Machine Learning System

Security Guidelines

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

機械学習工学研究会

MACHINE LEARNING SYSTEMS ENGINEERING

Contents

Machine Learning System
Security Guidelines,
Part I. “Security Measures Procedures”

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

MLSE 機械学習工学研究会 MACHINE LEARNING SYSTEMS ENGINEERING

I-1. Summary

I-1.1. Overview

This document organizes security procedures against machine learning-specific attacks (MLSAs) for developers and service providers of machine learning systems. The purpose of this document is to clarify effective countermeasures according to the situation and to help developers and service providers communicate with machine learning security experts when implementing security measures. This guideline is a self-assessment tool and does not necessarily need to be followed. Any damage resulting from the implementation of the contents of this document is not the responsibility of the authors.

These guidelines comprise "Security Measures Procedures," "Risk Assessment," and "Overview of Detection Techniques for Machine Learning-Specific Attacks" (Figure I- 1).

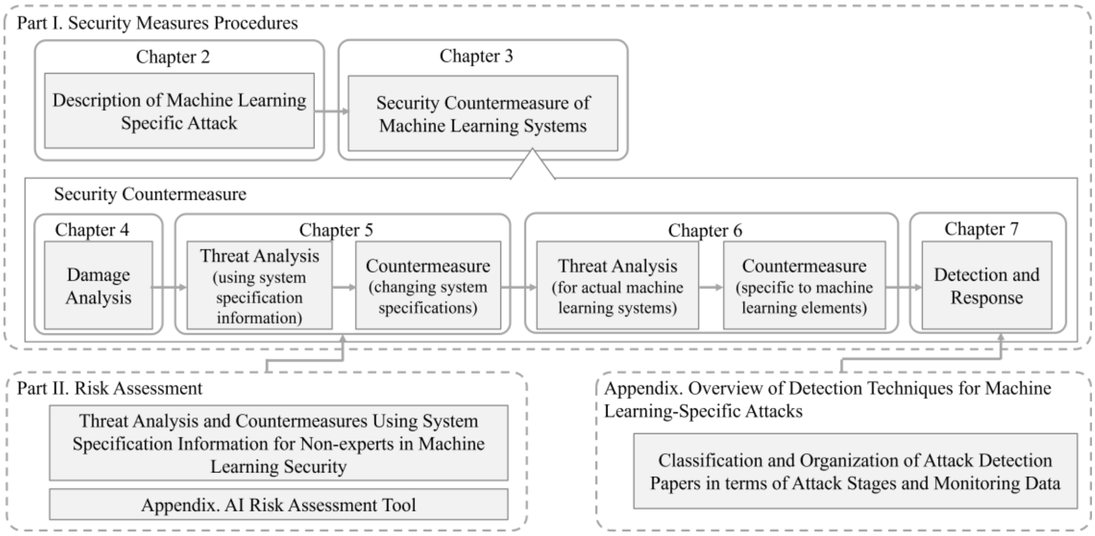

Figure I- 1. Overview of the Guidelines

In Part I, the security measures procedures, the procedures of security countermeasures against MLSAs are introduced. Chapter I-2 describes the MLSAs, and Chapter I-3 provides an overview of security countermeasures against MLSAs. Chapters I-4 to I-7 describe the specific implementation of each procedure introduced in Chapter I-3.

In Part II, the risk assessment, methods for system developers who may not have expertise in machine learning security to assess "threat analysis and countermeasures using system specification information" described in Chapter I-5 of Part I are introduced in detail.

The appendix, overview of detection techniques for machine learning-specific attacks, surveys papers related to detection technology against MLSAs, which are described in Chapter I-7 and classified and summarized in terms of "categories of attacks," "state of attack detection," and "data to be monitored." This document can be used as a reference when building a detection system.

I-1.2. Purposes and Background

Recently, with the development and widespread use of machine learning, many previously unobtainable machine learning-based systems with advanced functions, such as image recognition and natural language processing, have been developed (hereinafter called “AI systems”). In the future, such systems are expected to replace humans, such as self-driving cars and automated financial transactions. However, when they become an integral part of the social infrastructure, security risks increase.

Security assessments and countermeasures have been conducted even in conventional systems that do not use machine learning. However, machine learning has specific vulnerabilities, such as adversarial examples that intentionally malfunction train models; thus, security analysis and countermeasures specific to machine learning systems are required.

Since the discovery of the adversarial examples in 2014, extensive research has been conducted on MLSAs and defensive techniques against them. Although some papers collect research trends, the assumptions and evaluation metrics for the attacks and defenses proposed in each study have not been defined. Therefore, it is unclear how to determine what attacks are applicable and what countermeasures are necessary for system development. Furthermore, to analyze security specific to machine learning systems and implement countermeasures, special knowledge of machine learning and security is required. This requirement makes security measures difficult to implement in development workplaces where few machine learning security experts with knowledge in both areas are available.

Therefore, AI developers who may not know about information security need criteria (guidelines) to judge whether MLSAs will be applicable in their AI systems. These guidelines introduce typical attack methods and enable AI developers to judge whether the conditions required for those attacks to be executed on a developed AI system are satisfied. They will be able to determine whether they should work with machine learning security experts according to the results of their analysis of the likelihood of such attacks.

In organizing these guidelines, a Formulation Committee of Machine Learning System Security Guidelines was established in July 2021 within the Safety and Security Working Group on Machine Learning System of the Machine Learning Systems Engineering Research Group (MLSE). This committee comprises members from industry and academia selected from the public by this study group and aims to provide useful information for AI developers as guidelines. These guidelines summarize the results of the discussions and deliberations of this formulation committee.

