Machine Learning System Security Guidelines, Part II. “Risk Assessment”

Version 1.03
December 26, 2022

Editing Committee of Machine Learning System Security Guidelines
Security Working Group on Machine Learning System

Machine Learning Systems Engineering (MLSE)
Japan Society for Software Science and Technology

Contents

II-1. Introduction

Part II includes guidelines for machine learning system developers (AI developers) on an assessment method for self-analyzing security risks and vulnerabilities specific to machine learning. These guidelines are intended as reference information for AI developers (not mandatory). In these guidelines, “AI developers” are assumed to be general machine learning system developers who may not have machine learning security expertise. These guidelines correspond to the threat analysis by AI developers in the procedures described in the Part I of the “Machine Learning System Security Guidelines” and introduce concrete analysis methods. Please see “Appendix: An Overview of Detection Techniques for Machine Learning-Specific Attacks” for more information on how AI developers can consider attack detection techniques. Notably, the implementation example of the threat analysis described in these guidelines is the March 2022 version, and the described method must be reconsidered, including the possibility of being unable to respond if attack algorithms are advanced in the future. In addition, the implementation example was constructed by the authors of these guidelines and does not cover all published attacks.

II-2. Machine Learning Systems Covered in Part II

In this chapter, machine learning systems, which are focus of Part II, are explained.

II-2.1. Structure of the Machine Learning System

The machine learning system targeted in Part II is a system that uses machine learning. The machine learning processing part in the machine learning system generally comprises a training pipeline and an inference pipeline, represented in **Figure II- 1** and **Figure II- 2**. In some systems, training processing is performed externally and only the inference pipeline is comprised. Before the operation of the machine learning system, training processing is performed using much training data in a training pipeline to generate a trained model. Then, inference processing is performed using inference data and the trained model in an inference pipeline to obtain an inference result. Although the structure of a machine learning system is not necessarily identical to that of **Figure II- 1** and **Figure II- 2**, the contents of Part II can be applied to many machine learning systems.

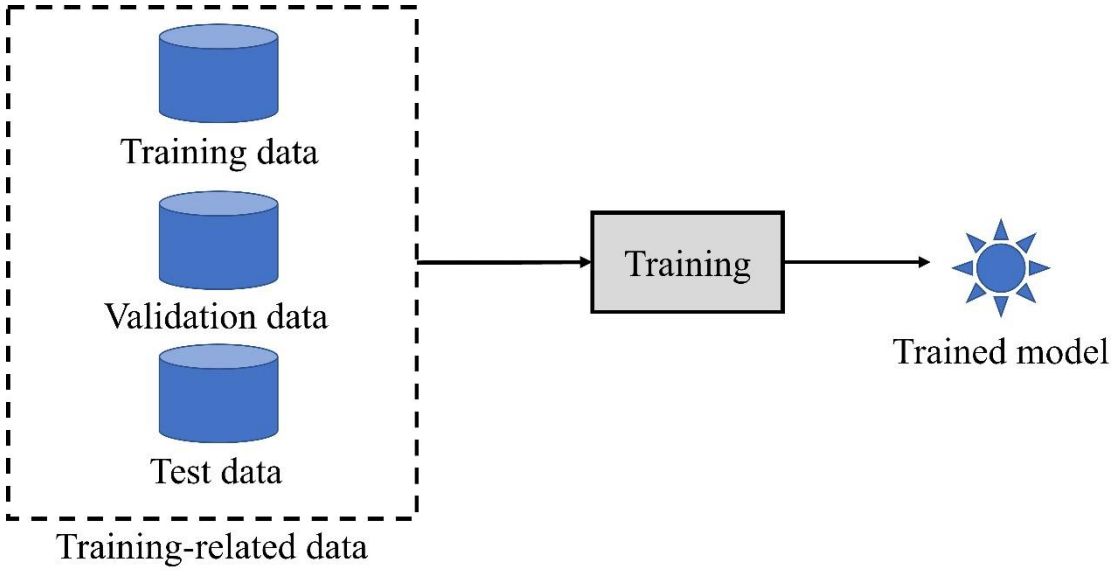

Figure II- 1. Training pipeline in a machine learning system

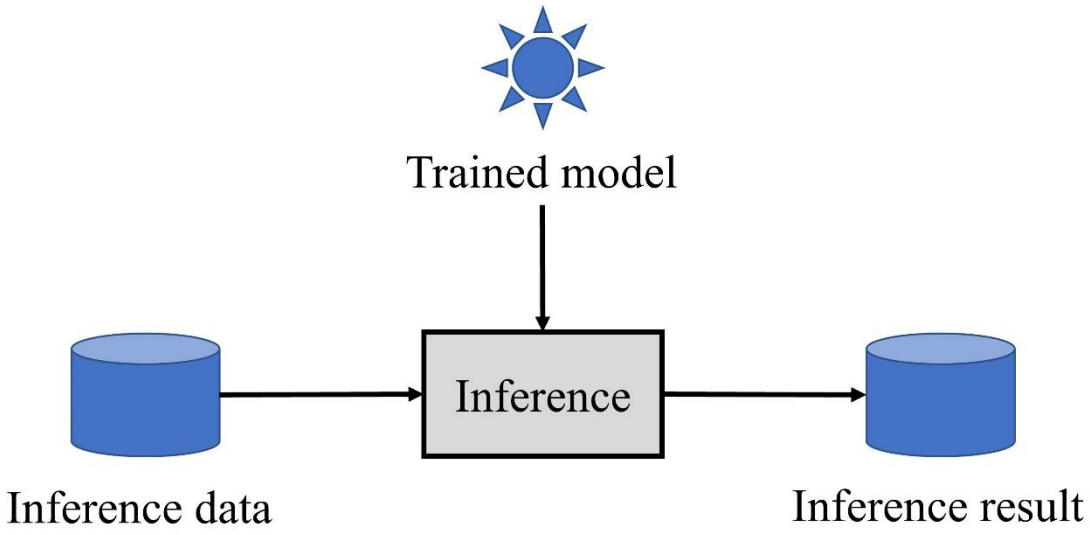

Figure II- 2. Inference pipeline in a machine learning system

II-2.2. Development Process of a Machine Learning System

Unlike the development of general IT systems, when developing a machine learning system, to develop a system that responds to customer demands, in many cases a prototype is made after the system design, and accuracy and performance are evaluated before formal development. If the prototype does not show the expected performance, the development may be restarted from the system design. An example of a development process in the machine learning processing part of a machine learning system, including such trials, is shown in Figure II- 3. This figure shows only AI construction part in the flows of AI utilization in the AI Utilization Guidelines [II-1] from the Conference toward AI Network Security of the Japanese Ministry of Internal Affairs and Communications referred to in Section I-1.3.1 of Part I of the Machine Learning System Security Guidelines. Part II considers inserting a security risk assessment phase in the development process. The results are explained in Chapter II-5.

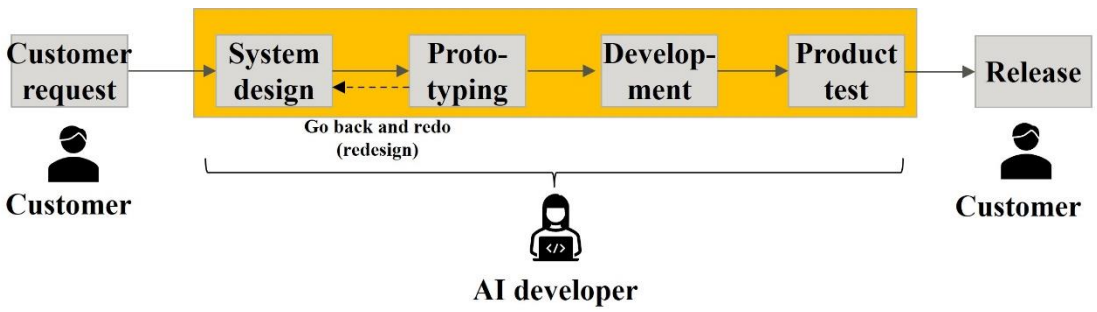

Figure II- 3. General development process of machine learning processing in the machine learning system

II-3. Overview of Machine Learning System Security

This chapter summarizes the attack methods and damages to machine learning systems that are covered in Part II.

II-3.1. Attack Method Against Machine Learning

Recently, machine learning-specific attacks on machine learning systems have been reported in many papers. These attacks involve machine learning making incorrect decisions and stealing training data and trained models, even though machine learning systems are accessed with legitimate authority. An attacker always has legitimate access authority and operates the system legitimately. In terms of the system side, this processing is normal and difficult to distinguish from legitimate system use. This point is a distinction from attacks in the general information security field (attacks in the information security field often cause systems to operate abnormally by inputting abnormal data into the systems.). Typical attacks on machine learning are summarized in Section I-2.2 of Part I of the Machine Learning System Security Guidelines and Table II- 1.

Table II- 1. Typical attacks on machine learning

Lead to misjudgment of a trained model	Evasion attack (adversarial examples)	To intentionally create an inference data/inference object that machine learning misjudges in inference.
	Poisoning attack	Train by inserting data given by an attacker into training data so that machine learning misjudges when the inference process is performed.
Leads to the theft of information from trained models	Model extraction	Performs the inference process of machine learning many times legitimately and replicates the trained model to the attacker.
	Model inversion	Performs the inference process of machine learning many times legitimately and recovers training data in the attacker’s environment.
	Membership inference	Performs the inference process of machine learning legitimately and infers whether data given by an attacker are included in training data or not.

II-3.2. Damage by Attacks

The damage caused by the attacks described in Table II- 1 is as follows.

- Evasion attack (adversarial example)

Data and objects created by an attacker can cause machine learning systems to misjudge. For example, in the case of a machine learning system that classifies road signs captured by a camera, such

as in self-driving, an attack such as putting a tape at a well-calculated position of a road sign to cause the misclassification of road signs is assumed [II-2]. A self-driving car takes pictures of this sign, misclassifies it as a different sign, and causes an accident.

- Poisoning attack

This attack succeeds when an attacker can intervene in the training phase of the machine learning system and trains the data created by the attacker. Thus, the accuracy of the machine learning system may be degraded, or misjudgment may occur. In addition, a risk exists that the machine learning system is trained to misjudge only when specific data are input. This specific data is called a “backdoor.”

- Model extraction

An attacker accesses a machine learning system many times and replicates the target system model. Consequently, the model of the machine learning system developed with much effort and cost is replicated by the attacker, who may use the model for free. An attacker may also deploy the service using a replicated model.

- Model inversion

An attacker accesses a machine learning system many times and recovers the training data of the target system. Thus, the training data used by the machine learning system during the training process may have leaked to an attacker, causing privacy problems. For example, in a face classification system, whose image is used for training may be leaked.

- Membership inference

An attacker accesses a machine learning system and infers whether data given by the attacker are included in training data. Consequently, training data may be leaked to an attacker, causing privacy problems. For example, in a machine learning system for addressing past medical history, an attacker may estimate whether the data of a specific person is included in the training data, and the attacker knows that this person has a disease listed in the past medical history.

II-4. Securing Machine Learning Systems

This chapter discusses strategies for preventing attacks on machine learning systems, as described in Chapter II-3, and the handling of general IT security.

II-4.1. Strategies for Protecting Machine Learning Systems

As shown in the Part I of the Machine Learning System Security Guidelines, systems can be protected from machine learning-specific attacks in two ways.

1. Dedicated defenses against attacks on machine learning systems
2. Operational defenses that make attacks difficult to apply to machine learning systems

In the above, the dedicated defenses are the protection methods against machine learning-specific attacks on the machine learning system described in Section II-3.1. Although many methods have been studied and proposed, as described in Chapter I-6 of the Part I of the Machine Learning System Security Guidelines, they have also noted that an attacker who knows which defense method is used to protect a system may be able to perform an attack that evades this defense. Thus, the silver bullet for protecting machine learning systems has yet to be established. Therefore, the preferred approach reduces the opportunities of applying attacks as much as possible before applying the dedicated defenses. One way to reduce the opportunities for applying attacks is to set system specifications appropriately and to protect systems by appropriate operation. For example, in an attack that generates an adversarial sample, an attacker can conduct the attack by performing the inference process many times. Therefore, limiting the number of performances of the inference process in a certain period is considered to protect against such attacks. Such protection is achieved by knowing which attacks can be performed on the system and adopting a system specification that prevents the attacker from satisfying the execution conditions required to perform the attack (in the above example, “an attacker can perform much inference processing,” and “an attacker can obtain inference results,” and “an attacker can obtain the data of the machine learning system,” etc.). Therefore, the attacks that can be performed and the conditions for performing these attacks must be known. A threat analysis is important as a knowing method for this information.

II-4.2. Relationship to General IT Security

As described in Section I-1.3.3 of the Part I of the Machine Learning System Security Guidelines, in addition to machine learning system-specific attacks, machine learning systems may be able to conduct attacks in the general IT security field. For example, attacks that intrude on a system and steal a machine learning model directly are considered. Securing machine learning systems must be protected from the vulnerabilities of traditional IT security and the machine learning-specific attacks

described in Part II, which describes only machine learning-specific attacks.

II-5. Risk Assessment on the Development Process of a Machine Learning System

In this chapter, the development process for taking measures against machine learning system-specific attacks and its problems are summarized, and the desired development process is described.

II-5.1. Development Process Considering Security Against Machine Learning System-Specific Attacks

As discussed in Chapter II-4, risk assessment is necessary to protect systems from the machine learning system-specific attacks. The risk assessment includes, in addition to the above-described threat analysis, an impact analysis that analyzes the impact of an attack. The assumed process in which security measures are considered for the development process of the machine learning processing unit in the general machine learning system shown in Figure II- 3 is as shown in Figure II- 4.

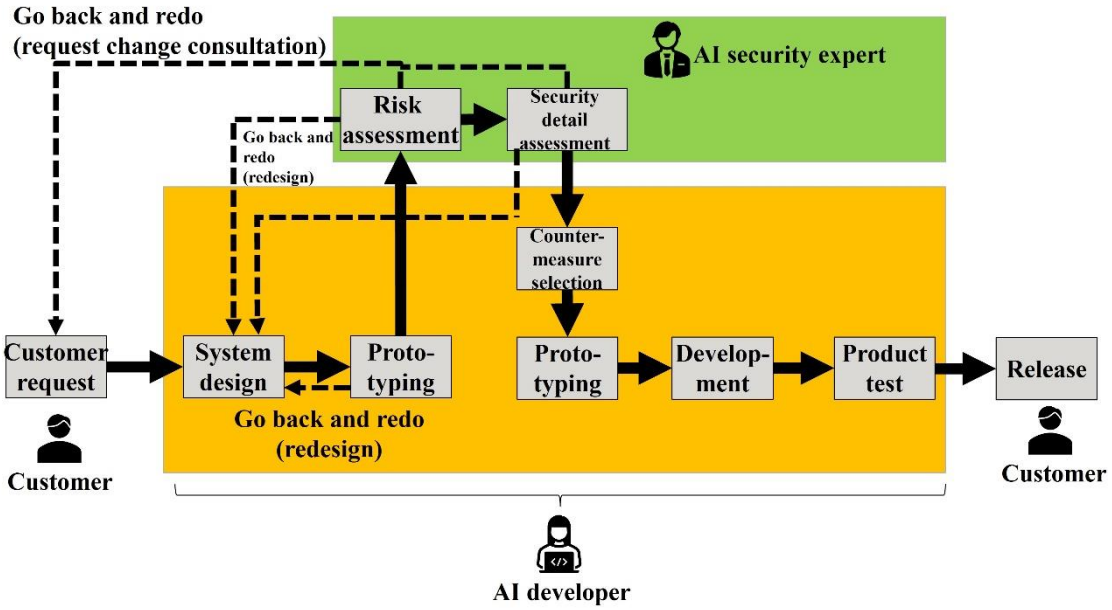

Figure II- 4. Development process of machine learning processing part considered for security against machine learning system-specific attacks

Currently, risk assessment against machine learning system-specific attacks is generally performed by machine learning security experts (AI security experts). On the request of AI developers, AI security experts conduct a risk assessment after hearing from AI developers and customers about which incidents on a system are considered problems (which attacks do the AI developers want protection from, and what damage is considered a problem?). If an attack can be conducted on a system, AI security experts consider the countermeasures of what type of specification and operation can provide protection and notify the AI developers of the countermeasures. The notified AI developer redesigns

the system and starts over from the Prototyping. Alternatively, if the attack can be conducted when the customer’s requests are satisfied, the system requests are redesigned after consulting with the AI developer and the customer. However, in the development process shown in Figure II- 4, the risk assessment by AI security experts may find many problems, and risk assessment and redesign may be repeated many times. Such reworking may reduce development efficiency, increase development costs, and delay delivery. Therefore, a more efficient development process is expected. To solve this problem, AI developers must be able to conduct risk assessments that AI security experts are currently required to perform. Additionally, AI security experts are not numerous, and not all companies have AI security experts, so a good solution is for AI developers themselves to assess their systems.

In part II, this kind of risk assessment done by AI developers themselves is called “**risk assessment for AI developers.**” With risk assessment for AI developers, AI developers can conduct their own risk assessment to guide secure specifications and operations and redesign will not result in as much repetition as the process shown in Figure II- 4. This result will also allow companies without AI security experts to conduct risk assessments (but if a vulnerability is discovered, AI security experts must be consulted). An example of introducing risk assessment for AI developers into the development process is shown in Figure II- 5. In Part II, we introduce the methods for implementing risk assessment for AI developers and an example of risk assessment for AI developers that we have actually designed. Please refer to Appendix of the Machine Learning System Security Guidelines for the attack detection technology that may be included in the proposed countermeasures presented after the detailed security evaluation.

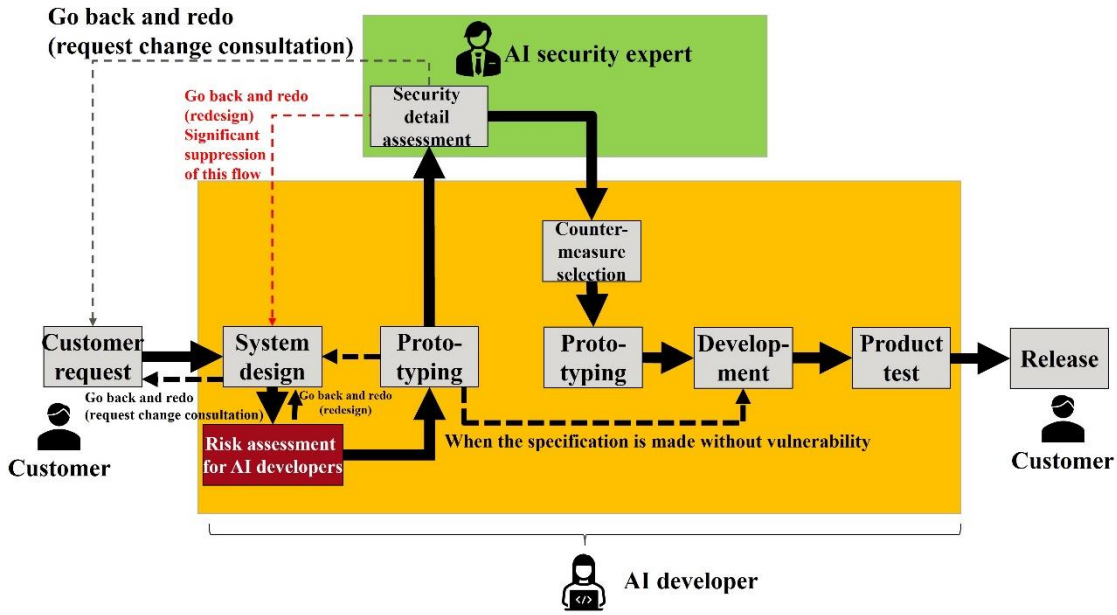

Figure II- 5. Development process of the machine learning processing unit that reduces rework due to security measures (desired model)

II-5.2. Threat Analysis for Machine Learning Systems

Machine learning threat analyzing technologies that AI developers and AI security experts are collaborate on are being proposed. In [II-3], the European Union Agency for Cybersecurity (ENISA) summarized the threats and assets to AI systems considering their life cycles. The report also outlined a five-level approach to threat modeling that includes asset identification, threat identification, and vulnerability identification. In [II-4], Microsoft outlined its ideas for modeling threats in AI. On this web page, a list of questions to check when developing AI is summarized. Many of these questions require AI security expertise. These technologies are considered so that they apply to security experts (Figure II- 4 and Figure II- 5) and can be used as references for threat analysis performed jointly by AI developers and AI security experts.

II-6. Risk Assessment for AI Developers

This chapter introduces an effective risk assessment method to the threat analysis part of risk assessment for AI developers. Although this threat analysis technology is intended to be used in the development process described in Chapter II-5, it does not necessarily assume such a process, and anyone, including analysts who are not AI security experts or AI security experts, can analyze machine learning system security from specification information by using this technology.

II-6.1. Overview of the Risk Assessment for AI Developers

Risk assessment for AI developers must identify (1) which attacks can be conducted on a system under development and (2) what damage will be caused by each attack. In addition, (3) what type of specification should be changed or what type of system operation should be conducted to prevent the attacks that are judged to be feasible is analyzed, and this information will be used as a reference for redesign. In Part II, this method is introduced as an analytical technique to solve (1) and (3).

In this method, an analyst (AI developer) performs the assessment by answering selective-choice questions previously prepared by AI security experts. After answering the questions, whether attack trees prepared by the AI security experts are satisfied is judged from the answers to the questions. Thus, which attack can be executed is clarified and (1) is solved. In this method, because the condition for not satisfying the satisfied tree is visualized, so (3) can also be solved. As for (2), because the threats are limited by the focus on the machine learning system-specific attacks, even analysts who are not AI security experts can clarify which attacks can be performed and what damage they cause. As for (2), please refer to Part I of the Machine Learning System Security Guidelines, which includes the method of performance. In Part II, this method is described in detail.

II-6.2. AI Security Risk Assessment Method

The requirements for this method are as follows.

1. AI developers who may not have machine learning security expertise can assess threats to machine learning systems.
2. Almost the same result is derived for a system no matter who assesses threats.
3. The results of the threat analysis have high acceptability.

An assessment method using an attack tree [II-5] is introduced as a technology satisfying the above requirements. In this technology, AI security experts prepare attack trees in the preparation phase and assessors self-assess whether the prepared trees are satisfied. After the preparation by AI security experts is completed, the assessment can be performed by AI developers who may not have AI security expertise. In this assessment, because the results are shown in an attack tree format, an easy understanding of the results is enabled. The procedure is described in detail below. Examples of attack trees, questions, etc. prepared by the authors are described in Chapter II-7. Examples of assessment

using these materials are described in Chapter II-8.

