributes, Proc. ICLR 2020, and also arXiv:1905.11742v3, 2020.

[179] Congzheng Song, Thomas Ristenpart, Vitaly Shmatikov. Machine Learning Models that Remember Too Much. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, (CCS'17)*, pp. 581-601, 2017.

[180] Liwei Song, Reza Shokri, Prateek Mittal. Membership Inference Attacks against Adversarially Robust Deep Learning Models. In *Proceedings of Deep Learning and Security Workshop (DLS)*, May 2019.

[181] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting, *Journal of Machine Learning Research* 15(56):1929–1958, 2014.

[182] Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. Certified Defenses for Data Poisoning Attacks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, pp.3520–3532, 2017. https://arxiv.org/abs/1706.03691

[183] Latanya Sweeney, Weaving Technology and Policy Together to Maintain Confidentiality, The Journal of Law, Medicine and Ethics 25(2-3), pp.98-110, 1997.

[184] Latabya Sweeney, k-Anonymity: A Model for Protecting Privacy, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5), pp.557-570, 2002.

[185] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, Rob Fergus. Intriguing properties of neural networks. arXiv preprint

arXiv:1312.6199 (2013). https://arxiv.org/abs/1312.6199

[186] H. Tian, M. Yu, and W. Wang. Continuum: A platform for cost aware, low latency continual learning, In *Proceedings of the ACM Symposium on Cloud Computing*. pp. 26–40, 2018.

[187] Y. Tian, K. Pei, S. Jana, and B. Ray. DeepTest: Automated Testing of Deep-Neural-Network-driven Autonomous Cars. In *Proceedings of the 40th International Conference on Software Engineering (ICSE 2018)*, pp. 303–314, 2018.

[188] Vincent Tjeng, Kai Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. In *International Conference on Learning Representations*, 2019.

[189] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing Machine Learning Models via Prediction APIs. In *Proceedings of the 25th USENIX Security Symposium (USENIX Security'16)*. https://arxiv.org/pdf/1609.02943.pdf

[190] Guillermo Valle-Pérez and Ard A. Louis. Generalization bounds for deep learning. arXiv:2012.04115v2, 2020. https://arxiv.org/abs/2012.04115

[191] A. Vostrikov and S. Chernyshev. Training sample generation software. In *Intelligent Decision Technologies 2019*, Smart Innovation, Systems and Technologies (SIST), 143:145–151, 2019.

[192] Isabel Wagner and David Eckhoff, Technical Privacy Metrics: A Systematic Survey, ACM Computing Survey 51(3), Article 57, 2018.

[193] Binghui Wang, Neil Zhenqiang Gong. Stealing hyperparameters in machine learning. In Proceedings of IEEE Symposium on Security and Privacy (SP), pages 36–52, 2018.

[194] J. Wang, J. Chen, Y. Sun, X. Ma, D. Wang, J. Sun, and P. Cheng. RoBOT: Robustness-Oriented Testing for Deep Learning Systems, In *Proceedings of the 43rd International Conference on Software Engineering (ICSE 2021)*, 2021.

[195] Tianhao Wang, Jeremiah Blocki, Ninghui Li, and Somesh Jha, Locally Differentially Private Protocols for Frequency Estimation, *Proc. 26th USENIX Security Symposium*, pp.729-745, 2017.

[196] Zhibo Wang, Jingjing Ma, Xue Wang, Jiahui Hu, Zhan Qin, and Kui Ren. Threats to Training: A Survey of Poisoning Attacks and Defenses on Machine Learning Systems. *ACM Computing Surveys*, 2022. https://doi.org/10.1145/3538707

[197] Tsui-Wei Weng, Pin-Yu Chen, Lam Nguyen, Mark Squillante, Akhilan Boopathy, Ivan Oseledets, and Luca Daniel. PROVEN: Verifying Robustness of Neural Networks with a Probabilistic Approach. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, PMLR 97:6727–6736, 2019.

[198] Tsui-Wei Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Duane Boning, Inderjit S. Dhillon, and Luca Daniel. Towards Fast Computation of Certified Robustness for ReLU Networks. In *Proceedings of the 35th International Conference on Machine*

Learning, PMLR 80:5276-5285, 2018.

[199] E. J. Weyuker. On Testing Non-testable Programs, Computer Journal, 25(4):465–470, 1982.

[200] Eric Wong and J. Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. *The 35th International Conference on Machine Learning (ICML 2018)*, PMLR vol. 80, pp. 5283–5292, 2018.