Some governments and international organizations have published guidelines on developing and using AI. In 2019, the OECD released the multi-country agreed-upon AI Principles [I-1], which include inclusive growth, sustainable development and well-being, human-centered values and equity, transparency and accountability, robustness, security and safety, and accountability. In Japan, the

"Human-centered AI Society Principles" [I-2] set forth the principles of human-centeredness, education and literacy, ensuring privacy, ensuring security, ensuring fair competition, fairness, accountability and transparency, and innovation.

Although the various elements that comprise the AI Principles are grouped in different ways, they fall into eight themes: privacy, accountability, safety and security, transparency and accountability, fairness and non-discrimination, human control of technology, professional responsibility, and promotion of human values.

To put the above principles into practice, METI's "The State of AI Governance in Japan" [I-3] organized the structure of AI governance as shown in Figure I- 2.

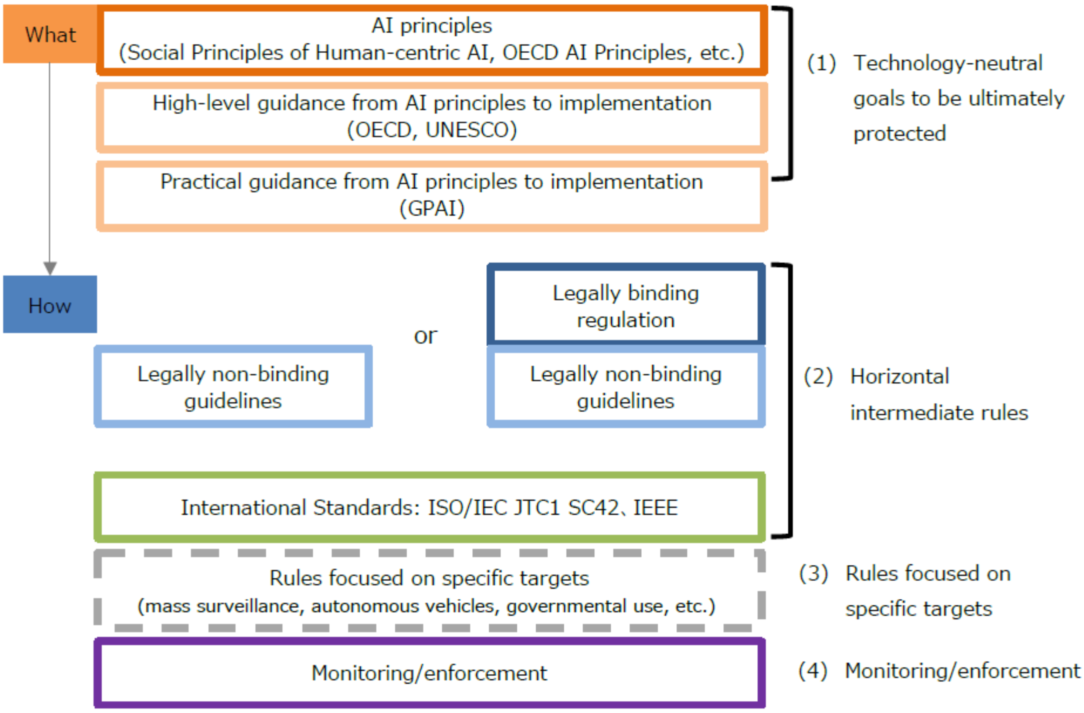

Figure I- 2. The structure of AI governance

These guidelines correspond to the "legally non-binding guidelines" shown in Figure I- 2. Additionally, there are an effort to emphasize "security" as addressed in the AI Principles.

I-1.3. Scope of the Guidelines

This section describes the intended audience for these guidelines, how the guidelines will be used, and their relationship to related literature.

I-1.3.1. Expected Target of the Guidelines

The intended audience for these guidelines includes developers of machine learning systems and service providers using machine learning systems who may not have expertise in machine learning security. In this section, what machine learning security experts should do about the security measures for machine learning systems and what non-experts should do are described separately.

Regarding the AI Utilization Guidelines of the Ministry of Internal Affairs and Communications [I-4], the flow of AI utilization is shown in Figure I- 3.