II-6.2.1.Preparation Procedures for Machine Learning Security Experts

First, AI security experts perform the following preparation phase for the assessment. This preparation phase only needs to be performed once.

II-6.2.1.1. Preparation of Attack Trees and Conditions for Attack Execution

This phase constructs attack trees for machine learning system-specific attacks and is performed by AI security experts. Threat analysis using attack trees is a type of analysis technology used in the general IT security field. An attack tree is generated in the tree structure with a threat as a top node, and realizing threat conditions are placed as a logical hierarchy. In general IT security, constructing an attack tree is difficult before a system specification is defined, because the tree representation has a high degree of freedom. However, in the case of machine learning systems, because the types of attacks that can be executed and the damage that can be caused are limited, an attack tree can be constructed before determining the system specifications. In general, a single attack category on machine learning systems has multiple attack scenarios (attack algorithms). When a tree is being constructed, attack scenarios to be assessed are determined, and the conditions for performing each attack scenario are extracted and placed in a node. The scenarios to be assessed are determined by the level of detail of the assessment; however, starting with constructing trees is preferable for typical scenarios. The conditions for performing attacks are determined by referring to papers. Examples of a part of the constructed tree are shown in Figure II- 6. These examples concern evasion attacks (adversarial examples). Four scenarios are prepared for executing an evasion attack (adversarial examples) (upper side of the figure). Either scenario being satisfied determines that the attack can be conducted. The lower part of the figure exemplifies attack scenario A1 of an evasion attack (adversarial examples). When the left side and the right sides of the tree are simultaneously satisfied, the attack scenario A1 is determined to be applicable (TRUE). The condition on the left side is TRUE when “Condition 6-2 OR (Condition 2-2 AND Condition 3-1)” is satisfied. The condition on the right side is TRUE when “Condition 4-2 OR Condition 7-1” is satisfied. These conditions written on each node are called “attack executable conditions”.

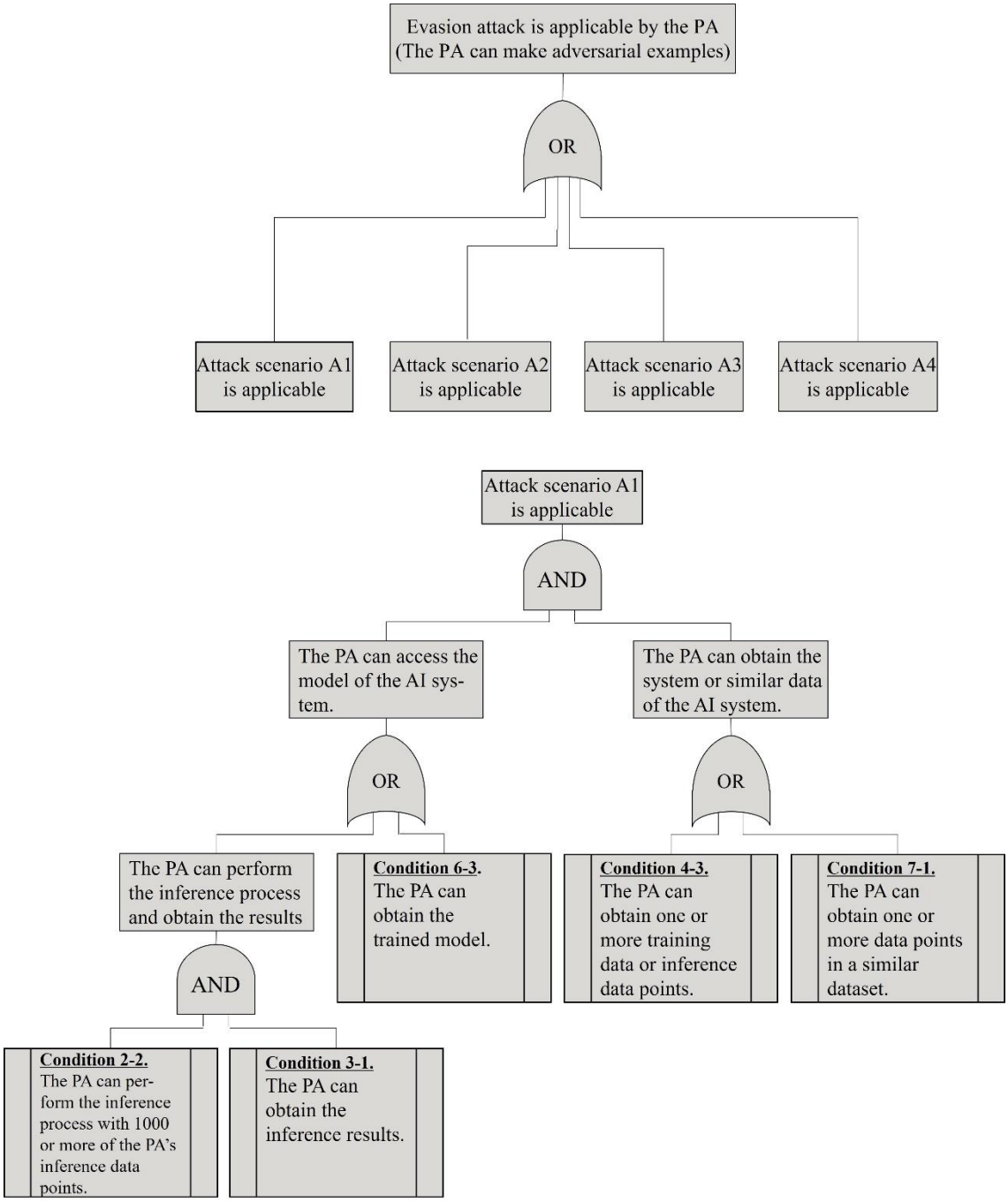

Figure II- 6. Examples of (a part of) constructed attack trees

II-6.2.1.2. Preparation of Questions

After the preparation of attack trees and attack executable conditions, questions for determining whether a given system specification satisfies the attack executable conditions are created. This task is also performed by AI security experts. Assessors are assumed to be AI developers and do not necessarily have expertise in machine learning security, so for easy answering, questions that are as

simple as possible and about specifications are preferable. Some examples should be provided with questions that are easy for the assessors to understand. The following example is our question. The potential attacker is described later.

Example:

Question: obtaining similar data

“Can the potential attacker (PA) obtain one or more data points of a similar dataset (similar objective and similar genre) to the AI system?”

1. The number of data points means the number of rows in the table dataset, or the number of images in the image dataset.

Example for “Yes”:

System type: face recognition system by training a face dataset

PA: a person who can be recorded by the camera

Notes: A PA can prepare a face dataset from the Internet, and it is a similar dataset.

Example for “Yes”:

System type: income prediction system

PA: a person who can perform the system

Notes: When a PA knows the attribute of the AI system and can prepare a dataset whose data distribution is similar to the original dataset, the selection is “Yes”.

II-6.2.1.3. Preparation of the Judgment Table for Determining the Satisfaction of Attack Executable Conditions

A table is prepared for determining whether the attack executable conditions (the conditions described in the nodes of attack trees) prepared in Section II-6.2.1.1 are satisfied from the answers to the questions prepared in Section II-6.2.1.2. For example, a judgment table such as Table II- 2 is prepared. This table also contains the system requirements for setting each condition to FALSE when it is TRUE. The system requirements to be set to FALSE are used to consider countermeasures.

Table II- 2. Example of a judgment table for determining the satisfaction of attack executable conditions

Condition	Conditions to be TRUE	Result of judgment	Countermeasure to be FALSE
Condition 1-1. The PA can perform the training process.	The answer to question 1-1A or question 1-1B is “Yes.”	TRUE or FALSE (filled by an assessor)	Prevent the PA from performing training operations.
Condition 2-1. The PA can perform the inference process with one or more of the PA’s inference data points.	The answer to question 2-1A or question 2-1B is “Yes.”	…	Prevent the PA from performing the inference process.
Condition 2-2. The PA can perform the inference process with 1000 or more of the PA’s inference data points.	The answer to question 2-2A or question 2-2B is “Yes.”		Prevent the PA from performing the inference process with 1000 or more inference data points.
Condition 2-3. The PA can perform the inference process with 10000 or more of the PA’s inference data points.	The answer to question 2-3A or question 2-3B is “Yes.”		Prevent the PA from performing the inference process with 10000 or more inference data points.
Condition 2-4. The PA can perform the inference process with 1000000 or more of the PA’s inference data.	The answer to question 2-4A or question 2-4B is “Yes.”		Prevent the PA from performing the inference process with 1000000 or more inference data points.
Condition 3-1. The PA can obtain the inference results.	The answer to question 3-1 or question 3-2 is “Yes.”		Prevent the PA from obtaining the inference results.
Condition 3-2. The PA can obtain confidence scores.	The answer to question 3-2 is “Yes.”		Prevent the PA from obtaining confidence scores.
…	…		…

II-6.2.2.Assessment Procedures for Assessors

The procedure is described below. These procedures can be repeated many times if the preparations described in Section II-6.2.1 are completed.

II-6.2.2.1. Clarification of the System Specifications to be Assessed and of the PAs of the System

As basic information for answering questions prepared by AI security experts, the assessor prepares definition material (this material contains information such as AI tasks, the training process performer/performing method/data input method, the inference process performer/performing method/data input method, the output content to be presented, and the presentation method/presentation destination) that describe system specifications to be assessed in as much detail as possible. This definition material includes information such as “Does the system show its user the output of the machine learning system?”, “How many data of queries to the system are allowed per hour?”, and “Will the training/inference data be open to the public?”.

Then, the PAs need to be identified. In the assessment, the authority of the PA is considered, and the assessment results are affected by who is assumed to be the PA. Therefore, a PA must be appropriately selected and set up. When a PA is assumed to have low relevance to the system, such as a person who

provides data to the system, the assessment assumes attacks from outside attackers; when a PA is assumed to have high relevance to the system, such as a system manager, the assessment assumes attacks from insider attackers. As suitable PAs, at least the following people should be assumed.

1. AI developer (when insider attackers are assessed)
2. Machine learning system administrator (when insider attackers are assessed)
3. End user of a machine learning system
4. Person whose data are used in a machine learning system (not necessarily a user)

II-6.2.2.2. Answering to Questions

When the system specifications and potential attackers have been clarified, the assessor answers YES or NO to the questions prepared by the AI security experts. Questions are prepared as described in Section II-6.2.1.2. Examples of all questions are provided in Chapter II-7.

II-6.2.2.3. Confirmation of the Satisfaction of the Attack Executable Conditions

The answers to the questions are used to determine whether each attack executable condition described in the node of the attack tree corresponding to each attack scenario (TRUE/FALSE) is satisfied. This task can be performed by preparing a judgment table as described in Section II-6.2.1.3, which can be uniquely determined from the answers to the corresponding questions. An example of a decision table is provided in Chapter II-7.

II-6.2.2.4. Confirmation of the Satisfaction of Attack Trees

Information (TRUE/FALSE) on whether the attack executable conditions determined based on the judgment table are satisfied is filled in the nodes of attack trees. Thus, whether each attack tree is satisfied, that is, whether attack scenarios are applicable, can be determined. An example of this work is provided in Section II-8.

II-6.2.2.5. Consideration of Countermeasures

A satisfied attack tree indicates that an attack can be conducted by a potential attacker. In this phase, measures are considered to prevent attacks by potential attackers. Specifically, according to the structure of the attack tree that has been satisfied, changes in specification are considered for making the attack executable condition on each node FALSE. An example of this consideration is shown in Figure II- 7. In this example, attack scenario A1 of an evasion attack (adversarial example) is applicable. Looking at the tree of this attack scenario described at the bottom of the figure, if the machine learning system specification is changed to not satisfy Condition 2-2, this attack scenario becomes difficult to execute. Specifically, the number of times a potential attacker can execute inference processing is limited to less than 1000 times in a fixed period. However, given the possibility

of attackers colluding, the number of times that the total number of times the inference processing can executed should be limited to less than 1000 times in a fixed period. A fixed period is the period when an attack is prevented, for example, the lifetime of a product. AI developers consider the acceptability of changing this specification so that this condition is no longer satisfied. Specifically, after considering which leaves of the attack tree should not be satisfied (FALSE) (almost identical to the consideration of specification change), the assessor considers whether the specification that the condition on the leaves can be FALSE can be changed by referring to the “Countermeasure to FALSE” column in the judgment table described in Table II- 2. If it is determined that the specification cannot be changed, the assessor should consider failing another condition. Otherwise, consult with AI security experts to implement specific countermeasures against machine learning-specific attacks.

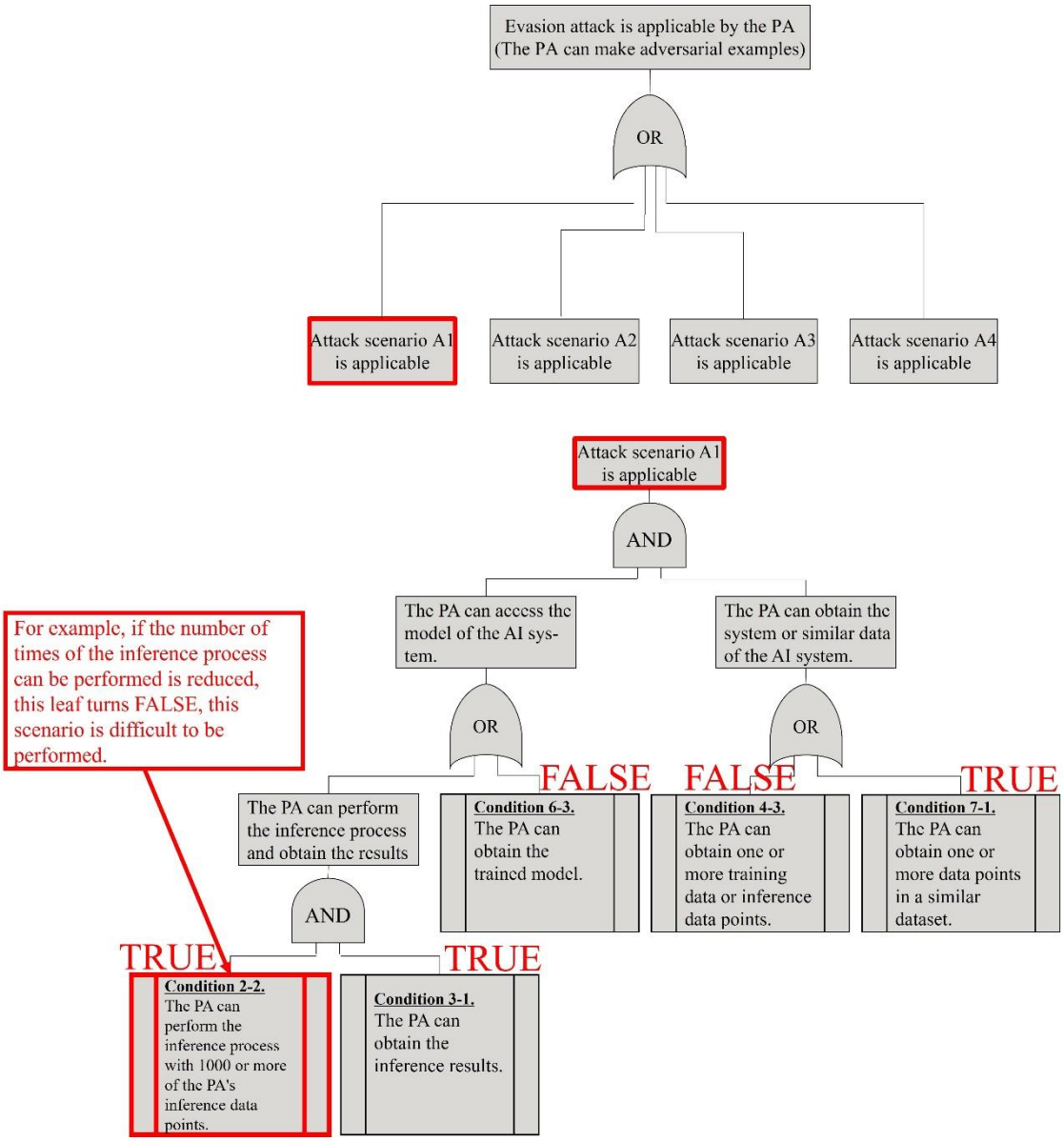

Figure II- 7. Example of the consideration of countermeasures

II-7. Realization Example of the risk assessment method

This chapter introduces a realization example of the risk analysis for AI developers explained in Chapter エラー! 参照元が見つかりません。.

II-7.1. Notes

[II-6] provides a realization example of this method explained in Chapter エラー! 参照元が見つかりません。. [II-6] contains attack trees for evasion attacks (adversarial examples), poisoning attacks, model extraction attacks, and model inversion attacks. Additionally, Part II describes attack trees for membership inference attacks, and provides selective questions that are easier to understand. In this example, typical attack algorithms for each attack are defined as scenarios and extracted as attack trees. However, this example was realized in March 2022, and not all attack scenarios discussed at academic conferences are covered. Notably, revision may be possible in the future when attacks and countermeasures are improved.

II-7.2. Attack Trees and Attack Executable Conditions

The attack trees and the conditions under which an attack can be executed corresponding to the tree leaves (attack executable conditions) shown in [II-6] are described below. Table II- 3 shows the viewpoints of extracted attack scenarios in this realization example. The letter in the table indicates the variation of attacks described below.

A: Evasion attack (adversarial examples), P: Poisoning attack, X: Model Extraction, I: Model Inversion, M: Membership inference.

Table II- 3. Viewpoints of extracted attack scenarios for constructing attack trees

Attack scenario	Viewpoint of the construction of attack trees
A1	Basic evasion attack conditions for black-box attacks.
A2	Evasion attack conditions for white-box attacks and attacks using model duplication technology, which is simpler than model extraction.
A3	Evasion attack conditions using model extraction.
A4	Evasion attack conditions using poisoning attacks.
P1	Basic poisoning attacks. Conditions for poisoning attacks.
P2	Poisoning attack conditions if a backdoor is included in the model when this model is reused from outside or inside the environment.
P3	Poisoning attack conditions using model extraction.
X1	Model extraction conditions for data-free attacks.
X2	Typical model extraction conditions related to [II-7].
X3	Typical model extraction conditions related to [II-8].
X4	Model extraction attack condition when the treated data are a table dataset.

X5	Model extraction attack condition when the treated data are not a table dataset (e.g., an image dataset).
X6	Model extraction attack conditions when an attacker can obtain the model itself.
I1	Basic model inversion conditions.
M1	First pattern of the typical membership inference conditions related to [II-9].
M2	Second pattern of the typical membership inference conditions related to [II-9].
M3	Third pattern of the typical membership inference conditions related to [II-9].
M4	Typical membership inference conditions related to [II-10].
M5	Typical membership inference conditions related to [II-11].
M6	Typical membership inference conditions related to [II-12].
M7	Typical membership inference conditions related to [II-13].
M8	Membership inference conditions when an attacker can obtain the training data.

II-7.2.1.Examples of Attack Trees and Attack Executable Conditions for Evasion Attacks (Adversarial Examples)

Examples of attack trees and attack executable conditions for evasion attacks (adversarial examples) are as follows.

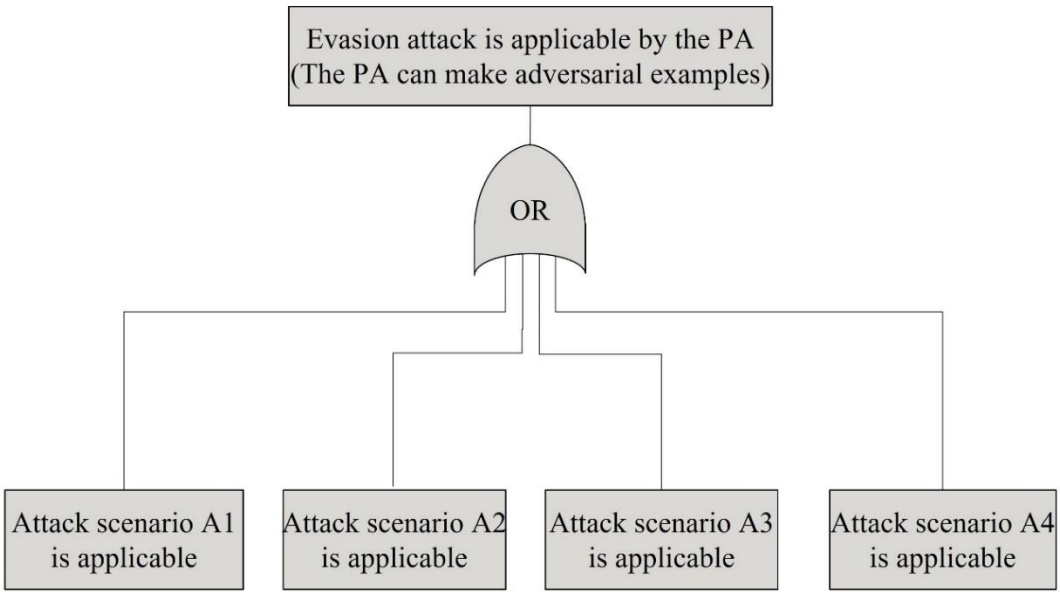

Figure II- 8. Example of the attack tree for evasion attacks (adversarial examples) (upper part)