[201] Min Wu, Matthew Wicker, Wenjie Ruan, Xiaowei Huang, and Marta Kwiatkowska. A Game-Based Approximate Verification of Deep Neural Networks with Provable Guarantees. In *Theoretical Computer Science,* 2019. https://doi.org/10.1016/j.tcs.2019.05.046

[202] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou, Differentially Private Generative Adversarial Network, arXiv: 1802.06739, 2018.

[203] Weilin Xu, David Evans, and Yanjun Qi. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. In *Network and Distributed Systems Security Symposium* (NDSS), 2018.

[204] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha, Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting, *IEEE Computer Security Foundations Symposium* (CSF), and also arXiv:1709.01604v5, 2018.

[205] Kota Yoshida and Takeshi Fujino. Hardware Security for Edge AI Devices. IEICE Fundamentals and Boundary Society Fundamentals Review, Vol. 15, No. 2, pp. 88-100, 2021. In Japanese

[206] B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating Unwanted Biases with Adversarial Learning. In *AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, December 2018. https://dl.acm.org/doi/pdf/10.1145/3278721.3278779.

[207] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals, Understanding Deep Learning Requires Rethinking Generalization, arXiv:1611.03530v2, 2017.

[208] Jie M. Zhang, Mark Harman, Lei Ma, and Yang Liu. Machine Learning Testing: Survey, Landscapes and Horizons. arXiv:1906.10742 [cs.LG]. https://arxiv.org/abs/1906.10742

[209] P. Zhang, Q. Dai, and P. Pelliccione. CAGFuzz: Coverage-Guided Adversarial Generative Fuzzing Testing of Deep Learning Systems. arXiv:1911.07931, 2019.

[210] P. Zhang, J. Wang, J. Sun, G. Dong, and X. Wang. White-box Fairness Testing through Adversarial Sampling. In *Proceedings of the 42nd International Conference on Software Engineering (ICSE 2020)*, pp. 1–12, 2020.

14.5. Miscellaneous

[211] Information-technology Promotion Agency, Japan . STAMP/STPA for beginners. June 2019.In Japanese https://www.ipa.go.jp/ikc/reports/20190329.html

[212] Security Center, Information-technology Promotion Agency, Japan. Guidance for quality assurance for connected worlds . June 2018.In Japanese https://www.ipa.go.jp/sec/publish/tn18-001.html

[213] Security Center, Information-technology Promotion Agency, Japan. Overview of Common Platform Enumeration (CPE)]. October 2008. In Japanese https://www.ipa.go.jp/security/vuln/CPE.html

[214] Security Center, Information-technology Promotion Agency, Japan. Overview of Common Vulnerability Enumeration (CVE)]. January 2009. In Japanese https://www.ipa.go.jp/security/vuln/CVE.html

[215] National Institute of Advanced Industrial Science and Technology. Report on Machine Learning Quality Evaluation and Improvement Techniques, Technical Report, Digital Architecture Research Center, Cyber Physical Security Research Center, Artificial Intelligence Research Center. In Japanese, DigiARC-TR-2021-02 / CPSEC-TR-2021002, 2021.

[216] National Institute of Advanced Industrial Science and Technology. Report on Machine Learning Quality Evaluation and Improvement Techniques (2nd Edition), Technical Report, Digital Architecture Research Center, Cyber Physical Security Research Center, Artificial Intelligence Research Center. In Japanese, DigiARC-TR-2022-06 and also CPSEC-TR-2022007, 2022.

[217] IBM Corporation. AI Fairness 360 - Resources. http://aif360.mybluemix.net/resources

[218] Information Comissioner's Office (United Kingdom), Guidance on the AI Auditing Framework – draft guidance for consultation, 2020.

[219] United States Census Bureau, Your Guide to the 2020 Census, 2020.

[220] United States Census Bureau, Disclosure Avoidance for the 2020 Census: An Introduction, 2021.

[221] National Institute of Advanced Industrial Science and Technology. Machine Learning Quality Management Reference Guide, Technical Report, Digital Architecture Research Center, Cyber Physical Security Research Center, Artificial Intelligence Research Center, In Japanese DigiARC-TR-2022-03 / CPSEC-TR-2022004, 2022

[222] Machine Learning Engineering Research Group. Machine Learning System Security Guidelines Version 1.0, 2022. In Japanese

[223] Taichi Kakinuma, The Impact of the Revision of the Copyright Law on AI Development. 2nd Symposium on Automatic Translation. March 2019 In Japanese https://h-bank.nict.go.jp/event/event_190306.html

15. Changes

15.1. The English version of the third edition (January 2023)

- Changes introduced in the Japanese version of the third edition have been incorporated.
- Section 7.6.1.7 has been reinstated together with three of the four bibliographic items, which had been added to the English version of the second edition but lost in the Japanese version of the third edition.