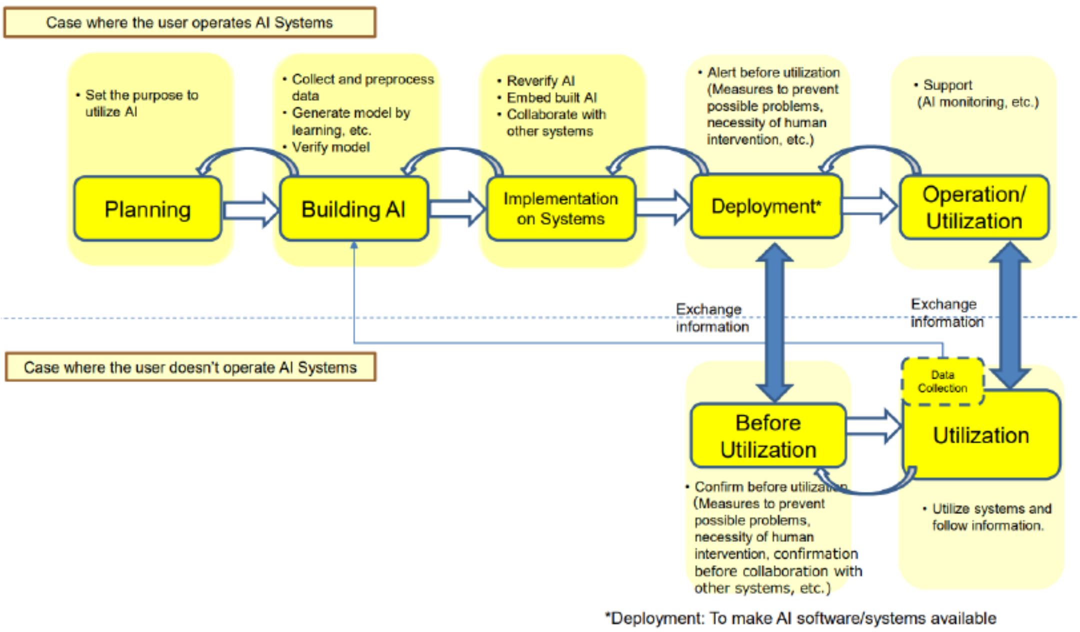

Figure I- 3. General flow of AI utilization for each entity

This document is intended for use in “planning,” “building AI,” and “operation/utilization.”

I-1.3.2. Relationship to Other Machine Learning Security Countermeasure Literature

Various studies on security countermeasures against MLSAs have been published by various government agencies and companies.

The ATLAS [I-5] published by MITRE organizes the tactics and methods of MLSAs. This report summarizes the methods used at each stage of an attack; such as the reconnaissance phase, initial access phase, attack execution phase, and information theft phase. It also summarizes actual attack examples as case studies.

For risk assessment of machine learning systems, ENISA's Artificial Intelligence Cybersecurity Challenges [I-6] discusses assets and damage associated with machine learning systems. Microsoft’s Threat Modeling AI/ML Systems and Dependencies [I-6] and ICO’s AI and data protection risk mitigation and management toolkit [I-8] also mention precautions to be taken when developing AI systems, and the documents help to connect the development system to a threat.

MLSAs and defenses against them have been published by organizations such as NIST [I-9], AIST [I-10], ENISA [I-6], and ICO [I-8], and companies such as Microsoft [I-6] and Mitsui Bussan Secure

Directions [I-11].

On the other hand, a series of security countermeasures in the existing security countermeasure literature for MLSAs had no systematic organization. Therefore, system developers and service providers implement security measures with difficulty.

Under these guidelines, security procedures are explained to developers and service providers of machine learning systems to help them understand what they need to do when implementing security measures, and to help them communicate with machine learning security experts when implementing security measures.

Additionally, to facilitate security measure performance, a method that enables “threat analysis and countermeasures using system specification information” in the security measures procedures to be performed even without expertise in machine learning security is introduced.

I-1.3.3. Relationship to General Information Security

This section describes the relationship between security against MLSAs and general information security. In these guidelines, “MLSAs” are defined as attacks that use machine learning system characteristics and have access to legitimate authority. For example, an attack that duplicates the model of the target machine learning system by repeatedly inputting data into the system and observing the output is included in the MLSAs. Security measures should be considered for general information security and security against MLSAs. It is desirable to consider countermeasures by combining the priorities of general information security and security against MLSAs, referring to the damage caused by attacks and the likelihood of attacks occurring. The ISO 27000 series, ISO/IEC 15408, NIST SP 800 series, NIST Cyber Security Framework, and the Information Technology Promotion Agency, Japan guidelines provide frameworks for general information security risk assessment and countermeasures.

I-1.4. Machine Learning Systems Used in These Guidelines

The target system in Part I is a system using machine learning. The machine learning processing part of a machine learning system generally comprises a training pipeline and an inference pipeline, as shown in Figure I- 4 and Figure I- 5. Some systems have only an inference pipeline because the training pipeline is performed externally. Before the machine learning system is operated, the training pipeline performs training processing using training-related data and generates trained models. Then, the inference pipeline performs inference processing using the inference target data and the trained models to obtain inference results.

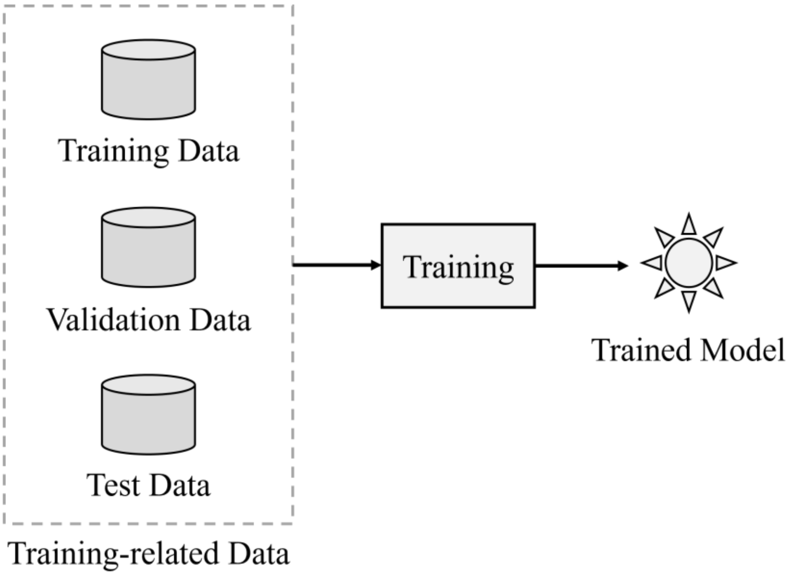

Figure I- 4. Training pipeline of a machine learning processing part

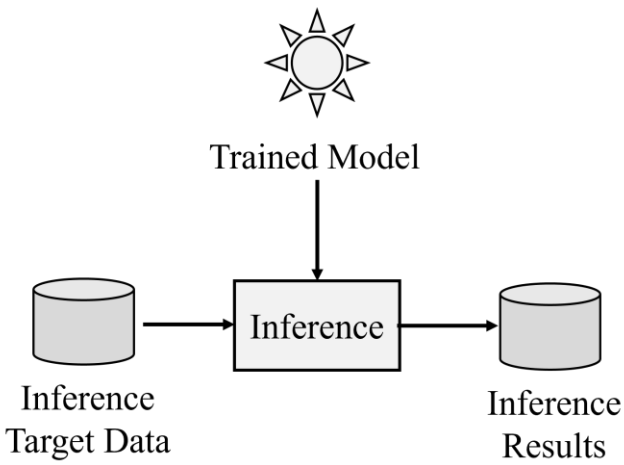

Figure I- 5. Inference pipeline of a machine learning processing part

I-2. Machine Learning Specific Attacks (MLSAs)