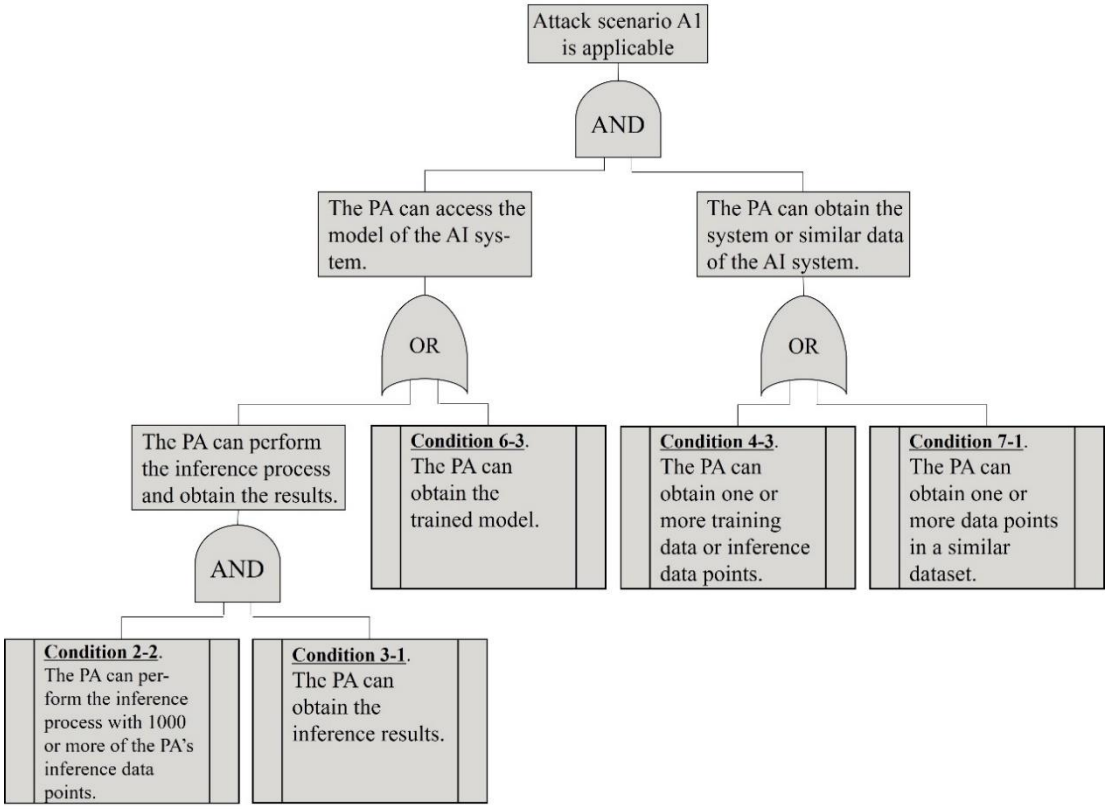

Figure II- 9. Attack tree and attack executable conditions for attack scenario A1 of evasion attacks (adversarial examples)

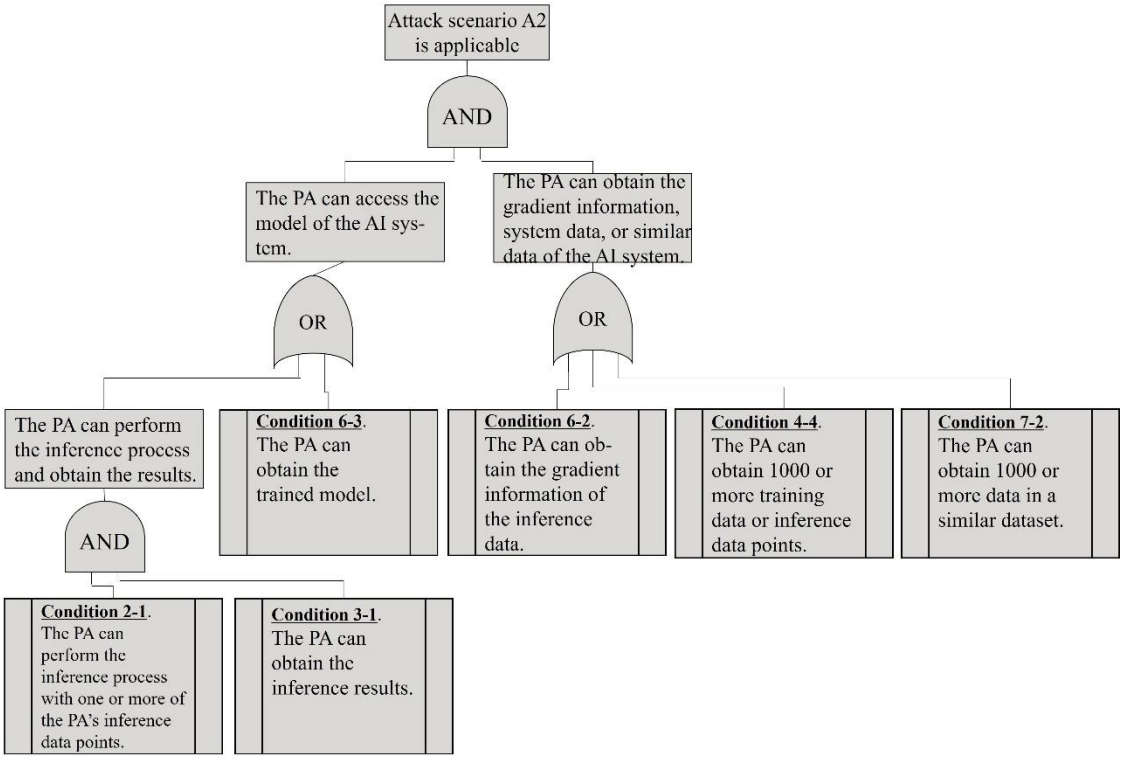

Figure II- 10. Attack tree and attack executable conditions for attack scenario A2 of evasion attacks (adversarial examples)

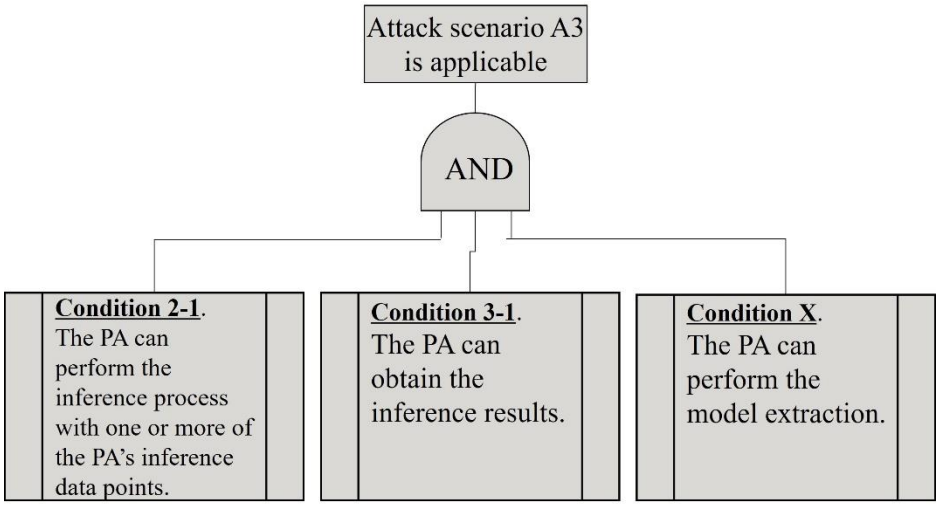

Figure II- 11. Attack tree and attack executable conditions for attack scenario A3 of evasion attacks (adversarial examples)

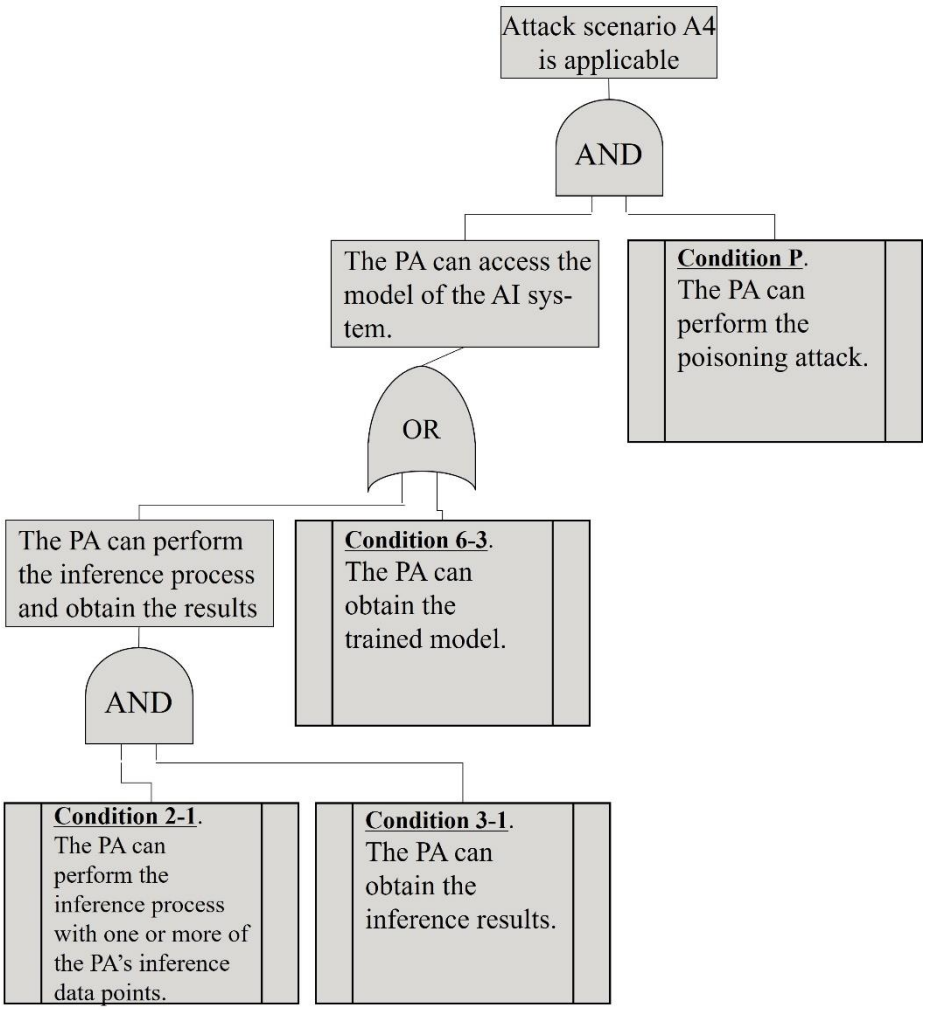

Figure II- 12. Attack tree and attack executable conditions for attack scenario A4 of evasion attacks (adversarial examples)

II-7.2.2.Examples of Attack Trees and Attack Executable Conditions for Poisoning Attacks

Examples of attack trees and attack executable conditions for poisoning attacks are provided as follows.

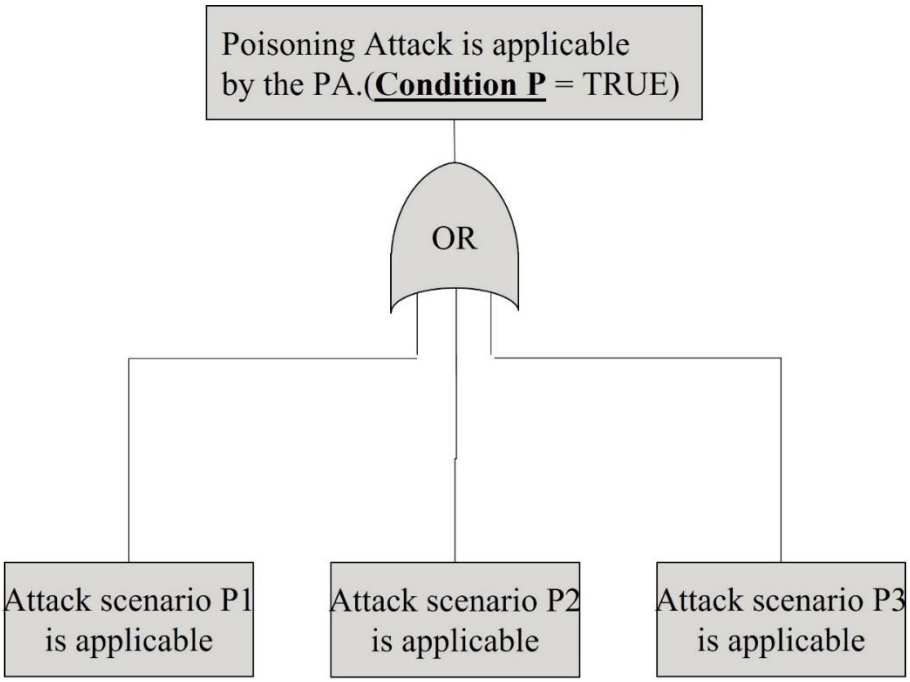

Figure II- 13. Example of the attack tree for poisoning attacks (upper part)

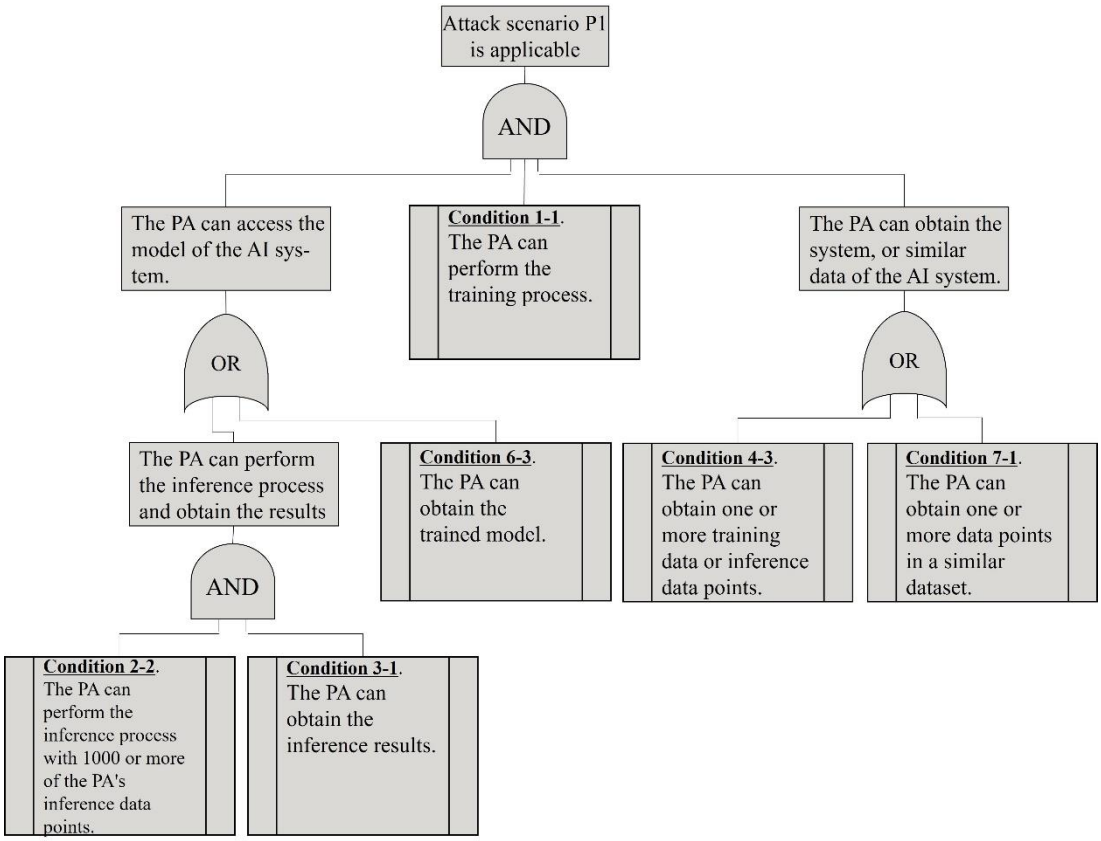

Figure II- 14. Attack tree and attack executable conditions for attack scenario P1 of poisoning attacks

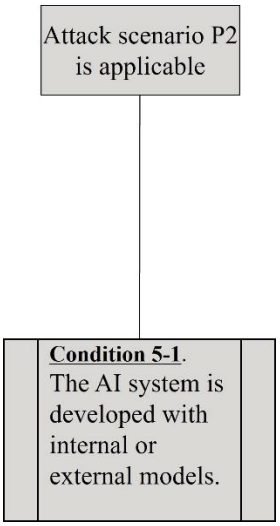

Figure II- 15. Attack tree and attack executable conditions for attack scenario P2 of poisoning attacks

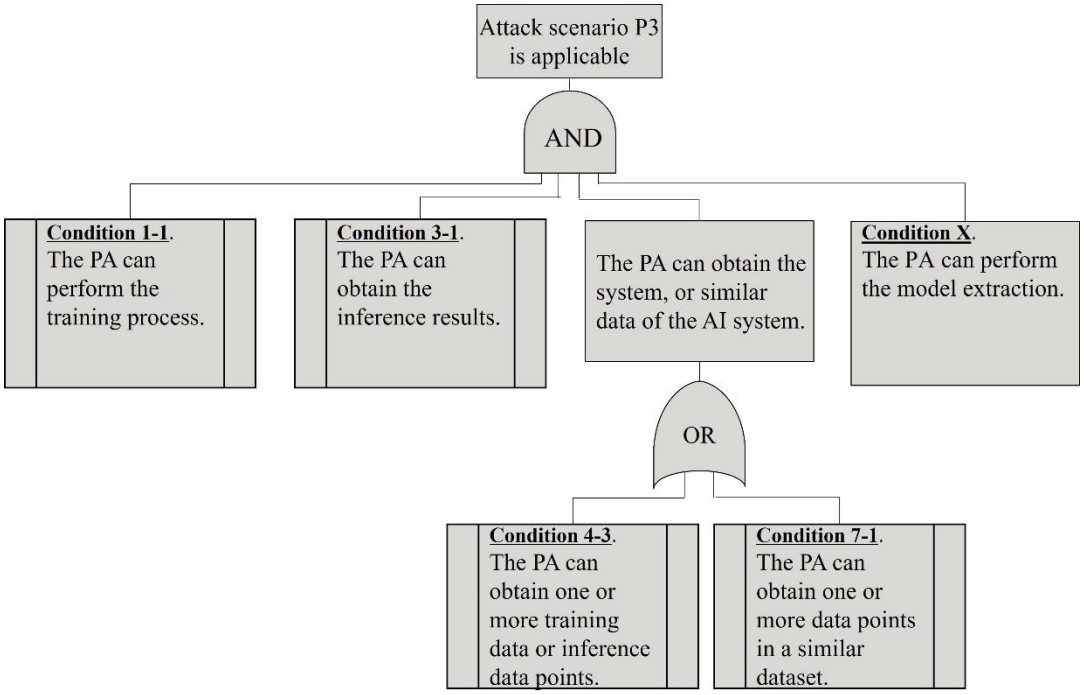

Figure II- 16. Attack tree and attack executable conditions for attack scenario P3 of poisoning attacks

II-7.2.3.Examples of Attack Trees and Attack Executable Conditions for Model Extraction

Examples of attack trees and attack executable conditions for model extraction are as follows.

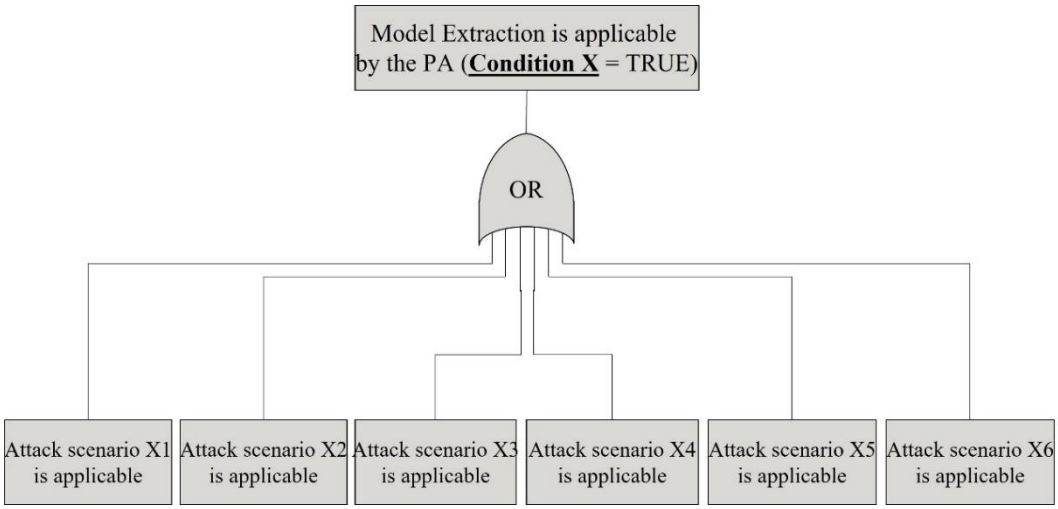

Figure II- 17. Example of the attack tree for model extraction (upper part)

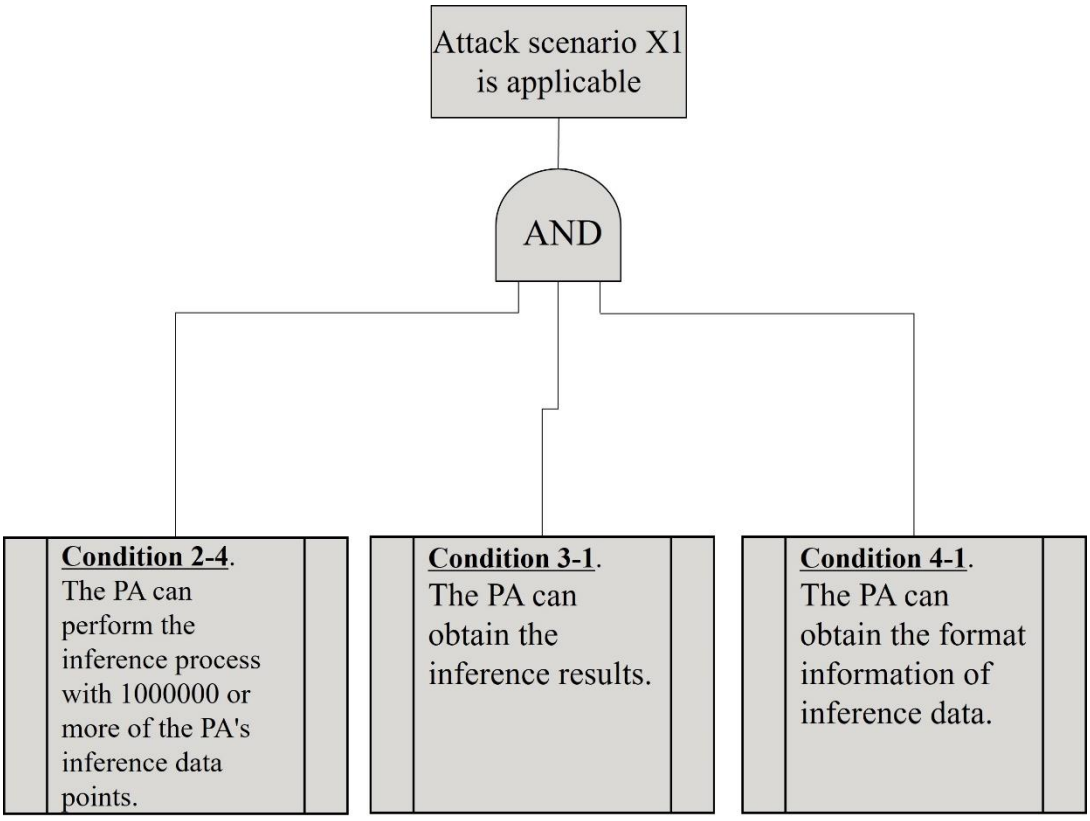

Figure II- 18. Attack tree and attack executable conditions for attack scenario X1 of model extraction