15.2. The Japanese version of the third edition (August 2022)

- The description of privacy as an external quality, including its analysis and handling methods, was considered and added.
- Security was positioned as an external quality, and its relationship with existing characteristics was organized, and the descriptions in the second edition were expanded.
- Reorganized the descriptions in the chapter on fairness.
- Updated other descriptions.
- Table numbers for Table 6 in Section 9.3 and all tables after that were corrected in Rev. 3.1.1.0077 issued on September 2022.

15.3. The English version of the second edition (Feb. 2022)

- Section 7.6.1 has been reorganized and extended with a new section, Section 7.6.1.7.
- Four new items have been added to the bibliography and cited in Section 7.6.1.

Editors and authors

Digital Architecture Research Center (DigiARC)/
Artificial Intelligence Research Center (AIRC)/
Cyber Physical Security Research Center (CPSEC),
National Institute of Advanced Industrial Science and Technology (AIST)

List of members of the Committee for machine learning quality management (FY2020-21)

Shin Nakajima	National Institute of Advanced Industrial Science and Technology/ National Institute of Informatics (chair)
Yoshiki Seo	National Institute of Advanced Industrial Science and Technology (vice-chair)
Takashi Egawa	NEC Corporation
Shintaro Fukushima	Toyota Motor Corporation
Chinami Hamatani	Ad-Sol Nissin Corporation
Kenichi Kobayashi	Fujitsu Laboratories Limited
Hiroshi Kuwajima	Denso Corporation
Yusei Nakashima	TechMatrix Corporation
Hideto Ogawa	Hitachi, Ltd.
Yutaka Oiwa	National Institute of Advanced Industrial Science and Technology
Tamao Okamoto	Panasonic Corporation
Naoto Sato	Hitachi, Ltd.
Tomomichi Suzuki	Tokyo University of Science
Shingo Takada	Keio University
Satoshi Tsuchiya	Fujitsu Limited
Naoya Wakamatsu	NEC Corporation
Atsushi Yamada	IBM Japan, Ltd.

List of authors of the Machine Learning Quality Management Guideline

- Japanese edition:
  - Overall structure, writing and editing: Yutaka Oiwa and Koichi Konishi (AIST)
  - Section 5.2: Chinami Hamatani (Ad-Sol Nisshin Corporation)
  - Section 7.6: Shin Nakajima (AIST)
  - Section 7.6.2: Yoshinao Isobe (AIST)
  - Section 7.8: Kenichi Kobayashi and Yoshihiro Okawa (Fujitsu Laboratories Limited)
  - Section 8: Yutaka Oiwa, (AIST), Chinami Hamatani (Ad-Sol Nisshin Corporation)
  - Section 9: Shin Nakajima (AIST)
  - Sections 10.1-10.5: Yutaka Oiwa, Yusuke Kawamoto, and Koichi Konishi (AIST), Kazumasa Miyake (SEI)
  - Section 10.6: Kazumasa Miyake (SEI)
  - Section 11.2: Contribution from Egawa Takashi (NEC Corporation) and Hiroshi Kuwajima (Denso)
  - Supervision: Members of
    - Committee for machine learning quality management
    - Guideline Taskforce
- English edition: members of the Guideline Taskforce
  editors: Yutaka Oiwa and Koichi Konishi (AIST)

List of members of the Guideline Taskforce (FY 2020/2021)

Yutaka Oiwa	National Institute of Advanced Industrial Science and Technology (leader and co-editor)
Koichi Konishi	National Institute of Advanced Industrial Science and Technology
Takashi Egawa	NEC Corporation
Shin Nakajima	National Institute of Informatics
Shintaro Fukushima	Toyota Motor Corporation
Chinami Hamatani	Ad-Sol Nissin Corporation
Yoshinao Isobe	National Institute of Advanced Industrial Science and Technology
Yusuke Kawamoto	National Institute of Advanced Industrial Science and Technology
Kenichi Kobayashi	Fujitsu Laboratories Ltd.
Hiroshi Kuwajima	Denso Corporation
Yosuke Motohashi	NEC Corporation
Kazumasa Miyake	Sumitomo Electric Industries, Ltd.
Takeshi Miyake	Cyber Creative Institute Company Limited
Yusei Nakashima	Techmatrix Corporation
Tamao Okamoto	Panasonic Corporation
Yoshiki Seo	National Institute of Advanced Industrial Science and Technology
Satoshi Tsuchiya	Fujitsu Limited