This chapter describes the threats posed by MLSAs and the specific attacks that pose threats and concludes with basic ideas for security measures.

MLSAs can be found in NIST’s NIST IR 8269 Draft [I-9], AIST’s Machine Learning Quality Management Guidelines [I-10], and Microsoft’s Threat Modeling AI/ML Systems and Dependencies [I-6].

I-2.1. Threats Specific to Machine Learning Systems

Threats specific to machine learning systems can be broadly classified into three categories. These threats can be caused by authorized access to the machine learning system.

- Model or system malfunctions
- Model theft
- Training data theft

I-2.1.1. Model or System Malfunctions

This threat is the malfunctioning of machine learning system models, which prevents the correct system operation. The malfunction of models and systems can lead to damage such as induced accidents in automatic driving systems (e.g., misidentification of sign recognition systems) and evading malware detection.

Two types of attacks can cause a model or system to malfunction: attacks that are performed by input to the model/system during inference (Section I-2.2.1) and attacks that contaminate training data and training models (Section I-2.2.2). In addition, attacks can not only intentionally change the decision results but also impair only the explanatory function of the model or the system. This result can affect the transparency of the system.

I-2.1.2. Model Theft

This attack is the theft of a model of a machine learning system by an attacker who creates a copy of the model or a model with similar performance. Model theft can lead not only to direct damage to Intellectual Property (IP), such as replicating services, but also to other attacks that use the thieved models.

One type of model theft attack is based on inputs to models and systems (Section I-2.2.3).

I-2.1.3. Training Data Theft

This threat is the theft by an attacker of data used to train models for machine learning systems or the theft of some of the information in the data used for training. This action can lead to privacy protection damage, such as personally identifiable information (PII) leakage.

One type of training data theft attack is based on inputs to models and systems (Section I-2.2.4, Section I-2.2.5).

I-2.2. Threat-causing Attacks

Five typical MLSAs cause the threats listed in Section I-2.1.

I-2.2.1. Evasion Attacks

Evasion attacks cause models and systems to malfunction (Section I-2.1.1).

Maliciously modifying the input data of a machine learning system can induce the system to behave in a way that was not intended. A well-known attack is called an adversarial example. This attack induces model misjudgments by adding a small amount of noise to the input data that are invisible to humans.

I-2.2.2. Poisoning Attacks

Poisoning attacks cause models and systems to malfunction (Section I-2.1.1).

An attacker can cause malfunctions by inserting crafted data and models into data and models used to train models of machine learning systems. This attack not only causes one label to be misjudged by another label, but also includes a backdoor attack that causes a given label to be misjudged by inputting data containing a specific pattern called a trigger.

I-2.2.3. Model Extraction Attacks

Model extraction attacks cause model theft (Section I-2.1.2).

This attack creates a model that has the same performance as the target system's model by analyzing the input-output relationship with the machine learning system.

I-2.2.4. Model Inversion Attacks

Model inversion attacks cause training data theft (Section I-2.1.3).

This attack reconstructs information contained in training data by analyzing the relationship between input and output data on machine learning systems.

I-2.2.5. Membership Inference Attacks

Membership inference attacks cause training data theft (Section I-2.1.3).

This attack identifies whether specific data are included in the training data of a model by analyzing the input and output data to the machine learning system. Unlike model inversion attacks, it does not reconstruct the training data.

I-3. Security of Machine Learning Systems

This chapter describes the procedures for security measures against MLSAs. In machine learning systems, security measures should be taken against MLSAs, in addition to conventional general information security. It is desirable to consider the priority of general information security and security against MLSAs, referring to the amount of damage and the possibility of attacks.

Notably, comprehensively understanding and taking countermeasures against all attacks is usually impossible.

I-3.1. Concept of Machine Learning Security

Machine learning security is considered in the same process as general information security measures, where the possibility of attacks is identified through “risk assessment” and then “countermeasures” are taken based on the results of that analysis. First, the impact of the threat is analyzed by identifying the assets related to the system and linking them to the threats that affect the assets (Section エラー! 参照元が見つかりません。). Next, threat analysis determines whether the system has vulnerabilities that allow attacks leading to threats (Section I-2.2).

Figure I- 6 shows the overall picture of the relationship between assets, threats, and attacks, damage analysis, threat analysis, and countermeasures in machine learning.