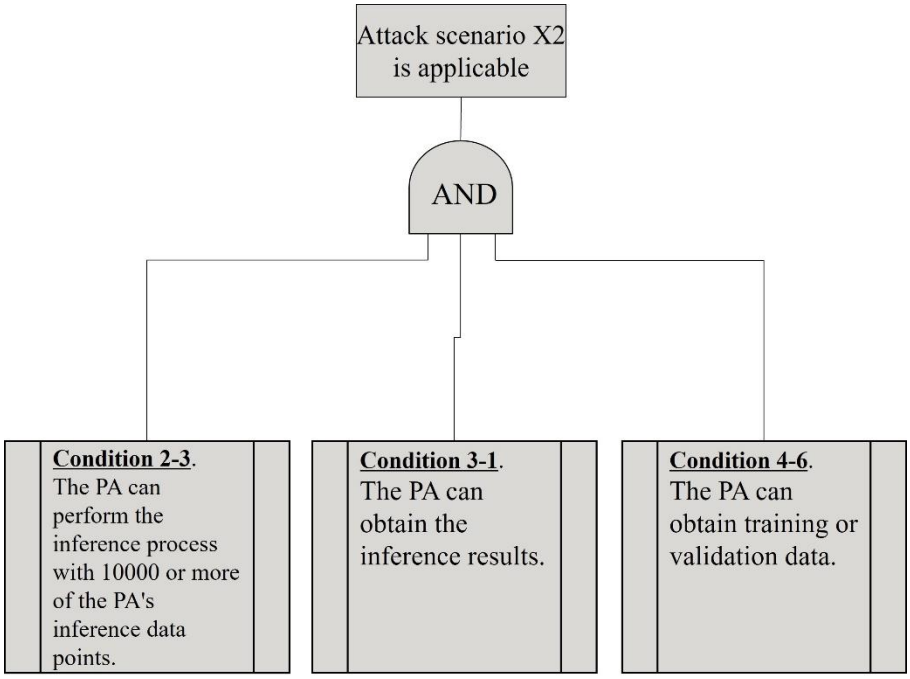

Figure II- 19. Attack tree and attack executable conditions for attack scenario X2 of model extraction

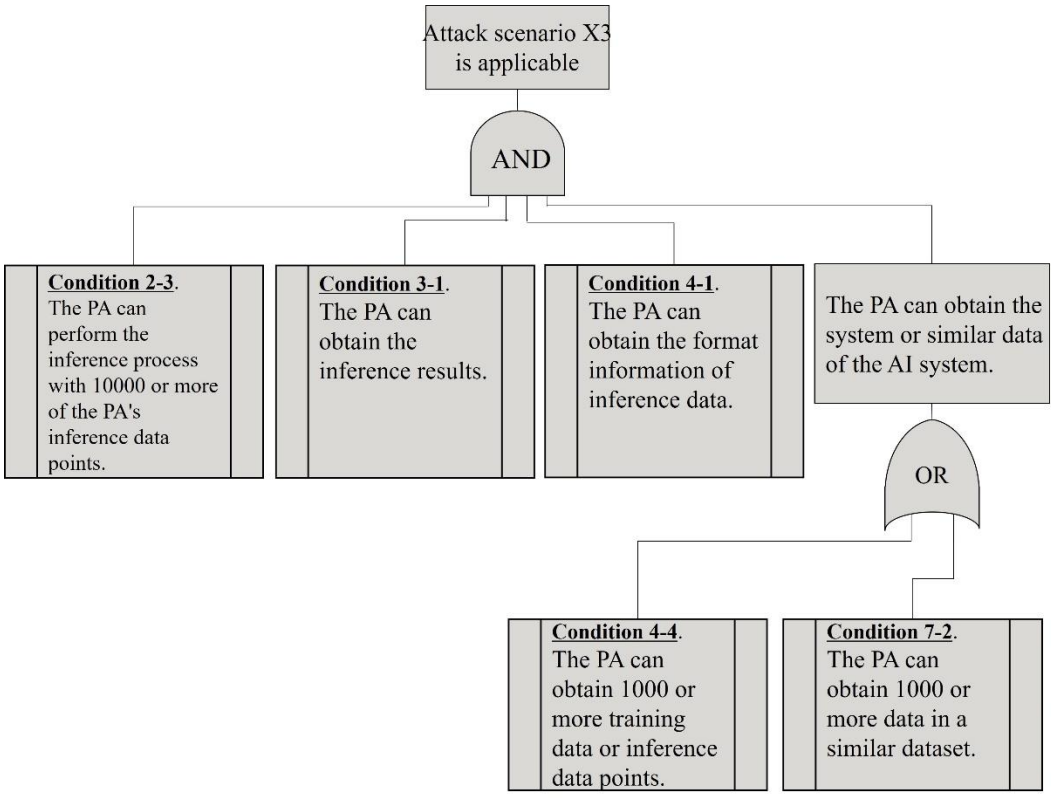

Figure II- 20. Attack tree and attack executable conditions for attack scenario X3 of model extraction

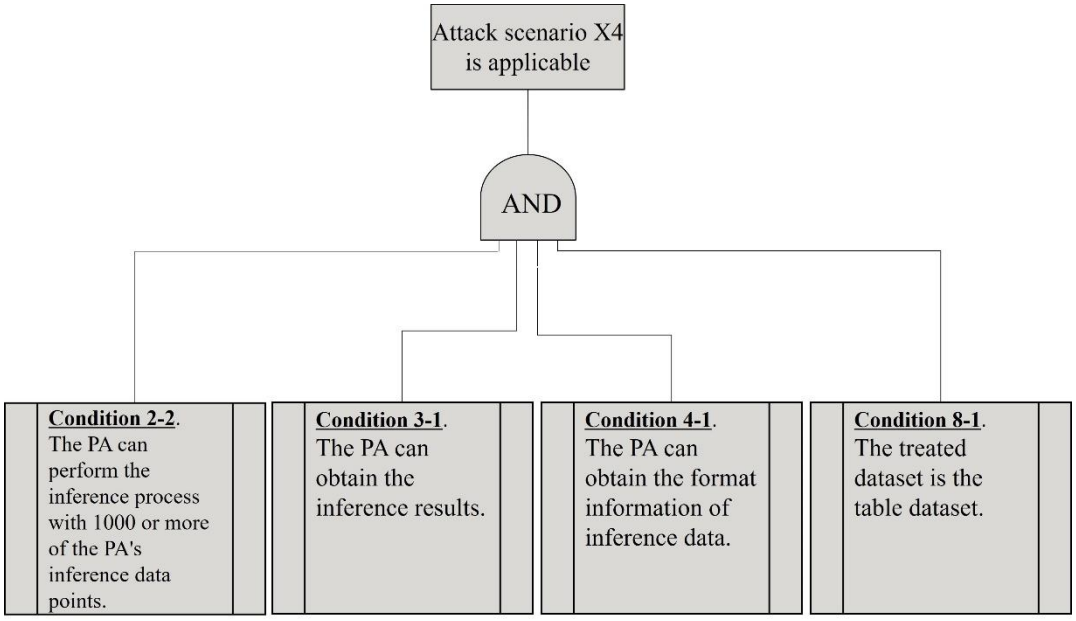

Figure II- 21. Attack tree and attack executable conditions for attack scenario X4 of model extraction

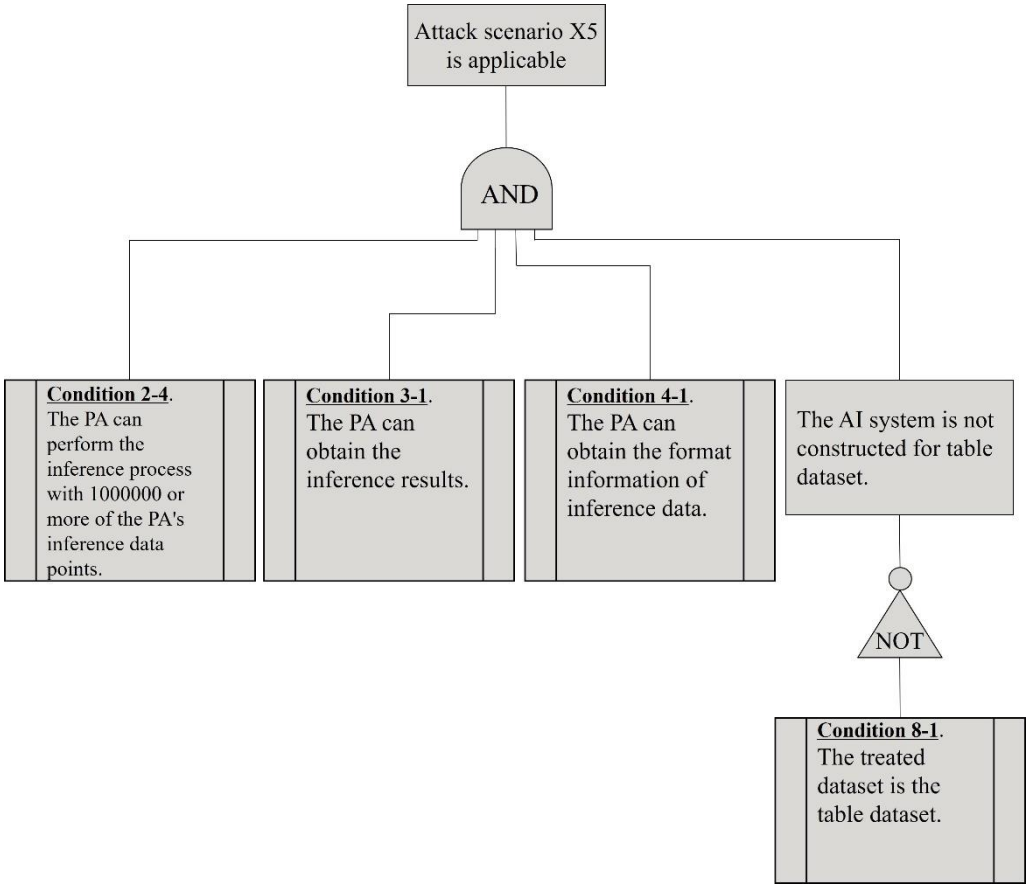

Figure II- 22. Attack tree and attack executable conditions for attack scenario X5 of model extraction

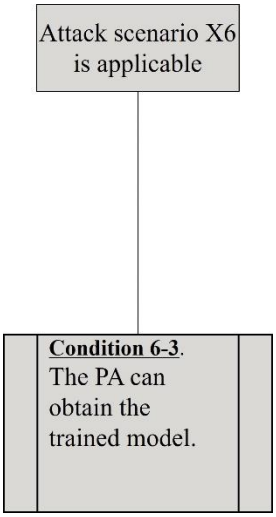

Figure II- 23. Attack tree and attack executable conditions for attack scenario X6 of model extraction

II-7.2.4.Examples of Attack Trees and Attack Executable Conditions for Model Inversion

Examples of attack trees and attack executable conditions for model inversion are as follows.

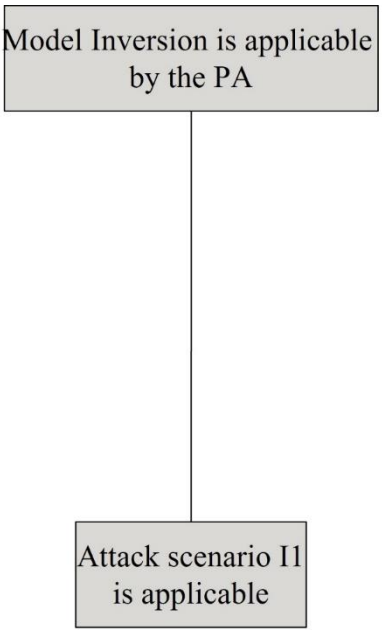

Figure II- 24. Example of the attack tree for model inversion (upper part)

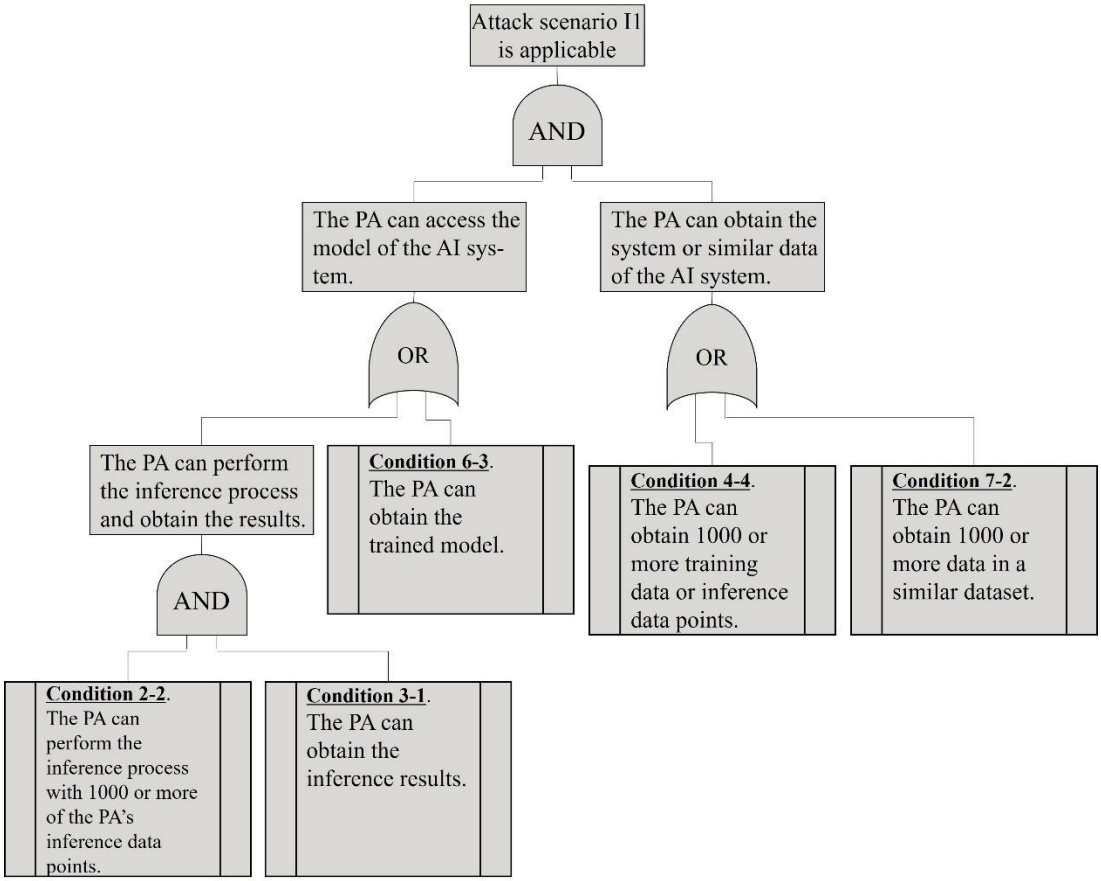

Figure II- 25. Attack tree and attack executable conditions for attack scenario I1 of model inversion

II-7.2.5.Examples of Attack Trees and Attack Executable Conditions for Membership Inference

Examples of attack trees and attack executable conditions for membership inference are as follows.

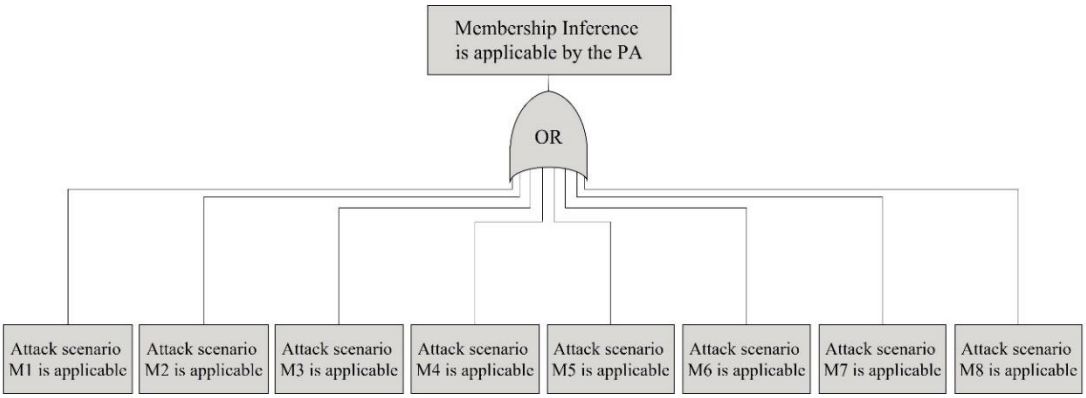

Figure II- 26. Example of the attack tree for membership inference (upper part)

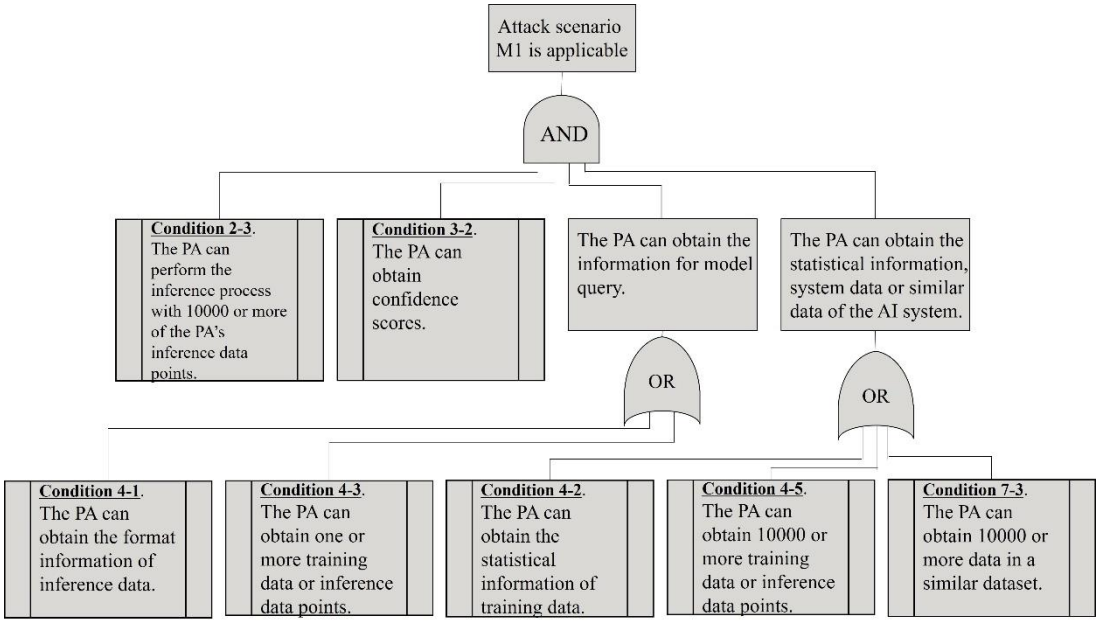

Figure II- 27. Attack tree and attack executable conditions for attack scenario M1 of membership inference

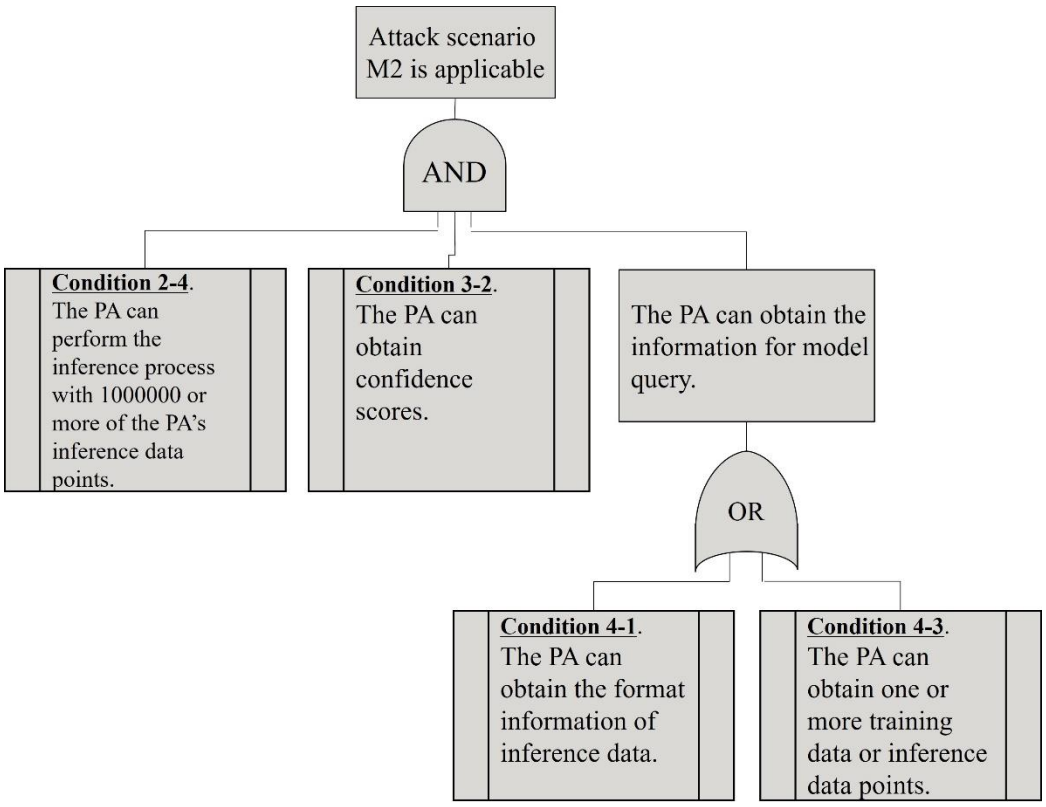

Figure II- 28. Attack tree and attack executable conditions for attack scenario M2 of membership inference

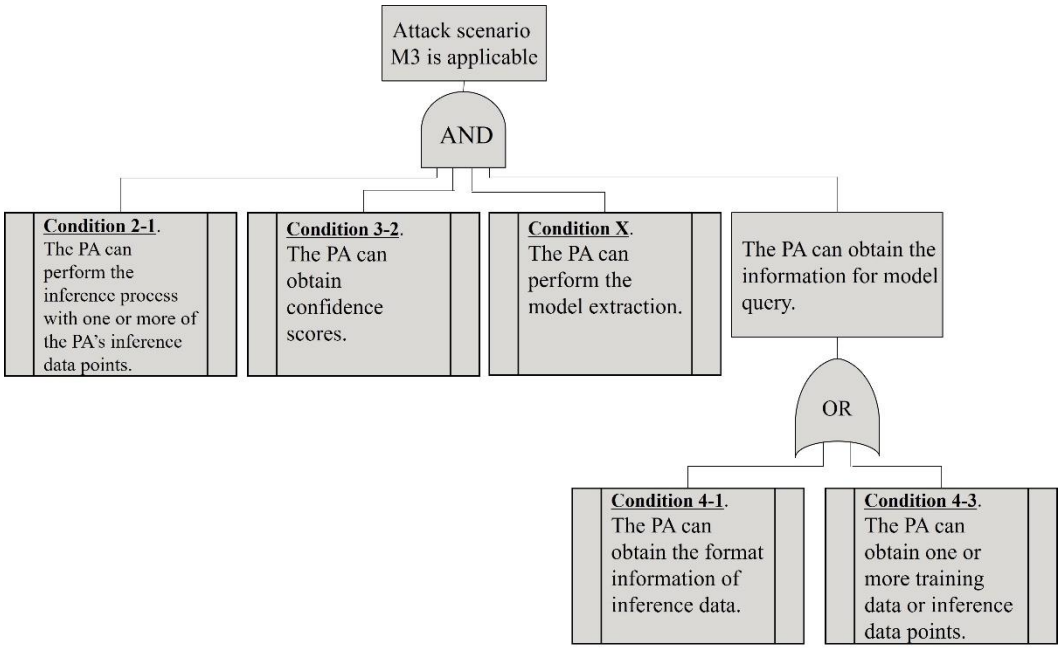

Figure II- 29. Attack tree and attack executable conditions for attack scenario M3 of membership inference

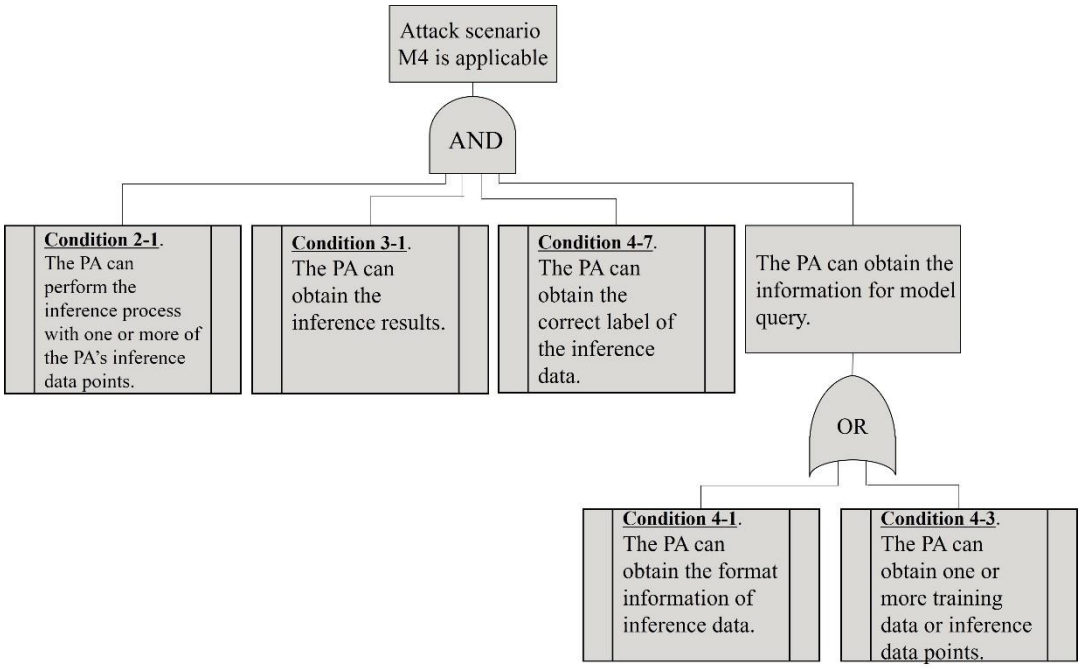

Figure II- 30. Attack tree and attack executable conditions for attack scenario M4 of membership inference

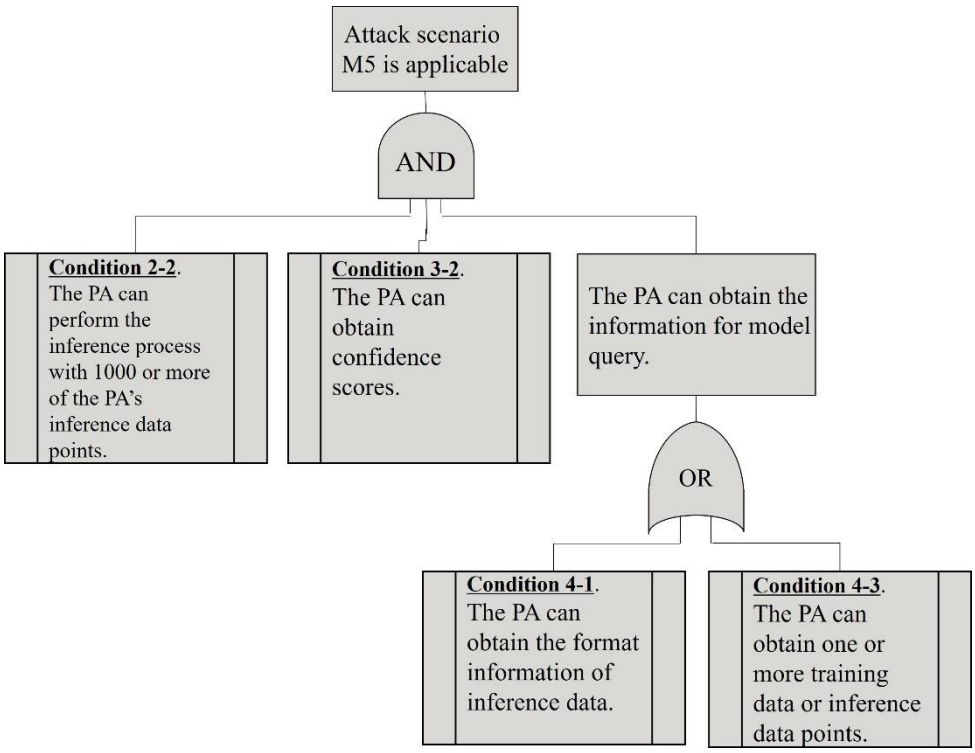

Figure II- 31. Attack tree and attack executable conditions for attack scenario M5 of membership inference

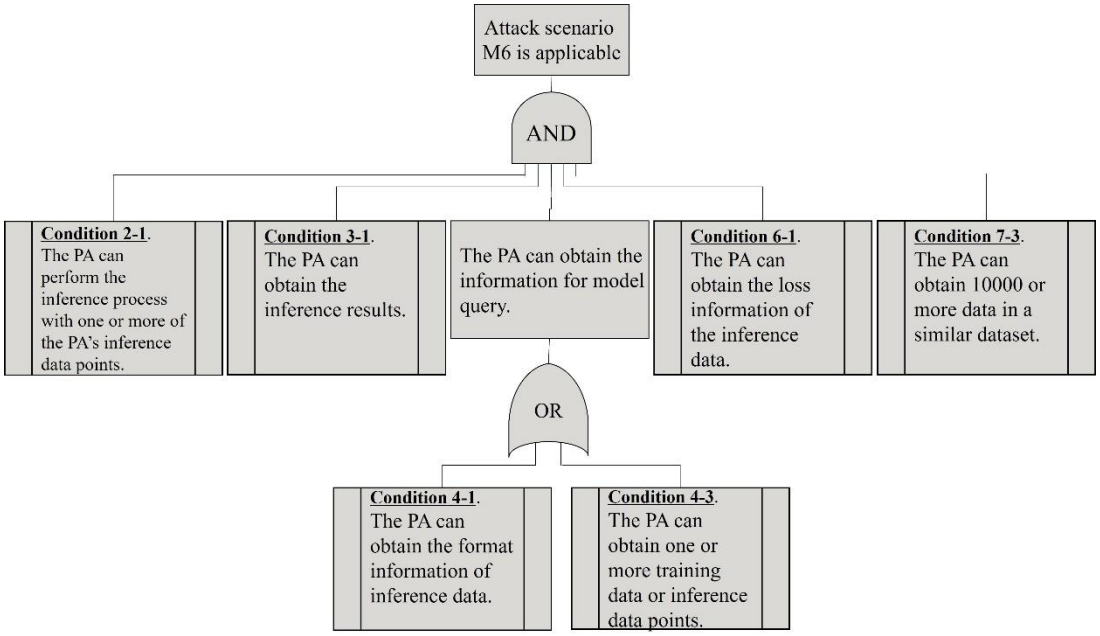

Figure II- 32. Attack tree and attack executable conditions for attack scenario M6 of membership inference

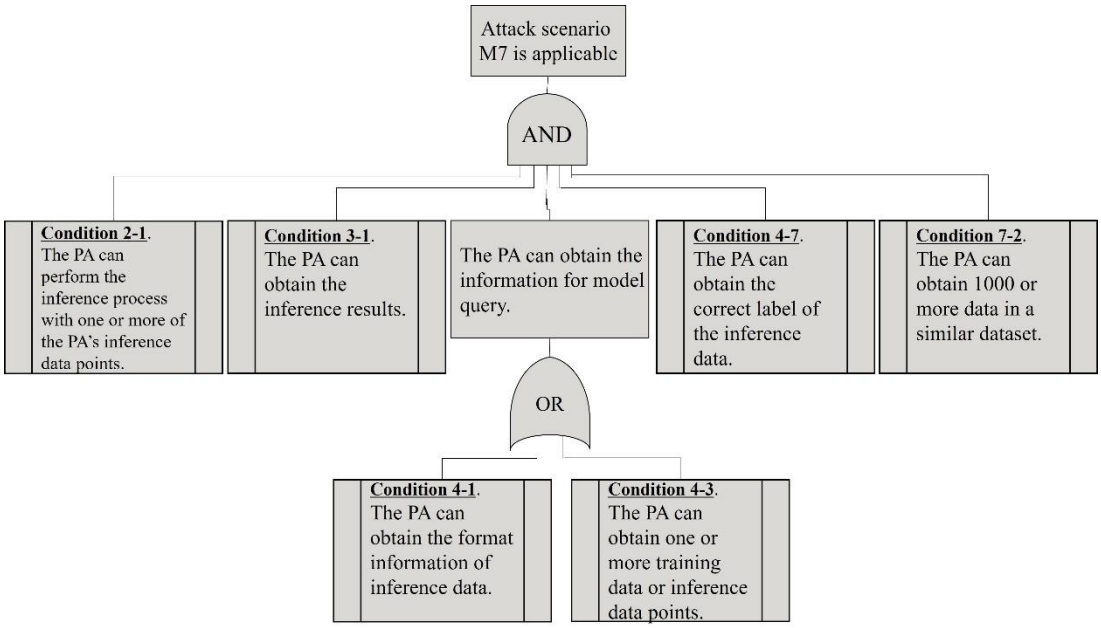

Figure II- 33. Attack tree and attack executable conditions for attack scenario M7 of membership inference

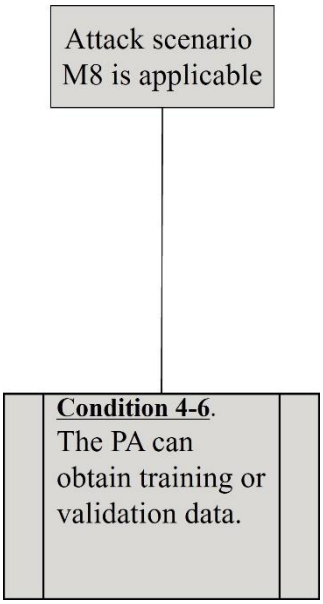

Figure II- 34. Attack tree and attack executable conditions for attack scenario M8 of membership inference

II-7.3. Selective Questions

Selective questions for membership inference are added to confirm whether the attack executable conditions shown in Section II-7.2 are satisfied. Additionally, the assessment was improved for the assessor to easily understand. The questions after the improvement are as follows.

1. Questions about model training

   Please answer question 1-1A when the PA can train the AI system model using PA's data; otherwise, please answer question 1-1B.

[Question 1-1A] When the PA can train the model, can the PA do so using several PA-prepared data?

[Question 1-1B] When the PA cannot train the model or when training data are automatically input into the training process, can the PA insert several pieces of PA-prepared data into the original training data? (For example, all training objects are trained by taking pictures of the objects that go through the factory lane.)

2. Questions about inferring

a) Please answer questions 2-1A to 2-4A when the inference process of the AI system is performed by the PA.

b) When the inference process of the AI system is performed regardless of the PA's will, please answer questions 2-1B to 2-4B.

c) When the AI system meets a) and b), please answer questions 2-1A to 2-4B.

[Question 2-1A] Can the PA perform the inference process using one or more of the PA-prepared data points?

i. The number of data points means the number of data points to be input at one time. For example, the number of data points is the number of rows in a table dataset or the number of images in an image dataset.

[Question 2-2A] Can the PA perform the inference process using 1000 or more of the PA-prepared data points?

i. The amount of data can be calculated by considering the operation period of the AI system and the interval of each inference process.

ii. If the PA can make several user accounts for the AI system, the number of data points is calculated by summing the amount of data for each account.

[Question 2-3A] Can the PA perform the inference process by using 10000 or more of the PA-prepared data points?

[Question 2-4A] Can the PA perform the inference process by using 1000000 or more of the PA-prepared data points?

[Question 2-1B] Can the PA insert one or more of the PA-prepared data points into the data of the inference target or replace one or more data points of the inference target with PA-prepared data when the inference process is performed?

i. The number of data points means the number of data points to be input at one time. For example, the number of data points is the number of rows in a table dataset or the number of images in an image dataset.

[Question 2-2B] Can the PA insert 1000 or more of the PA-prepared data points into the data of the inference target or replace 1000 or more data points of the inference target with PA-prepared data when the inference process is performed?

i. The amount of data can be calculated by considering the operation period of the AI system and the interval of each inference process.

[Question 2-3B] Can the PA insert 10000 or more of the PA-prepared data points into the data of the inference target or replace 10000 or more data points of the inference target with PA-prepared data when the inference process is performed?

[Question 2-4B] Can the PA insert 1000000 or more of the PA-prepared data points into the data of the inference target or replace 1000000 or more data points of the inference target with PA-prepared data when the inference process is performed?

3. Questions about the output of inference results

[Question 3-1] Does the AI system provide the inference results to the PA?

i. The inference result is the output result of the model. For example, the results are classification labels and a regression value when the model task is classification and logistic regression, respectively.

[Question 3-2] Does the AI system provide one or more confidence scores to the PA?

4. Questions about obtaining information on the system data

[Question 4-1] Can the PA know the format information (data type (table data or image data), matrix dimensions, or image data size) of inference target data or training data that can be used to prepare training data or inference target data?

i. The format information is the input format of the AI system. For example, it is the number of rows and columns and the order of the elements in the rows and the columns when the AI system treats a table dataset, and the number of vertical and horizontal pixels when the AI system treats an image dataset.

[Question 4-2] Can the PA know the statistical information of the training data?

i. The statistical information is the average or variance of each column in the table dataset.

[Question 4-3] Can the PA obtain one or more original data points (training, validation, or inference data points) of the AI system?

i. The number of data points is the number of rows in a table dataset, or the number of images in an image dataset.

ii. The answer is “yes” when the PA can prepare the inference target data. For example, when the PA knows the task of the MLS and the format information, the PA can create inference target data; therefore, the answer is “yes”.

[Question 4-4] Can the PA obtain 1000 or more original data points (training, validation, or inference data points) of the AI system?

i. When the sum of the number of original data points obtained from some PAs is 1000 or more, collusion should be considered, and the answer is “yes”.

[Question 4-5] Can the PA obtain 10000 or more original data points (training, validation, or inference data points) of the AI system?

i. When the sum of the number of original data points obtained from some PAs is 10000 or more, collusion should be considered, and the answer is “yes”.

[Question 4-6] Can the PA obtain one or more training, validation, or test data points?

i. The number of data points means the number of rows in a table dataset, or the number of images in an image dataset.

[Question 4-7] Can the PA obtain the correct input data label?

i. The correct label is the ground truth of the input data.

5. Question about reusing other models

[Question 5-1] Is the AI system constructed by diverting other trained models? (Is the model constructed using the transferability?)

i. The answer is “yes” when the model is constructed by reusing a model obtained from the Internet or an untrusted source.

6. Questions about obtaining information about the system model

[Question 6-1] Can the PA know the loss information of inference data?

i. The answer is “no” when the PA can obtain only the inference results or obtain no inference results.

ii. The answer is “yes” when the model has a function to obtain the loss information.

[Question 6-2] Can the PA know the gradient information of inference data?

i. The answer is “no” when the PA can obtain only the inference results or obtain no inference results.

ii. The answer is “yes” when the model has a function to obtain the gradient information.

[Question 6-3] Can the PA obtain the trained model of the AI system?

7. Questions about obtaining a similar dataset

[Question 7-1] Can the PA obtain one or more data points of a similar dataset (similar objective and similar genre) to the AI system?

i. The number of data points is the number of rows in a table dataset or the number of images in an image dataset.

[Question 7-2] Can the PA obtain 1000 or more data points of a similar dataset (similar objective and similar genre) to the AI system?

[Question 7-3] Can the PA obtain 10000 or more data points of a similar dataset (similar objective and similar genre) to the AI system?

8. Question about the data type treated in the AI system

[Question 8-1] Is the AI system constructed for a table dataset?

i. The answer is “yes” when the MLS has the pre-processes of inference and the data after the pre-processes are table data.

II-7.4. Judgment Table for Confirming the Satisfaction of the Attack Executable Conditions

Table II- 4 shows a judgment table for confirming whether the attack executable conditions extracted in Section II-7.2 are satisfied based on the answers to the selective questions shown in II-7.2.5. This table includes countermeasures for each condition, but if countermeasures corresponding to the conditions decided not to be satisfied are difficult to adopt, another condition for preventing the satisfaction of the attack trees should be derived, and this condition is set to FALSE. In addition, when it is difficult to take measures even if other conditions are selected, to use machine learning security-specific measures, consultation with experts is required.

Table II- 4. Example of a judgment table for confirming the satisfaction of attack executable conditions

Attack executable condition	Explanation	Conditions for TRUE	Result of judgment (TRUE/FALSE)	Suggestions for making this condition FALSE (Countermeasure).
Condition 1-1	The PA can perform the training process.	The answer to question 1-1A or question 1-1B is “Yes.”		Prevent the PA from performing training operations.
Condition 2-1	The PA can perform the inference process with one or more of the PA’s inference data points.	The answer to question 2-1A or question 2-1B is “Yes.”		Prevent the PA from performing the inference process.
Condition 2-2	The PA can perform the inference process with 1000 or more of the PA’s inference data points.	The answer to question 2-2A or question 2-2B is “Yes.”		Prevent the PA from performing the inference process with 1000 or more inference data points.
Condition 2-3	The PA can perform the inference process	The answer to question 2-3A		Prevent the PA from performing the inference process with 10000 or

	with 10000 or more of the PA’s inference data points.	or question 2-3B is “Yes.”		more inference data points.
Condition 2-4	The PA can perform the inference process with 1000000 or more of the PA’s inference data points.	The answer to question 2-4A or question 2-4B is “Yes.”		Prevent the PA from performing the inference process with 1000000 or more inference data points.
Condition 3-1	The PA can obtain the inference results.	The answer to question 3-1 or question 3-2 is “Yes.”		Prevent the PA from obtaining the inference results.
Condition 3-2	The PA can obtain confidence scores.	The answer to question 3-2 is “Yes.”		Prevent the PA from obtaining confidence scores.
Condition 4-1	The PA can obtain the format information of inference data.	The answer to question 4-1 is “Yes.”		Prevent the PA from obtaining and estimating the format information of inference data.
Condition 4-2	The PA can obtain the statistical information of training data.	The answer to question 4-2 is “Yes.”		Prevent the PA from obtaining and estimating the training-related data and the statistical information of the training data.
Condition 4-3	The PA can obtain one or more training data or inference data points.	The answer to question 4-3 is “Yes.”		Prevent the PA from obtaining training-related data or inferred data used in the past. Prevent the PA from obtaining or estimating the task information of the AI system and the format

				information of inference data.
Condition 4-4	The PA can obtain 1000 or more training data or inference data points.	The answer to question 4-4 is “Yes.”		Prevent the PA from obtaining training-related data or inferred data used in the past. Prevent the PA from obtaining or estimating the task information of the AI system and the format information of inference data.
Condition 4-5	The PA can obtain 10000 or more training data or inference data points.	The answer to question 4-5 is “Yes.”		Prevent the PA from obtaining training-related data or inferred data used in the past. Prevent the PA from obtaining or estimating the task information of the AI system and the format information of inference data.
Condition 4-6	The PA can obtain training or validation data.	The answer to question 4-6 is “Yes.”		Prevent the PA from obtaining training or validation data.
Condition 4-7	The PA can obtain the correct label of the inference data.	The answer to question 4-7 is “Yes.”		Do not expose detailed system specifications and prevent the PA from guessing the true correct label (ground truth)
Condition 5-1	The AI system is developed with internal or external models.	The answer to question 5-1 is “Yes.”		Only models from trusted sources are used in the system. Otherwise, do not use model transferability.

Condition 6-1	The PA can obtain the loss information of the inference data.	The answer to question 6-1 is “Yes.”		Prevent the PA from obtaining the loss information of the inference data.
Condition 6-2	The PA can obtain the gradient information of the inference data.	The answer to question 6-2 is “Yes.”		Prevent the PA from obtaining the gradient information of the inference data.
Condition 6-3	The PA can obtain the trained model.	The answer to question 6-3 is “Yes.”		The trained model should be strictly managed to ensure that it does not leak outside.
Condition 7-1	The PA can obtain one or more data points in a similar dataset.	The answer to question 7-1 is “Yes.”		Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.
Condition 7-2	The PA can obtain 1000 or more data in a similar dataset.	The answer to question 7-2 is “Yes.”		Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.
Condition 7-3	The PA can obtain 10000 or more data in a similar dataset.	The answer to question 7-3 is “Yes.”		Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.
Condition 8-1	The treated dataset is the table dataset.	The answer to question 8-1 is “Yes.”		Check whether it is an AI system that handles table data or image data, and whether the handled data are appropriate.

II-7.5. Risk Assessment Tool

The realization example of the risk assessment method described in this chapter as a software tool is published together with this document. This tool is implemented on Microsoft Excel. When Sheet I (Specification of AI system) and Sheet II (Questionnaire) are completed, the assessment result is displayed after Sheet IV (Result of Assessment). Please see the readme document of the tool and the instructions in the sheet of this tool for more information.