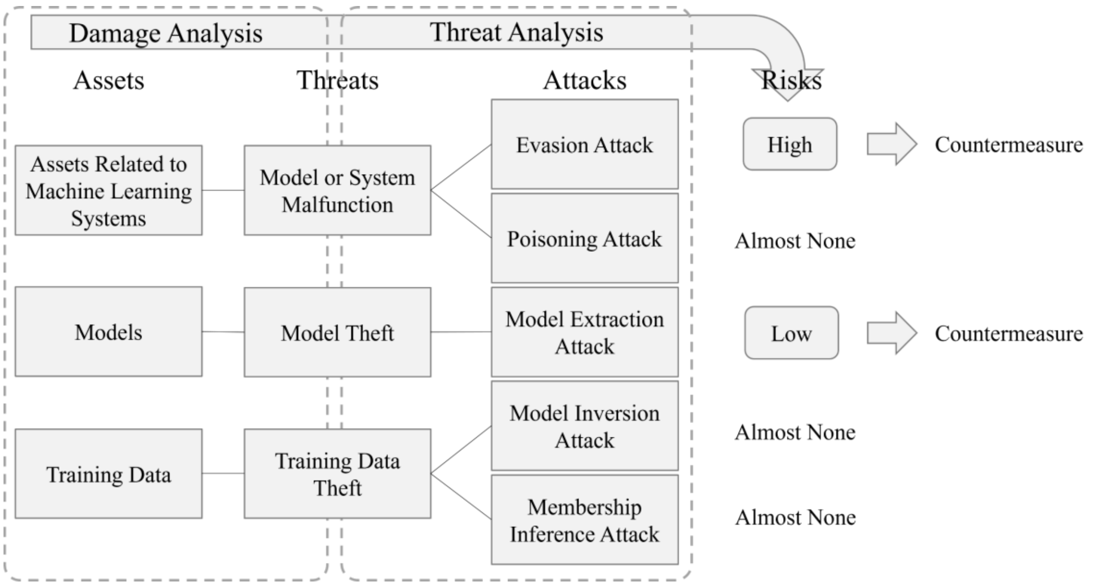

Figure I- 6. Overall picture of machine learning security

Countermeasures are divided into two major categories: “mitigation measures” to make attacks difficult or suppress their effects before operation/utilization, and “detection and response” to detect the occurrence of attacks during operation and take action. Basically, measures should be implemented through “mitigation measures” before operation as much as possible. “Detection and response” are performed for threats that cannot be prevented by mitigation measures, threats that require more focused countermeasures, and unknown threats that cannot be identified in advance.

I-3.2. Machine Learning Security Processes

Figure I- 7 shows an example of how to proceed with each process in the previous section.

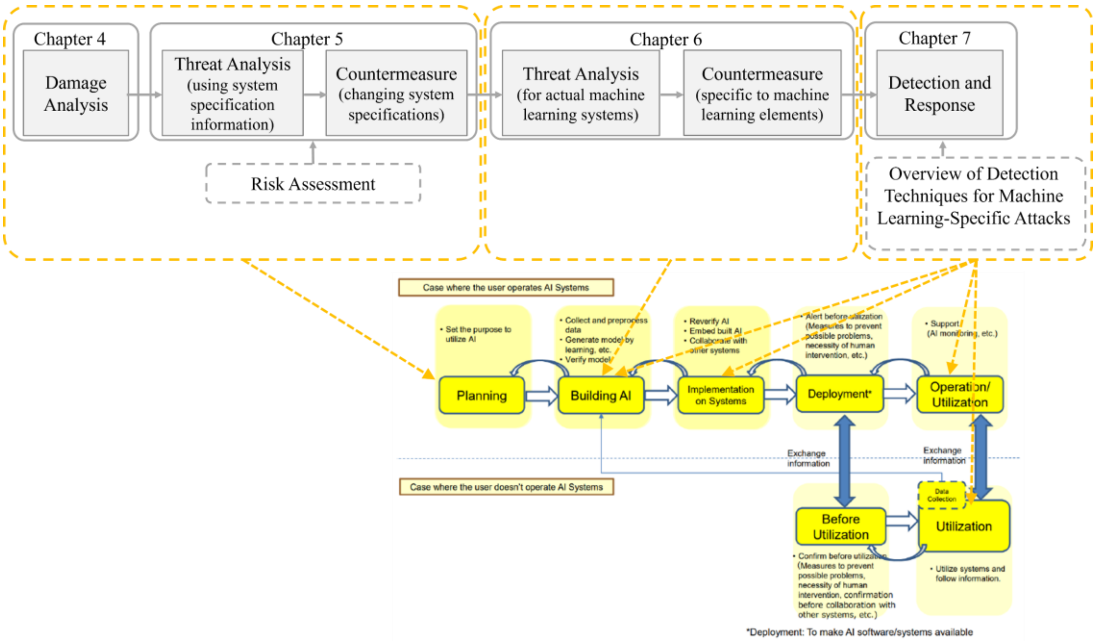

Figure I- 7. Machine learning security processes

First, “damage analysis” should be conducted, followed by “threat analysis and countermeasures using system specification information.” If threats are not sufficiently removed at this stage, “threat analysis and countermeasures for actual machine learning systems” should be conducted. Then, based on the results of the countermeasures, “detection and response” are designed and implemented as security measures for areas where the countermeasures are insufficient. Generally, “detection and response” are implemented during the operation/utilization of the system. However, in some cases, such as countermeasures against contamination of contaminated data and models, they are implemented at the stage of building AI and implementation on systems. In such cases, a system for “detection and response” must be established at the time of “building AI” in the “planning” stage.

In this document, the “Risk Assessment” part describes methods for system developers who do not have expertise in machine learning security to analyze threats in the “Threat Analysis and Countermeasures Using System Specification Information” process described above. In addition, “Overview of Detection Techniques for Machine Learning-Specific Attacks” introduces detection technology papers on MLSAs, which are categorized and organized with “target attack” and “its attack stage” in the above “detection and response” process.