II-8. Case Studies of the Risk Assessment Method

This chapter introduces several case studies of analysis of machine learning systems using the realization example introduced in Chapter II-7.

II-8.1. Overview of Case Studies

In [II-6], a case study for road sign identification AI is given. In addition to this study, the committee members conducted case studies on three use cases. The use cases are given as follows. The results of these case studies are presented in this chapter.

- Loan review AI
- Plant control AI
- Gender and age estimation AI

II-8.1.1. Load Review AI

The specification of AI: To predict whether loan applicants will be able to repay.

A data processing technician trains financial information and information of a loan applicant to construct a model. A financial officer provides information about loan applicants, and the AI predicts (categorizes) whether the loan applicant can repay. Only the financial officer can know the inference results. No results will be shown to loan applicants.

1. PA = data processing technician

(i) Answers to selective questions

Question No	Question	Answer
1-1A	When the PA can train the model, can the PA do so using several PA-prepared data?	Yes
1-1B	When the PA cannot train the model or when training data are automatically input into the training process, can the PA insert several pieces of PA-prepared data into the original training data?	-
2-1A	Can the PA perform the inference process using one or more of the PA-prepared data points?	Yes
2-2A	Can the PA perform the inference process using 1000 or more of the PA-prepared data points?	Yes
2-3A	Can the PA perform the inference process by using 10000 or more of the PA-prepared data points?	Yes
2-4A	Can the PA perform the inference process by using 1000000 or more of the PA-prepared data points?	Yes

2-1B	Can the PA insert one or more of the PA-prepared data points into the data of the inference target or replace one or more data points of the inference target with PA-prepared data when the inference process is performed?	-
2-2B	Can the PA insert 1000 or more of the PA-prepared data points into the data of the inference target or replace 1000 or more data points of the inference target with PA-prepared data when the inference process is performed?	-
2-3B	Can the PA insert 10000 or more of the PA-prepared data points into the data of the inference target or replace 10000 or more data points of the inference target with PA-prepared data when the inference process is performed?	-
2-4B	Can the PA insert 1000000 or more of the PA-prepared data points into the data of the inference target or replace 1000000 or more data points of the inference target with PA-prepared data when the inference process is performed?	-
3-1	Does the AI system provide the inference results to the PA?	Yes
3-2	Does the AI system provide one or more confidence scores to the PA?	Yes
4-1	Can the PA know the format information (data type (table data or image data), matrix dimensions, or image data size) of inference target data or training data that can be used to prepare training data or inference target data?	Yes
4-2	Can the PA know the statistical information of the training data?	Yes
4-3	Can the PA obtain one or more original data points (training, validation, or inference data points) of the AI system?	Yes
4-4	Can the PA obtain 1000 or more original data points (training, validation, or inference data points) of the AI system?	Yes
4-5	Can the PA obtain 10000 or more original data points (training, validation, or inference data points) of the AI system?	Yes
4-6	Can the PA obtain one or more training, validation, or test data points?	Yes
4-7	Can the PA obtain the correct input data label?	Yes
5-1	Is the AI system constructed by diverting other trained models?	No
6-1	Can the PA know the loss information of inference data?	Yes
6-2	Can the PA know the gradient information of inference data?	Yes
6-3	Can the PA obtain the trained model of the AI system?	Yes

7-1	Can the PA obtain one or more data points of a similar dataset (similar objective and similar genre) to the AI system?	Yes
7-2	Can the PA obtain 1000 or more data points of a similar dataset (similar objective and similar genre) to the AI system?	Yes
7-3	Can the PA obtain 10000 or more data points of a similar dataset (similar objective and similar genre) to the AI system?	Yes
8-1	Is the AI system constructed for a table dataset?	Yes

(ii) Judgment of the satisfaction of the attack executable conditions

Attack executable condition	Explanation	Conditions for TRUE	Result of judgment (TRUE/FALSE)	Suggestions for making this condition FALSE (Countermeasure).
Condition 1-1	The PA can perform the training process.	The answer to question 1-1A or question 1-1B is “Yes.”	TRUE	Prevent the PA from performing training operations.
Condition 2-1	The PA can perform the inference process with one or more of the PA’s inference data points.	The answer to question 2-1A or question 2-1B is “Yes.”	TRUE	Prevent the PA from performing the inference process.
Condition 2-2	The PA can perform the inference process with 1000 or more of the PA’s inference data points.	The answer to question 2-2A or question 2-2B is “Yes.”	TRUE	Prevent the PA from performing the inference process with 1000 or more inference data points.
Condition 2-3	The PA can perform the inference process with	The answer to question 2-3A or question 2-3B is “Yes.”	TRUE	Prevent the PA from performing the inference process with 10000 or more inference data points.

	10000 or more of the PA’s inference data points.			
Condition 2-4	The PA can perform the inference process with 1000000 or more of the PA’s inference data points.	The answer to question 2-4A or question 2-4B is “Yes.”	TRUE	Prevent the PA from performing the inference process with 1000000 or more inference data points.
Condition 3-1	The PA can obtain the inference results.	The answer to question 3-1 or question 3-2 is “Yes.”	TRUE	Prevent the PA from obtaining the inference results.
Condition 3-2	The PA can obtain confidence scores.	The answer to question 3-2 is “Yes.”	TRUE	Prevent the PA from obtaining confidence scores.
Condition 4-1	The PA can obtain the format information of inference data.	The answer to question 4-1 is “Yes.”	TRUE	Prevent the PA from obtaining and estimating the format information of inference data.
Condition 4-2	The PA can obtain the statistical information of training data.	The answer to question 4-2 is “Yes.”	TRUE	Prevent the PA from obtaining and estimating the training-related data and the statistical information of the training data.
Condition 4-3	The PA can obtain one or more training data or	The answer to question 4-3 is “Yes.”	TRUE	Prevent the PA from obtaining training-related data or inferred data used in the past.

	inference data points.			Prevent the PA from obtaining or estimating the task information of the AI system and the format information of inference data.
Condition 4-4	The PA can obtain 1000 or more training data or inference data points.	The answer to question 4-4 is “Yes.”	TRUE	Prevent the PA from obtaining training-related data or inferred data used in the past. Prevent the PA from obtaining or estimating the task information of the AI system and the format information of inference data.
Condition 4-5	The PA can obtain 10000 or more training data or inference data points.	The answer to question 4-5 is “Yes.”	TRUE	Prevent the PA from obtaining training-related data or inferred data used in the past. Prevent the PA from obtaining or estimating the task information of the AI system and the format information of inference data.
Condition 4-6	The PA can obtain training or validation data.	The answer to question 4-6 is “Yes.”	TRUE	Prevent the PA from obtaining training or validation data.
Condition 4-7	The PA can obtain the correct label of the inference data.	The answer to question 4-7 is “Yes.”	TRUE	Do not expose detailed system specifications and prevent the PA from guessing the true correct label (ground truth)

Condition 5-1	The AI system is developed with internal or external models.	The answer to question 5-1 is “Yes.”	FALSE	Only models from trusted sources are used in the system. Otherwise, do not use model transferability.
Condition 6-1	The PA can obtain the loss information of the inference data.	The answer to question 6-1 is “Yes.”	TRUE	Prevent the PA from obtaining the loss information of the inference data.
Condition 6-2	The PA can obtain the gradient information of the inference data.	The answer to question 6-2 is “Yes.”	TRUE	Prevent the PA from obtaining the gradient information of the inference data.
Condition 6-3	The PA can obtain the trained model.	The answer to question 6-3 is “Yes.”	TRUE	The trained model should be strictly managed to ensure that it does not leak outside.
Condition 7-1	The PA can obtain one or more data points in a similar dataset.	The answer to question 7-1 is “Yes.”	TRUE	Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.
Condition 7-2	The PA can obtain 1000 or more data in a similar dataset.	The answer to question 7-2 is “Yes.”	TRUE	Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.
Condition 7-3	The PA can obtain 10000 or more data in a similar dataset.	The answer to question 7-3 is “Yes.”	TRUE	Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.

Condition 8-1	The treated dataset is the table dataset.	The answer to question 8-1 is “Yes.”	TRUE	Check whether it is an AI system that handles table data or image data, and whether the handled data are appropriate.

(iii) Satisfied attack scenarios (judgment by the attack trees)

Evasion attack (adversarial examples): A1, A2, A3, A4

Poisoning attack: P1, P3

Model Extraction: X1, X2, X3, X4, X6

Model Inversion: I1

Membership Inference: M1, M2, M3, M4, M5, M6, M7, M8

(iv) Assessment results

All scenarios except P2 and X5 were judged to be applicable. The authority of the data processing technician is almost identical to that of the AI system developer. Thus, this result is considered to have been derived because of his strong authority.

2. PA = financial officer

(i) Answers to selective questions

Question No	Question	Answer
1-1A	When the PA can train the model, can the PA do so using several PA-prepared data?	Yes
1-1B	When the PA cannot train the model or when training data are automatically input into the training process, can the PA insert several pieces of PA-prepared data into the original training data?	-
2-1A	Can the PA perform the inference process using one or more of the PA-prepared data points?	Yes
2-2A	Can the PA perform the inference process using 1000 or more of the PA-prepared data points?	Yes
2-3A	Can the PA perform the inference process by using 10000 or more of the PA-prepared data points?	No
2-4A	Can the PA perform the inference process by using 1000000 or more of the PA-prepared data points?	No

2-1B	Can the PA insert one or more of the PA-prepared data points into the data of the inference target or replace one or more data points of the inference target with PA-prepared data when the inference process is performed?	-
2-2B	Can the PA insert 1000 or more of the PA-prepared data points into the data of the inference target or replace 1000 or more data points of the inference target with PA-prepared data when the inference process is performed?	-
2-3B	Can the PA insert 10000 or more of the PA-prepared data points into the data of the inference target or replace 10000 or more data points of the inference target with PA-prepared data when the inference process is performed?	-
2-4B	Can the PA insert 1000000 or more of the PA-prepared data points into the data of the inference target or replace 1000000 or more data points of the inference target with PA-prepared data when the inference process is performed?	-
3-1	Does the AI system provide the inference results to the PA?	Yes
3-2	Does the AI system provide one or more confidence scores to the PA?	Yes
4-1	Can the PA know the format information (data type (table data or image data), matrix dimensions, or image data size) of inference target data or training data that can be used to prepare training data or inference target data?	Yes
4-2	Can the PA know the statistical information of the training data?	Yes
4-3	Can the PA obtain one or more original data points (training, validation, or inference data points) of the AI system?	Yes
4-4	Can the PA obtain 1000 or more original data points (training, validation, or inference data points) of the AI system?	Yes
4-5	Can the PA obtain 10000 or more original data points (training, validation, or inference data points) of the AI system?	Yes
4-6	Can the PA obtain one or more training, validation, or test data points?	Yes
4-7	Can the PA obtain the correct input data label?	Yes
5-1	Is the AI system constructed by diverting other trained models?	No
6-1	Can the PA know the loss information of inference data?	No
6-2	Can the PA know the gradient information of inference data?	No
6-3	Can the PA obtain the trained model of the AI system?	Yes

7-1	Can the PA obtain one or more data points of a similar dataset (similar objective and similar genre) to the AI system?	Yes
7-2	Can the PA obtain 1000 or more data points of a similar dataset (similar objective and similar genre) to the AI system?	Yes
7-3	Can the PA obtain 10000 or more data points of a similar dataset (similar objective and similar genre) to the AI system?	Yes
8-1	Is the AI system constructed for a table dataset?	Yes

(ii) Judgment of the satisfaction of the attack executable conditions

Attack executable condition	Explanation	Conditions for TRUE	Result of judgment (TRUE/FALSE)	Suggestions for making this condition FALSE (Countermeasure).
Condition 1-1	The PA can perform the training process.	The answer to question 1-1A or question 1-1B is “Yes.”	TRUE	Prevent the PA from performing training operations.
Condition 2-1	The PA can perform the inference process with one or more of the PA’s inference data points.	The answer to question 2-1A or question 2-1B is “Yes.”	TRUE	Prevent the PA from performing the inference process.
Condition 2-2	The PA can perform the inference process with 1000 or more of the PA’s inference data points.	The answer to question 2-2A or question 2-2B is “Yes.”	TRUE	Prevent the PA from performing the inference process with 1000 or more inference data points.
Condition 2-3	The PA can perform the inference process with	The answer to question 2-3A or question 2-3B is “Yes.”	FALSE	Prevent the PA from performing the inference process with 10000 or more inference data points.

	10000 or more of the PA’s inference data points.			
Condition 2-4	The PA can perform the inference process with 1000000 or more of the PA’s inference data points.	The answer to question 2-4A or question 2-4B is “Yes.”	FALSE	Prevent the PA from performing the inference process with 1000000 or more inference data points.
Condition 3-1	The PA can obtain the inference results.	The answer to question 3-1 or question 3-2 is “Yes.”	TRUE	Prevent the PA from obtaining the inference results.
Condition 3-2	The PA can obtain confidence scores.	The answer to question 3-2 is “Yes.”	TRUE	Prevent the PA from obtaining confidence scores.
Condition 4-1	The PA can obtain the format information of inference data.	The answer to question 4-1 is “Yes.”	TRUE	Prevent the PA from obtaining and estimating the format information of inference data.
Condition 4-2	The PA can obtain the statistical information of training data.	The answer to question 4-2 is “Yes.”	TRUE	Prevent the PA from obtaining and estimating the training-related data and the statistical information of the training data.
Condition 4-3	The PA can obtain one or more training data or	The answer to question 4-3 is “Yes.”	TRUE	Prevent the PA from obtaining training-related data or inferred data used in the past.

	inference data points.			Prevent the PA from obtaining or estimating the task information of the AI system and the format information of inference data.
Condition 4-4	The PA can obtain 1000 or more training data or inference data points.	The answer to question 4-4 is “Yes.”	TRUE	Prevent the PA from obtaining training-related data or inferred data used in the past. Prevent the PA from obtaining or estimating the task information of the AI system and the format information of inference data.
Condition 4-5	The PA can obtain 10000 or more training data or inference data points.	The answer to question 4-5 is “Yes.”	TRUE	Prevent the PA from obtaining training-related data or inferred data used in the past. Prevent the PA from obtaining or estimating the task information of the AI system and the format information of inference data.
Condition 4-6	The PA can obtain training or validation data.	The answer to question 4-6 is “Yes.”	TRUE	Prevent the PA from obtaining training or validation data.
Condition 4-7	The PA can obtain the correct label of the inference data.	The answer to question 4-7 is “Yes.”	TRUE	Do not expose detailed system specifications and prevent the PA from guessing the true correct label (ground truth)

Condition 5-1	The AI system is developed with internal or external models.	The answer to question 5-1 is “Yes.”	FALSE	Only models from trusted sources are used in the system. Otherwise, do not use model transferability.
Condition 6-1	The PA can obtain the loss information of the inference data.	The answer to question 6-1 is “Yes.”	FALSE	Prevent the PA from obtaining the loss information of the inference data.
Condition 6-2	The PA can obtain the gradient information of the inference data.	The answer to question 6-2 is “Yes.”	FALSE	Prevent the PA from obtaining the gradient information of the inference data.
Condition 6-3	The PA can obtain the trained model.	The answer to question 6-3 is “Yes.”	TRUE	The trained model should be strictly managed to ensure that it does not leak outside.
Condition 7-1	The PA can obtain one or more data points in a similar dataset.	The answer to question 7-1 is “Yes.”	TRUE	Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.
Condition 7-2	The PA can obtain 1000 or more data in a similar dataset.	The answer to question 7-2 is “Yes.”	TRUE	Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.
Condition 7-3	The PA can obtain 10000 or more data in a similar dataset.	The answer to question 7-3 is “Yes.”	TRUE	Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.

Condition 8-1	The treated dataset is the table dataset.	The answer to question 8-1 is “Yes.”	TRUE	Check whether it is an AI system that handles table data or image data, and whether the handled data are appropriate.

(iii) Satisfied attack scenarios (judgment by the attack trees)

Evasion attack (adversarial examples): A1, A2, A3, A4

Poisoning attack: P1, P3

Model extraction: X4, X6

Model inversion: I1

Membership inference: M3, M4, M5, M7, M8

(iv) Assessment results

The answer to question 1 was marked “Yes” because a financial officer could reflect the results in financial information. Because much inference processing could not be performed outside his work, fewer scenarios could be applied compared to the case of the data processing technician.

3. PA = other people (ex: loan applicant)

(i) Answers to selective questions

Question No	Question	Answer
1-1A	When the PA can train the model, can the PA do so using several PA-prepared data?	-
1-1B	When the PA cannot train the model or when training data are automatically input into the training process, can the PA insert several pieces of PA-prepared data into the original training data?	No
2-1A	Can the PA perform the inference process using one or more of the PA-prepared data points?	Yes
2-2A	Can the PA perform the inference process using 1000 or more of the PA-prepared data points?	No
2-3A	Can the PA perform the inference process by using 10000 or more of the PA-prepared data points?	No
2-4A	Can the PA perform the inference process by using 1000000 or more of the PA-prepared data points?	No

2-1B	Can the PA insert one or more of the PA-prepared data points into the data of the inference target or replace one or more data points of the inference target with PA-prepared data when the inference process is performed?	-
2-2B	Can the PA insert 1000 or more of the PA-prepared data points into the data of the inference target or replace 1000 or more data points of the inference target with PA-prepared data when the inference process is performed?	-
2-3B	Can the PA insert 10000 or more of the PA-prepared data points into the data of the inference target or replace 10000 or more data points of the inference target with PA-prepared data when the inference process is performed?	-
2-4B	Can the PA insert 1000000 or more of the PA-prepared data points into the data of the inference target or replace 1000000 or more data points of the inference target with PA-prepared data when the inference process is performed?	-
3-1	Does the AI system provide the inference results to the PA?	No
3-2	Does the AI system provide one or more confidence scores to the PA?	No
4-1	Can the PA know the format information (data type (table data or image data), matrix dimensions, or image data size) of inference target data or training data that can be used to prepare training data or inference target data?	No
4-2	Can the PA know the statistical information of the training data?	No
4-3	Can the PA obtain one or more original data points (training, validation, or inference data points) of the AI system?	No
4-4	Can the PA obtain 1000 or more original data points (training, validation, or inference data points) of the AI system?	No
4-5	Can the PA obtain 10000 or more original data points (training, validation, or inference data points) of the AI system?	No
4-6	Can the PA obtain one or more training, validation, or test data points?	No
4-7	Can the PA obtain the correct input data label?	No
5-1	Is the AI system constructed by diverting other trained models?	No
6-1	Can the PA know the loss information of inference data?	No
6-2	Can the PA know the gradient information of inference data?	No
6-3	Can the PA obtain the trained model of the AI system?	No

7-1	Can the PA obtain one or more data points of a similar dataset (similar objective and similar genre) to the AI system?	No
7-2	Can the PA obtain 1000 or more data points of a similar dataset (similar objective and similar genre) to the AI system?	No
7-3	Can the PA obtain 10000 or more data points of a similar dataset (similar objective and similar genre) to the AI system?	No
8-1	Is the AI system constructed for a table dataset?	Yes

(ii) Judgment of the satisfaction of the attack executable conditions

Attack executable condition	Explanation	Conditions for TRUE	Result of judgment (TRUE/FALSE)	Suggestions for making this condition FALSE (Countermeasure).
Condition 1-1	The PA can perform the training process.	The answer to question 1-1A or question 1-1B is “Yes.”	FALSE	Prevent the PA from performing training operations.
Condition 2-1	The PA can perform the inference process with one or more of the PA’s inference data points.	The answer to question 2-1A or question 2-1B is “Yes.”	TRUE	Prevent the PA from performing the inference process.
Condition 2-2	The PA can perform the inference process with 1000 or more of the PA’s inference data points.	The answer to question 2-2A or question 2-2B is “Yes.”	FALSE	Prevent the PA from performing the inference process with 1000 or more inference data points.
Condition 2-3	The PA can perform the inference process with	The answer to question 2-3A or question 2-3B is “Yes.”	FALSE	Prevent the PA from performing the inference process with 10000 or more inference data points.

	10000 or more of the PA’s inference data points.			
Condition 2-4	The PA can perform the inference process with 1000000 or more of the PA’s inference data points.	The answer to question 2-4A or question 2-4B is “Yes.”	FALSE	Prevent the PA from performing the inference process with 1000000 or more inference data points.
Condition 3-1	The PA can obtain the inference results.	The answer to question 3-1 or question 3-2 is “Yes.”	FALSE	Prevent the PA from obtaining the inference results.
Condition 3-2	The PA can obtain confidence scores.	The answer to question 3-2 is “Yes.”	FALSE	Prevent the PA from obtaining confidence scores.
Condition 4-1	The PA can obtain the format information of inference data.	The answer to question 4-1 is “Yes.”	FALSE	Prevent the PA from obtaining and estimating the format information of inference data.
Condition 4-2	The PA can obtain the statistical information of training data.	The answer to question 4-2 is “Yes.”	FALSE	Prevent the PA from obtaining and estimating the training-related data and the statistical information of the training data.
Condition 4-3	The PA can obtain one or more training data or	The answer to question 4-3 is “Yes.”	FALSE	Prevent the PA from obtaining training-related data or inferred data used in the past.

	inference data points.			Prevent the PA from obtaining or estimating the task information of the AI system and the format information of inference data.
Condition 4-4	The PA can obtain 1000 or more training data or inference data points.	The answer to question 4-4 is “Yes.”	FALSE	Prevent the PA from obtaining training-related data or inferred data used in the past. Prevent the PA from obtaining or estimating the task information of the AI system and the format information of inference data.
Condition 4-5	The PA can obtain 10000 or more training data or inference data points.	The answer to question 4-5 is “Yes.”	FALSE	Prevent the PA from obtaining training-related data or inferred data used in the past. Prevent the PA from obtaining or estimating the task information of the AI system and the format information of inference data.
Condition 4-6	The PA can obtain training or validation data.	The answer to question 4-6 is “Yes.”	FALSE	Prevent the PA from obtaining training or validation data.
Condition 4-7	The PA can obtain the correct label of the inference data.	The answer to question 4-7 is “Yes.”	FALSE	Do not expose detailed system specifications and prevent the PA from guessing the true correct label (ground truth)

Condition 5-1	The AI system is developed with internal or external models.	The answer to question 5-1 is “Yes.”	FALSE	Only models from trusted sources are used in the system. Otherwise, do not use model transferability.
Condition 6-1	The PA can obtain the loss information of the inference data.	The answer to question 6-1 is “Yes.”	FALSE	Prevent the PA from obtaining the loss information of the inference data.
Condition 6-2	The PA can obtain the gradient information of the inference data.	The answer to question 6-2 is “Yes.”	FALSE	Prevent the PA from obtaining the gradient information of the inference data.
Condition 6-3	The PA can obtain the trained model.	The answer to question 6-3 is “Yes.”	FALSE	The trained model should be strictly managed to ensure that it does not leak outside.
Condition 7-1	The PA can obtain one or more data points in a similar dataset.	The answer to question 7-1 is “Yes.”	FALSE	Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.
Condition 7-2	The PA can obtain 1000 or more data in a similar dataset.	The answer to question 7-2 is “Yes.”	FALSE	Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.
Condition 7-3	The PA can obtain 10000 or more data in a similar dataset.	The answer to question 7-3 is “Yes.”	FALSE	Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.

Condition 8-1	The treated dataset is the table dataset.	The answer to question 8-1 is “Yes.”	TRUE	Check whether it is an AI system that handles table data or image data, and whether the handled data are appropriate.

(iii) Satisfied attack scenarios (judgment by the attack trees)

Evasion attack (adversarial examples): No scenario can be applicable

Poisoning attack: No scenario can be applicable

Model extraction: No scenario can be applicable

Model inversion: No scenario can be applicable

Membership inference: No scenario can be applicable

(iv) Assessment results

Third parties, including loan applicants, can perform the inference process of the AI system indirectly and enter their data. However, the results cannot be obtained. Therefore, all assumed attack scenarios were determined to be difficult to conduct.

II-8.1.2.Plant Control AI

The specification of AI: To determine the oxygen supply to the plant.

The oxygen supply amount is determined based on the information obtained from sensors. Plant-related person conduct the training process. The inference process is performed periodically and does not involve humans. At this time, whether an outside attacker can perform the attack was analyzed. The attacker sends data to the AI system by replacing the sensors with his self-made sensors. In addition, assuming that the plant is patrolled once a day, it was decided that abnormal data could only be transmitted for a maximum of 24 hours.

1. PA = outside person

(i) Answers to selective questions

Question No	Question	Answer
1-1A	When the PA can train the model, can the PA do so using several PA-prepared data?	-
1-1B	When the PA cannot train the model or when training data are automatically input into the training process, can the PA insert several	No

	pieces of PA-prepared data into the original training data?	
2-1A	Can the PA perform the inference process using one or more of the PA-prepared data points?	-
2-2A	Can the PA perform the inference process using 1000 or more of the PA-prepared data points?	-
2-3A	Can the PA perform the inference process by using 10000 or more of the PA-prepared data points?	-
2-4A	Can the PA perform the inference process by using 1000000 or more of the PA-prepared data points?	-
2-1B	Can the PA insert one or more of the PA-prepared data points into the data of the inference target or replace one or more data points of the inference target with PA-prepared data when the inference process is performed?	Yes
2-2B	Can the PA insert 1000 or more of the PA-prepared data points into the data of the inference target or replace 1000 or more data points of the inference target with PA-prepared data when the inference process is performed?	Yes
2-3B	Can the PA insert 10000 or more of the PA-prepared data points into the data of the inference target or replace 10000 or more data points of the inference target with PA-prepared data when the inference process is performed?	Yes
2-4B	Can the PA insert 1000000 or more of the PA-prepared data points into the data of the inference target or replace 1000000 or more data points of the inference target with PA-prepared data when the inference process is performed?	No
3-1	Does the AI system provide the inference results to the PA?	No
3-2	Does the AI system provide one or more confidence scores to the PA?	No
4-1	Can the PA know the format information (data type (table data or image data), matrix dimensions, or image data size) of inference target data or training data that can be used to prepare training data or inference target data?	No
4-2	Can the PA know the statistical information of the training data?	No
4-3	Can the PA obtain one or more original data points (training, validation, or inference data points) of the AI system?	Yes
4-4	Can the PA obtain 1000 or more original data points (training, validation, or inference data points) of the AI system?	Yes

4-5	Can the PA obtain 10000 or more original data points (training, validation, or inference data points) of the AI system?	Yes
4-6	Can the PA obtain one or more training, validation, or test data points?	No
4-7	Can the PA obtain the correct input data label?	No
5-1	Is the AI system constructed by diverting other trained models?	No
6-1	Can the PA know the loss information of inference data?	No
6-2	Can the PA know the gradient information of inference data?	No
6-3	Can the PA obtain the trained model of the AI system?	No
7-1	Can the PA obtain one or more data points of a similar dataset (similar objective and similar genre) to the AI system?	No
7-2	Can the PA obtain 1000 or more data points of a similar dataset (similar objective and similar genre) to the AI system?	No
7-3	Can the PA obtain 10000 or more data points of a similar dataset (similar objective and similar genre) to the AI system?	No
8-1	Is the AI system constructed for a table dataset?	Yes

(ii) Judgment of the satisfaction of the attack executable conditions

Attack executable condition	Explanation	Conditions for TRUE	Result of judgment (TRUE/FALSE)	Suggestions for making this condition FALSE (Countermeasure).
Condition 1-1	The PA can perform the training process.	The answer to question 1-1A or question 1-1B is “Yes.”	FALSE	Prevent the PA from performing training operations.
Condition 2-1	The PA can perform the inference process with one or more of the PA’s inference data points.	The answer to question 2-1A or question 2-1B is “Yes.”	TRUE	Prevent the PA from performing the inference process.
Condition 2-2	The PA can perform the inference	The answer to question 2-2A	TRUE	Prevent the PA from performing the inference

	process with 1000 or more of the PA’s inference data points.	or question 2-2B is “Yes.”		process with 1000 or more inference data points.
Condition 2-3	The PA can perform the inference process with 10000 or more of the PA’s inference data points.	The answer to question 2-3A or question 2-3B is “Yes.”	TRUE	Prevent the PA from performing the inference process with 10000 or more inference data points.
Condition 2-4	The PA can perform the inference process with 1000000 or more of the PA’s inference data points.	The answer to question 2-4A or question 2-4B is “Yes.”	FALSE	Prevent the PA from performing the inference process with 1000000 or more inference data points.
Condition 3-1	The PA can obtain the inference results.	The answer to question 3-1 or question 3-2 is “Yes.”	FALSE	Prevent the PA from obtaining the inference results.
Condition 3-2	The PA can obtain confidence scores.	The answer to question 3-2 is “Yes.”	FALSE	Prevent the PA from obtaining confidence scores.
Condition 4-1	The PA can obtain the format information of inference data.	The answer to question 4-1 is “Yes.”	FALSE	Prevent the PA from obtaining and estimating the format information of inference data.

Condition 4-2	The PA can obtain the statistical information of training data.	The answer to question 4-2 is “Yes.”	FALSE	Prevent the PA from obtaining and estimating the training-related data and the statistical information of the training data.
Condition 4-3	The PA can obtain one or more training data or inference data points.	The answer to question 4-3 is “Yes.”	TRUE	Prevent the PA from obtaining training-related data or inferred data used in the past. Prevent the PA from obtaining or estimating the task information of the AI system and the format information of inference data.
Condition 4-4	The PA can obtain 1000 or more training data or inference data points.	The answer to question 4-4 is “Yes.”	TRUE	Prevent the PA from obtaining training-related data or inferred data used in the past. Prevent the PA from obtaining or estimating the task information of the AI system and the format information of inference data.
Condition 4-5	The PA can obtain 10000 or more training data or inference data points.	The answer to question 4-5 is “Yes.”	TRUE	Prevent the PA from obtaining training-related data or inferred data used in the past. Prevent the PA from obtaining or estimating the task information of the AI system and the format information of inference data.

Condition 4-6	The PA can obtain training or validation data.	The answer to question 4-6 is “Yes.”	FALSE	Prevent the PA from obtaining training or validation data.
Condition 4-7	The PA can obtain the correct label of the inference data.	The answer to question 4-7 is “Yes.”	FALSE	Do not expose detailed system specifications and prevent the PA from guessing the true correct label (ground truth)
Condition 5-1	The AI system is developed with internal or external models.	The answer to question 5-1 is “Yes.”	FALSE	Only models from trusted sources are used in the system. Otherwise, do not use model transferability.
Condition 6-1	The PA can obtain the loss information of the inference data.	The answer to question 6-1 is “Yes.”	FALSE	Prevent the PA from obtaining the loss information of the inference data.
Condition 6-2	The PA can obtain the gradient information of the inference data.	The answer to question 6-2 is “Yes.”	FALSE	Prevent the PA from obtaining the gradient information of the inference data.
Condition 6-3	The PA can obtain the trained model.	The answer to question 6-3 is “Yes.”	FALSE	The trained model should be strictly managed to ensure that it does not leak outside.
Condition 7-1	The PA can obtain one or more data points in a similar dataset.	The answer to question 7-1 is “Yes.”	FALSE	Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.

Condition 7-2	The PA can obtain 1000 or more data in a similar dataset.	The answer to question 7-2 is “Yes.”	FALSE	Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.
Condition 7-3	The PA can obtain 10000 or more data in a similar dataset.	The answer to question 7-3 is “Yes.”	FALSE	Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.
Condition 8-1	The treated dataset is the table dataset.	The answer to question 8-1 is “Yes.”	TRUE	Check whether it is an AI system that handles table data or image data, and whether the handled data are appropriate.

(iii) Satisfied attack scenarios (judgment by the attack trees)

Evasion attack (adversarial examples): No scenario can be applicable

Poisoning attack: No scenario can be applicable

Model extraction: No scenario can be applicable

Model inversion: No scenario can be applicable

Membership inference: No scenario can be applicable

(iv) Assessment results

It is assumed that an outside attacker can tweak the external sensor of the plant and input the desired data. However, since the results were not shown to him, all assumed attack scenarios were judged to be difficult to conduct.

II-8.1.3.Gender and Age Estimation AI

The specification of AI: To predict the gender and age of people in recorded images

This system is a combination of the two AIs of object recognition and gender and age prediction. The training process was performed by manually labeling the recorded image. The training process was conducted by a reliable person. The inference process is performed by extracting an image from a camera image recorded in the store and inputting it into the model. The results are only shown to the analyst and can be used for promotional and other purposes.

1. PA = developer of AI system

(i) Answers to selective questions

Question No	Question	Answer
1-1A	When the PA can train the model, can the PA do so using several PA-prepared data?	Yes
1-1B	When the PA cannot train the model or when training data are automatically input into the training process, can the PA insert several pieces of PA-prepared data into the original training data?	-
2-1A	Can the PA perform the inference process using one or more of the PA-prepared data points?	Yes
2-2A	Can the PA perform the inference process using 1000 or more of the PA-prepared data points?	Yes
2-3A	Can the PA perform the inference process by using 10000 or more of the PA-prepared data points?	Yes
2-4A	Can the PA perform the inference process by using 1000000 or more of the PA-prepared data points?	Yes
2-1B	Can the PA insert one or more of the PA-prepared data points into the data of the inference target or replace one or more data points of the inference target with PA-prepared data when the inference process is performed?	-
2-2B	Can the PA insert 1000 or more of the PA-prepared data points into the data of the inference target or replace 1000 or more data points of the inference target with PA-prepared data when the inference process is performed?	-
2-3B	Can the PA insert 10000 or more of the PA-prepared data points into the data of the inference target or replace 10000 or more data points of the inference target with PA-prepared data when the inference process is performed?	-
2-4B	Can the PA insert 1000000 or more of the PA-prepared data points into the data of the inference target or replace 1000000 or more data points of the inference target with PA-prepared data when the inference process is performed?	-
3-1	Does the AI system provide the inference results to the PA?	Yes
3-2	Does the AI system provide one or more confidence scores to the PA?	Yes

4-1	Can the PA know the format information (data type (table data or image data), matrix dimensions, or image data size) of inference target data or training data that can be used to prepare training data or inference target data?	Yes
4-2	Can the PA know the statistical information of the training data?	Yes
4-3	Can the PA obtain one or more original data points (training, validation, or inference data points) of the AI system?	Yes
4-4	Can the PA obtain 1000 or more original data points (training, validation, or inference data points) of the AI system?	Yes
4-5	Can the PA obtain 10000 or more original data points (training, validation, or inference data points) of the AI system?	Yes
4-6	Can the PA obtain one or more training, validation, or test data points?	Yes
4-7	Can the PA obtain the correct input data label?	Yes
5-1	Is the AI system constructed by diverting other trained models?	Yes
6-1	Can the PA know the loss information of inference data?	Yes
6-2	Can the PA know the gradient information of inference data?	Yes
6-3	Can the PA obtain the trained model of the AI system?	Yes
7-1	Can the PA obtain one or more data points of a similar dataset (similar objective and similar genre) to the AI system?	Yes
7-2	Can the PA obtain 1000 or more data points of a similar dataset (similar objective and similar genre) to the AI system?	Yes
7-3	Can the PA obtain 10000 or more data points of a similar dataset (similar objective and similar genre) to the AI system?	Yes
8-1	Is the AI system constructed for a table dataset?	No

(ii) Judgment of the satisfaction of the attack executable conditions

Attack executable condition	Explanation	Conditions for TRUE	Result of judgment (TRUE/FALSE)	Suggestions for making this condition FALSE (Countermeasure).
Condition 1-1	The PA can perform the training process.	The answer to question 1-1A or question 1-1B is “Yes.”	TRUE	Prevent the PA from performing training operations.

Condition 2-1	The PA can perform the inference process with one or more of the PA’s inference data points.	The answer to question 2-1A or question 2-1B is “Yes.”	TRUE	Prevent the PA from performing the inference process.
Condition 2-2	The PA can perform the inference process with 1000 or more of the PA’s inference data points.	The answer to question 2-2A or question 2-2B is “Yes.”	TRUE	Prevent the PA from performing the inference process with 1000 or more inference data points.
Condition 2-3	The PA can perform the inference process with 10000 or more of the PA’s inference data points.	The answer to question 2-3A or question 2-3B is “Yes.”	TRUE	Prevent the PA from performing the inference process with 10000 or more inference data points.
Condition 2-4	The PA can perform the inference process with 1000000 or more of the PA’s inference data points.	The answer to question 2-4A or question 2-4B is “Yes.”	TRUE	Prevent the PA from performing the inference process with 1000000 or more inference data points.
Condition 3-1	The PA can obtain the inference results.	The answer to question 3-1 or question 3-2 is “Yes.”	TRUE	Prevent the PA from obtaining the inference results.

Condition 3-2	The PA can obtain confidence scores.	The answer to question 3-2 is “Yes.”	TRUE	Prevent the PA from obtaining confidence scores.
Condition 4-1	The PA can obtain the format information of inference data.	The answer to question 4-1 is “Yes.”	TRUE	Prevent the PA from obtaining and estimating the format information of inference data.
Condition 4-2	The PA can obtain the statistical information of training data.	The answer to question 4-2 is “Yes.”	TRUE	Prevent the PA from obtaining and estimating the training-related data and the statistical information of the training data.
Condition 4-3	The PA can obtain one or more training data or inference data points.	The answer to question 4-3 is “Yes.”	TRUE	Prevent the PA from obtaining training-related data or inferred data used in the past. Prevent the PA from obtaining or estimating the task information of the AI system and the format information of inference data.
Condition 4-4	The PA can obtain 1000 or more training data or inference data points.	The answer to question 4-4 is “Yes.”	TRUE	Prevent the PA from obtaining training-related data or inferred data used in the past. Prevent the PA from obtaining or estimating the task information of the AI system and the format information of inference data.

Condition 4-5	The PA can obtain 10000 or more training data or inference data points.	The answer to question 4-5 is “Yes.”	TRUE	Prevent the PA from obtaining training-related data or inferred data used in the past. Prevent the PA from obtaining or estimating the task information of the AI system and the format information of inference data.
Condition 4-6	The PA can obtain training or validation data.	The answer to question 4-6 is “Yes.”	TRUE	Prevent the PA from obtaining training or validation data.
Condition 4-7	The PA can obtain the correct label of the inference data.	The answer to question 4-7 is “Yes.”	TRUE	Do not expose detailed system specifications and prevent the PA from guessing the true correct label (ground truth)
Condition 5-1	The AI system is developed with internal or external models.	The answer to question 5-1 is “Yes.”	TRUE	Only models from trusted sources are used in the system. Otherwise, do not use model transferability.
Condition 6-1	The PA can obtain the loss information of the inference data.	The answer to question 6-1 is “Yes.”	TRUE	Prevent the PA from obtaining the loss information of the inference data.
Condition 6-2	The PA can obtain the gradient information of the inference data.	The answer to question 6-2 is “Yes.”	TRUE	Prevent the PA from obtaining the gradient information of the inference data.

Condition 6-3	The PA can obtain the trained model.	The answer to question 6-3 is “Yes.”	TRUE	The trained model should be strictly managed to ensure that it does not leak outside.
Condition 7-1	The PA can obtain one or more data points in a similar dataset.	The answer to question 7-1 is “Yes.”	TRUE	Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.
Condition 7-2	The PA can obtain 1000 or more data in a similar dataset.	The answer to question 7-2 is “Yes.”	TRUE	Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.
Condition 7-3	The PA can obtain 10000 or more data in a similar dataset.	The answer to question 7-3 is “Yes.”	TRUE	Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.
Condition 8-1	The treated dataset is the table dataset.	The answer to question 8-1 is “Yes.”	FALSE	Check whether it is an AI system that handles table data or image data, and whether the handled data are appropriate.

(iii) Satisfied attack scenarios (judgment by the attack trees)

Evasion attack (adversarial examples): A1, A2, A3, A4

Poisoning attack: P1, P2, P3

Model extraction: X1, X2, X3, X5, X6

Model inversion: I1

Membership inference: M1, M2, M3, M4, M5, M6, M7, M8

(iv) Assessment results

All scenarios except X4 were determined to be applicable. This result was derived from the developer's strong authority. X4 is a scenario when the treated dataset is a table dataset, so this scenario was difficult to be applicable.

2. PA = labeler

(i) Answers to selective questions

Question No	Question	Answer
1-1A	When the PA can train the model, can the PA do so using several PA-prepared data?	-
1-1B	When the PA cannot train the model or when training data are automatically input into the training process, can the PA insert several pieces of PA-prepared data into the original training data?	No
2-1A	Can the PA perform the inference process using one or more of the PA-prepared data points?	Yes
2-2A	Can the PA perform the inference process using 1000 or more of the PA-prepared data points?	Yes
2-3A	Can the PA perform the inference process by using 10000 or more of the PA-prepared data points?	Yes
2-4A	Can the PA perform the inference process by using 1000000 or more of the PA-prepared data points?	Yes
2-1B	Can the PA insert one or more of the PA-prepared data points into the data of the inference target or replace one or more data points of the inference target with PA-prepared data when the inference process is performed?	-
2-2B	Can the PA insert 1000 or more of the PA-prepared data points into the data of the inference target or replace 1000 or more data points of the inference target with PA-prepared data when the inference process is performed?	-
2-3B	Can the PA insert 10000 or more of the PA-prepared data points into the data of the inference target or replace 10000 or more data points of the inference target with PA-prepared data when the inference process is performed?	-
2-4B	Can the PA insert 1000000 or more of the PA-prepared data points into the data of the inference target or replace 1000000 or more data points of the inference target with PA-prepared data when the inference process is performed?	-
3-1	Does the AI system provide the inference results to the PA?	No

3-2	Does the AI system provide one or more confidence scores to the PA?	No
4-1	Can the PA know the format information (data type (table data or image data), matrix dimensions, or image data size) of inference target data or training data that can be used to prepare training data or inference target data?	No
4-2	Can the PA know the statistical information of the training data?	No
4-3	Can the PA obtain one or more original data points (training, validation, or inference data points) of the AI system?	No
4-4	Can the PA obtain 1000 or more original data points (training, validation, or inference data points) of the AI system?	No
4-5	Can the PA obtain 10000 or more original data points (training, validation, or inference data points) of the AI system?	No
4-6	Can the PA obtain one or more training, validation, or test data points?	No
4-7	Can the PA obtain the correct input data label?	Yes
5-1	Is the AI system constructed by diverting other trained models?	Yes
6-1	Can the PA know the loss information of inference data?	No
6-2	Can the PA know the gradient information of inference data?	No
6-3	Can the PA obtain the trained model of the AI system?	No
7-1	Can the PA obtain one or more data points of a similar dataset (similar objective and similar genre) to the AI system?	Yes
7-2	Can the PA obtain 1000 or more data points of a similar dataset (similar objective and similar genre) to the AI system?	Yes
7-3	Can the PA obtain 10000 or more data points of a similar dataset (similar objective and similar genre) to the AI system?	Yes
8-1	Is the AI system constructed for a table dataset?	No

(ii) Judgment of the satisfaction of the attack executable conditions

Attack executable condition	Explanation	Conditions for TRUE	Result of judgment (TRUE/FALSE)	Suggestions for making this condition FALSE (Countermeasure).
Condition 1-1	The PA can perform the training process.	The answer to question 1-1A or question 1-1B is “Yes.”	FALSE	Prevent the PA from performing training operations.

Condition 2-1	The PA can perform the inference process with one or more of the PA’s inference data points.	The answer to question 2-1A or question 2-1B is “Yes.”	TRUE	Prevent the PA from performing the inference process.
Condition 2-2	The PA can perform the inference process with 1000 or more of the PA’s inference data points.	The answer to question 2-2A or question 2-2B is “Yes.”	TRUE	Prevent the PA from performing the inference process with 1000 or more inference data points.
Condition 2-3	The PA can perform the inference process with 10000 or more of the PA’s inference data points.	The answer to question 2-3A or question 2-3B is “Yes.”	TRUE	Prevent the PA from performing the inference process with 10000 or more inference data points.
Condition 2-4	The PA can perform the inference process with 1000000 or more of the PA’s inference data points.	The answer to question 2-4A or question 2-4B is “Yes.”	TRUE	Prevent the PA from performing the inference process with 1000000 or more inference data points.
Condition 3-1	The PA can obtain the inference results.	The answer to question 3-1 or question 3-2 is “Yes.”	FALSE	Prevent the PA from obtaining the inference results.

Condition 3-2	The PA can obtain confidence scores.	The answer to question 3-2 is “Yes.”	FALSE	Prevent the PA from obtaining confidence scores.
Condition 4-1	The PA can obtain the format information of inference data.	The answer to question 4-1 is “Yes.”	FALSE	Prevent the PA from obtaining and estimating the format information of inference data.
Condition 4-2	The PA can obtain the statistical information of training data.	The answer to question 4-2 is “Yes.”	FALSE	Prevent the PA from obtaining and estimating the training-related data and the statistical information of the training data.
Condition 4-3	The PA can obtain one or more training data or inference data points.	The answer to question 4-3 is “Yes.”	FALSE	Prevent the PA from obtaining training-related data or inferred data used in the past. Prevent the PA from obtaining or estimating the task information of the AI system and the format information of inference data.
Condition 4-4	The PA can obtain 1000 or more training data or inference data points.	The answer to question 4-4 is “Yes.”	FALSE	Prevent the PA from obtaining training-related data or inferred data used in the past. Prevent the PA from obtaining or estimating the task information of the AI system and the format information of inference data.