I-3.3. Implementation of Machine Learning Security Processes

The implementation of each process described in the previous section requires the cooperation of two types of personnel, one responsible for actual analysis and countermeasures and the other responsible for deciding on countermeasures and detection in each process.

The person who performs the detection and countermeasures will basically be the system developer. However, in the case of “detection and response,” preliminary preparations such as log design are done in the “planning” and “implementation on systems” phases, whereas actual detection and response are performed in the “operation/utilization” phase. The former phase is often performed by the system developer and the latter phase by the service provider, so communication between these two is necessary. In addition, since expertise in machine learning security is essential implementing each phase, cooperation with experts is also necessary.

The responsible party is the person who takes responsibility for the risk when it materializes, such as the requester of the system development or the service provider. The results of the damage analysis and threat analysis presented by the implementer are used to determine the threats to be addressed, their priority, and constraints.

I-4. Damage Analysis

This chapter describes the damage analysis for MLSAs. The first step in damage analysis is to identify the assets to be protected related to the machine learning system. Then, the identified assets are linked to the threats posed by MLSAs, and the expected damage is calculated.

I-4.1. Identification of Assets That Should Be Protected

Assets related to the target machine learning system should be identified. Two major types of assets are related to machine learning systems: “assets that constitute machine learning” such as models and training data, and “assets related to machine learning systems” that are affected by the output results of the models.

Examples of “assets related to machine learning systems” for each system are listed below.

- Sign recognition systems ... Automatic driving
- X-ray diagnostic imaging systems ... Medical judgment
- Face recognition gates ... Security of gate installation location
- SNS image filters ... SNS policy
- Malware detection ...Information security of installed organizations and terminals

Assets associated with machine learning systems are also discussed in ENISA's Artificial Intelligence Cybersecurity Challenges [I-6]. Mainly “assets comprising machine learning” are classified and enumerated into six categories, such as data, models, and stakeholders. This document can be used as a reference to identify assets that should be protected.

I-4.2. Organization of Related Entities

Organize the stakeholders connected to these assets and identify these entities affected by the threats. For example, in the case of training data, the data provider is a stakeholder. Notably, “assets related to machine learning systems” may affect parties other than the system users. In the case of an X-ray image diagnosis system, the direct user of the system is the doctor, but the doctor and the patient to be diagnosed are affected by the judgment result of the system.

For reference, Figure I- 8 shows the entities involved in using AI according to the AI Utilization Guidelines [I-4].

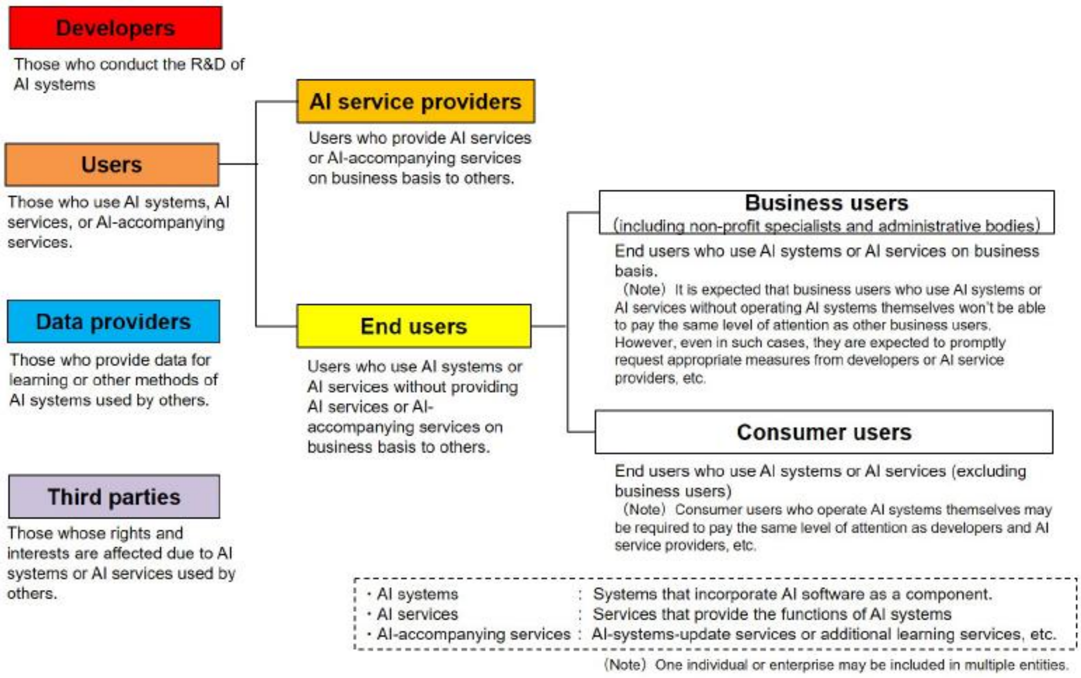

Figure I- 8. Classification of related entities in the use of AI

I-4.3. Damage Analysis of Threats Specific to Machine Learning Systems

The damage analysis identifies the possible damage by linking the assets and related entities identified in the work up to the previous section with the threats shown in Section エラー! 参照元が見つかりません。. An example of the linkages is shown in Figure I- 9.

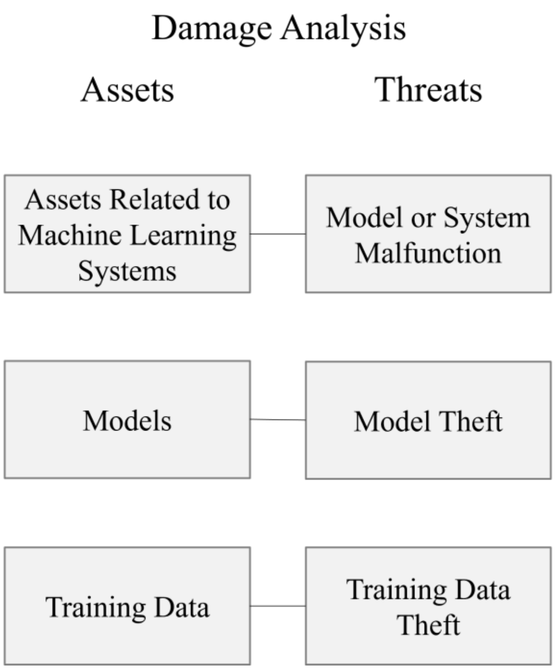

Figure I- 9. Examples of linking assets and threats

To link "assets related to machine learning systems" to "model or system malfunction," specific threat scenarios should be envisioned. Examples are listed below.

- Misjudging sign recognition system causes accidents in automated vehicles
- Misjudging X-ray image judgment systems causes medical errors
- Misjudging the facial recognition gate results allows non-registered persons to pass through the gate
- Avoiding social networking service's image filtering causes the uploading and publishing of images prohibited by terms and conditions
- Avoiding malware detection allows the target to download malware

AIST's Machine Learning Quality Management Guidelines [I-10] defines seven AI safety levels by dividing risk avoidance into human risks such as injury to humans and other economic risks as damage analysis. Please refer to that document as well.

I-5. Threat Analysis and Countermeasures Using System Specification Information

This chapter describes threat analysis and countermeasures from the specification information of machine learning systems. This analysis and these countermeasures are assumed to be conducted in the “planning” phase (Figure I- 3) in the flow of AI utilization in the Guidelines for AI Utilization [I-4] of the Ministry of Internal Affairs and Communications.

On the basis of the results of the damage analysis in Chapter I-4, threat analysis is conducted for attacks that cannot be overlooked based on the expected damage. By analyzing and taking countermeasures from specification information, the target attacks can be narrowed down and the analysis and countermeasures can be omitted at the time of “building AI” and the “implementation on systems” phase.

The specific implementation of this chapter is summarized in the “Risk Assessment” part as a method that enables system developers and service providers who do not have expertise in machine learning security.

I-5.1. Threat Analysis Using System Specification Information

The target machine learning system is analyzed to determine whether it is vulnerable to attacks that cause threats specific to machine learning systems (Section I-2.2) because of the system specifications. First, a potential attacker is established (Section I-5.1.1) from the entities (Section I-4.2) related to the target machine learning system. Next, the capabilities of the potential attacker are analyzed (Section I-5.1.2) to determine whether they satisfy the conditions for the target attack to be successful. Each process is described below.

I-5.1.1. Setting Potential Attackers

First, the potential attacker is set from the entities related to the target machine learning system (Section I-4.2). Potential attackers are the entities related to the target machine learning system that may become attackers. Each party has different privileges in the target machine learning system. For example, system operators can perform training processes, but system users can only perform inference processes. Therefore, multiple attackers are set up, and each attacker is analyzed to determine if it satisfies the attack success conditions described in Section I-5.1.2.

There are four potential attackers: “system administrators and operators (developers and service providers),” “training data providers,” “system users,” and “third parties that do not directly use the system.” For example, in the case of an X-ray diagnostic imaging system, the following attacker candidates are considered: “system administrator,” “provider of training X-ray image data,” “doctor (user of the system),” and “patient (a third party who does not directly use the system).” Figure I- 8 lists entities that are assumed to be involved using AI [I-4], which can be used as a reference when comprehensively identifying potential attackers.

I-5.1.2. Analysis of Conditions for Attack Formation

Next, the capability of the potential attacker is analyzed regarding whether attacks satisfy the conditions to be successful.

For the five attacks listed in Section I-2.2, the conditions for an attack to be successful are basically the number of inferences and the number of responses of inference results made by the machine learning system. In addition, for poisoning attacks, the training process and the degree of intervention in the training data and model are additional conditions. Factors that mitigate the conditions for the success of each attack include the content of the inference results, the ease of obtaining training data and similar data, and public information about the system.

The specific training and inference conditions for an attack to be successful must be calculated based on the latest research trends. Research that uses the same machine learning algorithm as the target machine learning system or that handles similar data can be used as a reference. Table I- 1 shows examples of factors used to determine the capabilities of a potential attacker.

Table I- 1. Examples of factors for determining potential attacker capability

Decision Factors	Examples
Degree of intervention in training	- Amount of arbitrary data that can be mixed into the training data - Number of arbitrary models that can be mixed into the model used for training - Feasibility of replacing trained models with arbitrary models by the attacker (For example, if the potential attacker can be entrusted with model training, the model can be replaced. In this case, poisoning attacks are possible regardless of the inference conditions.)
Number of processing inference and obtaining inference result	- Availability of trained models (e.g., in-vehicle systems can freely perform inference processing) - Content of inference results that can be obtained ✓ Output label ✓ Confidence score ✓ Internal output of the model during inference ✓ Inference results guessed from system behavior (Even if the inference result cannot be obtained directly, it may possibly by inferred from the system behavior, similar to a sign recognition system)
Availability of relevant data	- Part of training data - Data similar to training data - Statistical information of training data (The presence of such information facilitates model extraction attacks, model inversion attacks, and membership inference attacks.)
Publicly available information on machine learning systems	- Types of using machine learning algorithms - Specification of input data - Preprocessing of model input - Parameters for training

In the analysis, the likelihood of an attack is calculated not only by determining whether the potential attacker's ability satisfies the conditions for a successful attack but also by considering the feasibility of the attack, such as the time required to execute an inference to make the attack possible.