Condition 4-5	The PA can obtain 10000 or more training data or inference data points.	The answer to question 4-5 is “Yes.”	FALSE	Prevent the PA from obtaining training-related data or inferred data used in the past. Prevent the PA from obtaining or estimating the task information of the AI system and the format information of inference data.
Condition 4-6	The PA can obtain training or validation data.	The answer to question 4-6 is “Yes.”	FALSE	Prevent the PA from obtaining training or validation data.
Condition 4-7	The PA can obtain the correct label of the inference data.	The answer to question 4-7 is “Yes.”	TRUE	Do not expose detailed system specifications and prevent the PA from guessing the true correct label (ground truth)
Condition 5-1	The AI system is developed with internal or external models.	The answer to question 5-1 is “Yes.”	TRUE	Only models from trusted sources are used in the system. Otherwise, do not use model transferability.
Condition 6-1	The PA can obtain the loss information of the inference data.	The answer to question 6-1 is “Yes.”	FALSE	Prevent the PA from obtaining the loss information of the inference data.
Condition 6-2	The PA can obtain the gradient information of the inference data.	The answer to question 6-2 is “Yes.”	FALSE	Prevent the PA from obtaining the gradient information of the inference data.

Condition 6-3	The PA can obtain the trained model.	The answer to question 6-3 is “Yes.”	FALSE	The trained model should be strictly managed to ensure that it does not leak outside.
Condition 7-1	The PA can obtain one or more data points in a similar dataset.	The answer to question 7-1 is “Yes.”	TRUE	Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.
Condition 7-2	The PA can obtain 1000 or more data in a similar dataset.	The answer to question 7-2 is “Yes.”	TRUE	Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.
Condition 7-3	The PA can obtain 10000 or more data in a similar dataset.	The answer to question 7-3 is “Yes.”	TRUE	Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.
Condition 8-1	The treated dataset is the table dataset.	The answer to question 8-1 is “Yes.”	FALSE	Check whether it is an AI system that handles table data or image data, and whether the handled data are appropriate.

(iii) Satisfied attack scenarios (judgment by the attack trees)

Evasion attack (adversarial examples): No scenario can be applicable

Poisoning attack: P2

Model extraction: No scenario can be applicable

Model inversion: No scenario can be applicable

Membership inference: No scenario can be applicable

(iv) Assessment results

Only P2 was judged as applicable. The PA only labels the training data and cannot obtain the inference results. Thus, most attacks were judged as difficult to conduct. P2 depends on the structure of the model, and the judgment was made with concern that the diverted part may have

been poisoned when an outer model was diverted to build the AI system.

3. PA = analyst

(i) Answers to selective questions

Question No	Question	Answer
1-1A	When the PA can train the model, can the PA do so using several PA-prepared data?	-
1-1B	When the PA cannot train the model or when training data are automatically input into the training process, can the PA insert several pieces of PA-prepared data into the original training data?	No
2-1A	Can the PA perform the inference process using one or more of the PA-prepared data points?	Yes
2-2A	Can the PA perform the inference process using 1000 or more of the PA-prepared data points?	Yes
2-3A	Can the PA perform the inference process by using 10000 or more of the PA-prepared data points?	Yes
2-4A	Can the PA perform the inference process by using 1000000 or more of the PA-prepared data points?	Yes
2-1B	Can the PA insert one or more of the PA-prepared data points into the data of the inference target or replace one or more data points of the inference target with PA-prepared data when the inference process is performed?	-
2-2B	Can the PA insert 1000 or more of the PA-prepared data points into the data of the inference target or replace 1000 or more data points of the inference target with PA-prepared data when the inference process is performed?	-
2-3B	Can the PA insert 10000 or more of the PA-prepared data points into the data of the inference target or replace 10000 or more data points of the inference target with PA-prepared data when the inference process is performed?	-
2-4B	Can the PA insert 1000000 or more of the PA-prepared data points into the data of the inference target or replace 1000000 or more data points of the inference target with PA-prepared data when the inference process is performed?	-

3-1	Does the AI system provide the inference results to the PA?	Yes
3-2	Does the AI system provide one or more confidence scores to the PA?	Yes
4-1	Can the PA know the format information (data type (table data or image data), matrix dimensions, or image data size) of inference target data or training data that can be used to prepare training data or inference target data?	No
4-2	Can the PA know the statistical information of the training data?	No
4-3	Can the PA obtain one or more original data points (training, validation, or inference data points) of the AI system?	Yes
4-4	Can the PA obtain 1000 or more original data points (training, validation, or inference data points) of the AI system?	No
4-5	Can the PA obtain 10000 or more original data points (training, validation, or inference data points) of the AI system?	No
4-6	Can the PA obtain one or more training, validation, or test data points?	Yes
4-7	Can the PA obtain the correct input data label?	No
5-1	Is the AI system constructed by diverting other trained models?	Yes
6-1	Can the PA know the loss information of inference data?	No
6-2	Can the PA know the gradient information of inference data?	No
6-3	Can the PA obtain the trained model of the AI system?	No
7-1	Can the PA obtain one or more data points of a similar dataset (similar objective and similar genre) to the AI system?	Yes
7-2	Can the PA obtain 1000 or more data points of a similar dataset (similar objective and similar genre) to the AI system?	Yes
7-3	Can the PA obtain 10000 or more data points of a similar dataset (similar objective and similar genre) to the AI system?	Yes
8-1	Is the AI system constructed for a table dataset?	No

(ii) Judgment of the satisfaction of the attack executable conditions

Attack executable condition	Explanation	Conditions for TRUE	Result of judgment (TRUE/FALSE)	Suggestions for making this condition FALSE (Countermeasure).
Condition 1-1	The PA can perform the	The answer to question 1-1A	FALSE	Prevent the PA from performing training operations.

	training process.	or question 1-1B is “Yes.”		
Condition 2-1	The PA can perform the inference process with one or more of the PA’s inference data points.	The answer to question 2-1A or question 2-1B is “Yes.”	TRUE	Prevent the PA from performing the inference process.
Condition 2-2	The PA can perform the inference process with 1000 or more of the PA’s inference data points.	The answer to question 2-2A or question 2-2B is “Yes.”	TRUE	Prevent the PA from performing the inference process with 1000 or more inference data points.
Condition 2-3	The PA can perform the inference process with 10000 or more of the PA’s inference data points.	The answer to question 2-3A or question 2-3B is “Yes.”	TRUE	Prevent the PA from performing the inference process with 10000 or more inference data points.
Condition 2-4	The PA can perform the inference process with 1000000 or more of the PA’s inference data points.	The answer to question 2-4A or question 2-4B is “Yes.”	TRUE	Prevent the PA from performing the inference process with 1000000 or more inference data points.

Condition 3-1	The PA can obtain the inference results.	The answer to question 3-1 or question 3-2 is “Yes.”	TRUE	Prevent the PA from obtaining the inference results.
Condition 3-2	The PA can obtain confidence scores.	The answer to question 3-2 is “Yes.”	TRUE	Prevent the PA from obtaining confidence scores.
Condition 4-1	The PA can obtain the format information of inference data.	The answer to question 4-1 is “Yes.”	FALSE	Prevent the PA from obtaining and estimating the format information of inference data.
Condition 4-2	The PA can obtain the statistical information of training data.	The answer to question 4-2 is “Yes.”	FALSE	Prevent the PA from obtaining and estimating the training-related data and the statistical information of the training data.
Condition 4-3	The PA can obtain one or more training data or inference data points.	The answer to question 4-3 is “Yes.”	TRUE	Prevent the PA from obtaining training-related data or inferred data used in the past. Prevent the PA from obtaining or estimating the task information of the AI system and the format information of inference data.
Condition 4-4	The PA can obtain 1000 or more training data or inference data points.	The answer to question 4-4 is “Yes.”	FALSE	Prevent the PA from obtaining training-related data or inferred data used in the past. Prevent the PA from obtaining or estimating the task information of the AI

				system and the format information of inference data.
Condition 4-5	The PA can obtain 10000 or more training data or inference data points.	The answer to question 4-5 is “Yes.”	FALSE	Prevent the PA from obtaining training-related data or inferred data used in the past. Prevent the PA from obtaining or estimating the task information of the AI system and the format information of inference data.
Condition 4-6	The PA can obtain training or validation data.	The answer to question 4-6 is “Yes.”	TRUE	Prevent the PA from obtaining training or validation data.
Condition 4-7	The PA can obtain the correct label of the inference data.	The answer to question 4-7 is “Yes.”	FALSE	Do not expose detailed system specifications and prevent the PA from guessing the true correct label (ground truth)
Condition 5-1	The AI system is developed with internal or external models.	The answer to question 5-1 is “Yes.”	TRUE	Only models from trusted sources are used in the system. Otherwise, do not use model transferability.
Condition 6-1	The PA can obtain the loss information of the inference data.	The answer to question 6-1 is “Yes.”	FALSE	Prevent the PA from obtaining the loss information of the inference data.
Condition 6-2	The PA can obtain the gradient information of	The answer to question 6-2 is “Yes.”	FALSE	Prevent the PA from obtaining the gradient information of the inference data.

	the inference data.			
Condition 6-3	The PA can obtain the trained model.	The answer to question 6-3 is “Yes.”	FALSE	The trained model should be strictly managed to ensure that it does not leak outside.
Condition 7-1	The PA can obtain one or more data points in a similar dataset.	The answer to question 7-1 is “Yes.”	TRUE	Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.
Condition 7-2	The PA can obtain 1000 or more data in a similar dataset.	The answer to question 7-2 is “Yes.”	TRUE	Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.
Condition 7-3	The PA can obtain 10000 or more data in a similar dataset.	The answer to question 7-3 is “Yes.”	TRUE	Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.
Condition 8-1	The treated dataset is the table dataset.	The answer to question 8-1 is “Yes.”	FALSE	Check whether it is an AI system that handles table data or image data, and whether the handled data are appropriate.

(iii) Satisfied attack scenarios (judgment by the attack trees)

Evasion attack (adversarial examples): A1, A2, A3, A4

Poisoning attack: P2

Model extraction: X2

Model inversion: I1

Membership inference: M1, M2, M3, M5, M8

(iv) Assessment results

All scenarios for evasion attacks (adversarial examples) were judged as applicable. This result is

derived because the analyst can perform the inference process an arbitrary number of times and obtain the inference result. P2 is an attack that can be performed when an outer model is diverted for building, and X2 is an attack that can be performed when training data can be obtained. It has also been suggested that several other attack scenarios can be applicable. Many scenarios for membership inference were determined to be feasible because the confidence scores were provided to the PA.

4. PA = people being recorded (people inside the store)

(i) Answers to selective questions

Question No	Question	Answer
1-1A	When the PA can train the model, can the PA do so using several PA-prepared data?	-
1-1B	When the PA cannot train the model or when training data are automatically input into the training process, can the PA insert several pieces of PA-prepared data into the original training data?	No
2-1A	Can the PA perform the inference process using one or more of the PA-prepared data points?	Yes
2-2A	Can the PA perform the inference process using 1000 or more of the PA-prepared data points?	Yes
2-3A	Can the PA perform the inference process by using 10000 or more of the PA-prepared data points?	Yes
2-4A	Can the PA perform the inference process by using 1000000 or more of the PA-prepared data points?	No
2-1B	Can the PA insert one or more of the PA-prepared data points into the data of the inference target or replace one or more data points of the inference target with PA-prepared data when the inference process is performed?	-
2-2B	Can the PA insert 1000 or more of the PA-prepared data points into the data of the inference target or replace 1000 or more data points of the inference target with PA-prepared data when the inference process is performed?	-
2-3B	Can the PA insert 10000 or more of the PA-prepared data points into the data of the inference target or replace 10000 or more data points of the inference target with PA-prepared data when the inference	-

	process is performed?	
2-4B	Can the PA insert 1000000 or more of the PA-prepared data points into the data of the inference target or replace 1000000 or more data points of the inference target with PA-prepared data when the inference process is performed?	-
3-1	Does the AI system provide the inference results to the PA?	No
3-2	Does the AI system provide one or more confidence scores to the PA?	No
4-1	Can the PA know the format information (data type (table data or image data), matrix dimensions, or image data size) of inference target data or training data that can be used to prepare training data or inference target data?	No
4-2	Can the PA know the statistical information of the training data?	No
4-3	Can the PA obtain one or more original data points (training, validation, or inference data points) of the AI system?	No
4-4	Can the PA obtain 1000 or more original data points (training, validation, or inference data points) of the AI system?	No
4-5	Can the PA obtain 10000 or more original data points (training, validation, or inference data points) of the AI system?	No
4-6	Can the PA obtain one or more training, validation, or test data points?	No
4-7	Can the PA obtain the correct input data label?	No
5-1	Is the AI system constructed by diverting other trained models?	Yes
6-1	Can the PA know the loss information of inference data?	No
6-2	Can the PA know the gradient information of inference data?	No
6-3	Can the PA obtain the trained model of the AI system?	No
7-1	Can the PA obtain one or more data points of a similar dataset (similar objective and similar genre) to the AI system?	Yes
7-2	Can the PA obtain 1000 or more data points of a similar dataset (similar objective and similar genre) to the AI system?	Yes
7-3	Can the PA obtain 10000 or more data points of a similar dataset (similar objective and similar genre) to the AI system?	Yes
8-1	Is the AI system constructed for a table dataset?	No

(ii) Judgment of the satisfaction of the attack executable conditions

Attack executable condition	Explanation	Conditions for TRUE	Result of judgment (TRUE/FALSE)	Suggestions for making this condition FALSE (Countermeasure).
Condition 1-1	The PA can perform the training process.	The answer to question 1-1A or question 1-1B is “Yes.”	FALSE	Prevent the PA from performing training operations.
Condition 2-1	The PA can perform the inference process with one or more of the PA’s inference data points.	The answer to question 2-1A or question 2-1B is “Yes.”	TRUE	Prevent the PA from performing the inference process.
Condition 2-2	The PA can perform the inference process with 1000 or more of the PA’s inference data points.	The answer to question 2-2A or question 2-2B is “Yes.”	TRUE	Prevent the PA from performing the inference process with 1000 or more inference data points.
Condition 2-3	The PA can perform the inference process with 10000 or more of the PA’s inference data points.	The answer to question 2-3A or question 2-3B is “Yes.”	TRUE	Prevent the PA from performing the inference process with 10000 or more inference data points.
Condition 2-4	The PA can perform the inference process with	The answer to question 2-4A or question 2-4B is “Yes.”	FALSE	Prevent the PA from performing the inference process with 1000000 or more inference data points.

	1000000 or more of the PA’s inference data points.			
Condition 3-1	The PA can obtain the inference results.	The answer to question 3-1 or question 3-2 is “Yes.”	FALSE	Prevent the PA from obtaining the inference results.
Condition 3-2	The PA can obtain confidence scores.	The answer to question 3-2 is “Yes.”	FALSE	Prevent the PA from obtaining confidence scores.
Condition 4-1	The PA can obtain the format information of inference data.	The answer to question 4-1 is “Yes.”	FALSE	Prevent the PA from obtaining and estimating the format information of inference data.
Condition 4-2	The PA can obtain the statistical information of training data.	The answer to question 4-2 is “Yes.”	FALSE	Prevent the PA from obtaining and estimating the training-related data and the statistical information of the training data.
Condition 4-3	The PA can obtain one or more training data or inference data points.	The answer to question 4-3 is “Yes.”	FALSE	Prevent the PA from obtaining training-related data or inferred data used in the past. Prevent the PA from obtaining or estimating the task information of the AI system and the format information of inference data.
Condition 4-4	The PA can obtain 1000 or more training	The answer to question 4-4 is “Yes.”	FALSE	Prevent the PA from obtaining training-related

	data or inference data points.			data or inferred data used in the past. Prevent the PA from obtaining or estimating the task information of the AI system and the format information of inference data.
Condition 4-5	The PA can obtain 10000 or more training data or inference data points.	The answer to question 4-5 is “Yes.”	FALSE	Prevent the PA from obtaining training-related data or inferred data used in the past. Prevent the PA from obtaining or estimating the task information of the AI system and the format information of inference data.
Condition 4-6	The PA can obtain training or validation data.	The answer to question 4-6 is “Yes.”	FALSE	Prevent the PA from obtaining training or validation data.
Condition 4-7	The PA can obtain the correct label of the inference data.	The answer to question 4-7 is “Yes.”	FALSE	Do not expose detailed system specifications and prevent the PA from guessing the true correct label (ground truth)
Condition 5-1	The AI system is developed with internal or external models.	The answer to question 5-1 is “Yes.”	TRUE	Only models from trusted sources are used in the system. Otherwise, do not use model transferability.
Condition 6-1	The PA can obtain the loss information of	The answer to question 6-1 is “Yes.”	FALSE	Prevent the PA from obtaining the loss information of the inference data.

	the inference data.			
Condition 6-2	The PA can obtain the gradient information of the inference data.	The answer to question 6-2 is “Yes.”	FALSE	Prevent the PA from obtaining the gradient information of the inference data.
Condition 6-3	The PA can obtain the trained model.	The answer to question 6-3 is “Yes.”	FALSE	The trained model should be strictly managed to ensure that it does not leak outside.
Condition 7-1	The PA can obtain one or more data points in a similar dataset.	The answer to question 7-1 is “Yes.”	TRUE	Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.
Condition 7-2	The PA can obtain 1000 or more data in a similar dataset.	The answer to question 7-2 is “Yes.”	TRUE	Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.
Condition 7-3	The PA can obtain 10000 or more data in a similar dataset.	The answer to question 7-3 is “Yes.”	TRUE	Prevent the PA from obtaining or estimating the purpose of the AI system and detailed specifications of the training-related data.
Condition 8-1	The treated dataset is the table dataset.	The answer to question 8-1 is “Yes.”	FALSE	Check whether it is an AI system that handles table data or image data, and whether the handled data are appropriate.

(iii) Satisfied attack scenarios (judgment by the attack trees)

Evasion attack (adversarial example): No scenario can be applicable

Poisoning attack: P2

Model extraction: No scenario can be applicable
Model inversion: No scenario can be applicable
Membership inference: No scenario can be applicable

(iv) Assessment results

Only P2 was judged as applicable. P2 can be applicable because of the structure of the model. P2 is an attack that could be performed if the diverted model were poisoned, and the PA cannot actually poison it. Thus, the PA is unlikely to attack.

II-9. Conclusion

In Part II, we introduce the assessment technology for AI developers to conduct threat analysis by themselves. In addition, realization examples of this method and trial results are shown. It is possible to determine whether an attack scenario corresponding to the satisfied attack tree can be applicably based on the information on the satisfied/unsatisfied nature of the attack tree obtained by this technology. To make the attack corresponding to the satisfied attack tree difficult to perform on the AI system, the conditions for making the attack tree unsatisfied are considered, and the specification is changed according to these conditions. This technology supports AI developers in assessing the security risks of their AI system, and it can be used to consider countermeasures or to consult with AI security experts on its assessment results as reference materials. If the proposed countermeasures obtained from this technology cannot be adopted for any reason, please consult with an AI security expert and consider machine learning-specific countermeasures. Future work on this technology should be expected to be extended to further attack scenarios. We would like to see this technology used to strengthen AI system security.

II-10. References

[II-1] The Conference toward AI Network Security, “AI Utilization Guidelines -Practical Reference for AI utilization”, https://www.soumu.go.jp/main_content/000658284.pdf

[II-2] K. Eykholt, I. Evtimov, E. Fernades, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, D. Song, “Robust Physical-World Attacks on Deep Learning Models”, CVPR 2018.

[II-3] European Union Agency for Cybersecurity (ENISA), “Artificial Intelligence Cybersecurity Challenges”, 2020. https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges

[II-4] Microsoft, “Threat Modeling AI/ML Systems and Dependencies”, 2019. https://docs.microsoft.com/en-us/security/engineering/threat-modeling-aiml

[II-5] J. Yajima, T. Shimizu, I. Morikawa, T. Okubo, “A Study on Analysis Method of AI Security in Machine Learning System”, 2021 Symposium on Cryptography and Information Security.

[II-6] J. Yajima, T. Oikawa, I. Morikawa, F. Kasahara, M. Inui, N. Yoshioka, “A Threat Analysis Method on Machine Learning Security for System Development Engineers”, 2022 Symposium on Cryptography and Information Security.

[II-7] M. Juuti, S. Szyller, S. Marchal, N. Asokan, “PRADA: Protecting against DNN Model Stealing Attacks”, the 4th IEEE European Symposium on Security and Privacy (EuroS&P 2019)

[II-8] T. Orekondy, B. Schiele. M. Fritz, “Knockoff Nets: Stealing Functionality of Black-Box Models”, arXiv https://arxiv.org/abs/1812.02766

[II-9] R. Shokri, M. Stronati, C. Song, V. Shmatikov, “Membership Inference Attacks Against Machine Learning Models”, 2017 IEEE Symposium on Security and Privacy (S&P).

[II-10] S. Yeom, I. Giacomelli, M. Fredrikson, S. Jha, “Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting”, 2018 IEEE 31st Computer Security Foundations Symposium (CSP).

[II-11] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, M. Backes, “ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models”, The Network and Distributed System Security 2019 (NDSS 2019).

[II-12] A. Sablayrolles, M. Douze, C. Schmid, Y. Ollivier, Herve Jegou, “White-box vs Black-box: Bayes Optimal Strategies for Membership Inference”, the 36th International Conference on Machine Learning.

[II-13] Z. Li, Y. Zhang, “Membership Leakage in Label-Only Exposures”, 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS’2021).

Members of the Editing Committee of Machine Learning System Security Guidelines

<Current Members>
Yoshikazu Hanatani (Toshiba Corporation)
Masazumi Hayashi (Teikyo Heisei University)
Maki Inui (Fujitsu Limited)
Fumiyoshi Kasahara (Fujitsu Limited)
Kei Kureishi (Toshiba Corporation)
Takao Okubo (Institute of Information Security)
Kentaro Tsuji (Fujitsu Limited)
Jun Yajima (Fujitsu Limited)
Nobukazu Yoshioka (Waseda University)

<Former Members>
Daiki Ichihara (NTT DATA Corporation)
Tomoko Kaneko (National Institute of Informatics)
Ikuya Morikawa (Fujitsu Limited)
Takanori Oikawa (Fujitsu Limited)
Kenji Taguchi (National Institute of Informatics)