I-5.2. Countermeasures by Changing System Specification

Combining the damage analysis results in Chapter I-4 with the threat analysis results in Section I-5.1, countermeasures against the attacks that shall be prevent are introduced into the system specifications. Basically, a higher priority should be given to countermeasures against attacks that are expected to cause substantial damage or are easy to execute.

Two major types of countermeasures that changes system specifications are taken: system-wide and development process mitigation measures. Both countermeasures should be implemented so that the potential attacker cannot satisfy the attack conditions analyzed in Section I-5.1. Measures to mitigate damage by changing the system usage conditions are also taken. Table I- 2 shows examples of each type of mitigation measure.

Table I- 2. Mitigation measures by changing system specifications

Classification of mitigation measures	Examples
System-wide mitigation measures	- Reduce the number of times a potential attacker can touch the system or model ✓ Limit the number of data inputs - Stop outputting information that is not requested ✓ Limit output, e.g., output only the top decision label ✓ Output only decision labels, not confidence scores - Reduce the amount of public information about the system/model - Limit the number of people who can input training data
Mitigation measures in the development process	- Use data and models from trusted providers for training
Damage mitigation	- Change the system design to allow for risk ✓ Add human judgment between the output and the next process in a system that automatically executes the next process from model output

If countermeasures cannot be taken by changing system specifications, threat analysis and countermeasures against the actual machine learning systems described in the next chapter should be performed.

I-6. Threat Analysis and Countermeasures for Actual Machine Learning Systems

This chapter describes threat analysis and countermeasures for actual machine learning systems. They are assumed to be performed in the “building AI” phase of the flow of AI utilization in the AI Utilization Guidelines of the Ministry of Internal Affairs and Communications [I-4].

The targets for implementation are attacks (Section I-2.2) for which sufficient countermeasures could not be taken in the threat analysis and countermeasures using the system specification information described in Chapter I-5. This countermeasure evaluates the feasibility of such attacks and implements countermeasures specific to machine learning elements.

I-6.1. Threat Analysis for Real Models

Threat analysis for real models is performed by analyzing whether attacks that pose threats specific to machine learning systems (Section I-2.2) are actually feasible for the target machine learning system.

First, specific attack methods that can be applied are extracted from published papers based on the machine learning algorithm of the target machine learning system and the type of handling data. Next, the extracted attack methods are implemented and the success rate of the attacks is evaluated. Several tools have been published to assist in implementing attacks, including the Adversarial Robustness Toolbox (ART) [I-12], CleverHans [I-13], and Counterfit [I-14].

I-6.2. Countermeasures Specific to Machine Learning Elements

On the basis of the results of the previous section, mitigation measures specific to machine learning elements are implemented for attacks that require countermeasures. Examples of mitigation measures for each type of attack are shown in Table I- 3. After implementing mitigation measures, the threat analysis described in Section I-6.1 is reperformed to evaluate and confirm the effectiveness of the countermeasures. The effectiveness of the countermeasures will be evaluated and verified.

Table I- 3. Examples of mitigation measures specific to machine learning elements

Attacks	Mitigation Examples
Evasion attack	- Adversarial Training Method to create models that are difficult to make adversarial examples for by training developer-created adversarial examples with correct labels in advance - Certified Robustness Techniques that determine the size of the perturbation are to be guaranteed in advance and ensure that no adversarial examples exist within that range. - Smoothing Techniques to make it harder for adversarial examples to exist by smoothing decision boundaries
Poisoning attack	- Robustness Methods to make it harder to apply poisoning attacks by devising training methods such as removing tainted datasets
Model extraction attack	- Differential Privacy Techniques that make it difficult to attack individual data by disrupting the algorithm that produces the output
Model inversion attack	- Differential Privacy Techniques that make it difficult to attack individual data by disrupting the algorithm that produces the output
Membership inference attacks	- Differential Privacy Techniques that make it difficult to attack individual data by disrupting the algorithm that produces the output

I-7. Detection and Response

This chapter describes how to detect and respond with attacks in the operation of machine learning systems. It is assumed that these technologies are drafted in “planning” and “implementation on systems” and implemented in “building AI” and “operation/utilization” in the flow of AI utilization in the Guidelines for AI Utilization [I-4] issued by the Ministry of Internal Affairs and Communications. These technologies are countermeasures against threats that could not be adequately addressed by the security measures described in Chapters I-5 and I-6.

In addition, technical papers on detection techniques for MLSAs are surveyed and categorized. These papers are organized by "attack events to be detected" and "data to be monitored." The results are summarized in "Overview of Detection Techniques for Machine Learning-Specific Attacks."

I-7.1. Detection and Response in Machine Learning System Security

Detection and response to attacks in machine learning systems are performed in the same way as in general cyber-attack countermeasures [I-15]. In the "detection" phase, the sign of the precursor and the indicator of an attack are monitored. In the "response" phase, containment, eradication, and restoration are performed after an attack or damage is detected.

I-7.2. Detection

Detection includes "precursor detection" and "indicator detection." "Precursor detection" is the detection of event signs that signal future attacks. "Indicator detection" is the detection of events that indicate attacks have already occurred or are currently occurring. In both cases, the attacks and events must be determined and the necessary logs must be recorded for detection.

MITRE ATT&CK [I-16], which analyzes and models attacker behavior, is a framework used for detection and response in general cyber-attack countermeasures. For countermeasures against attacks on machine learning systems, ATLAS [I-5], also published by MITRE, provides a list of tactics and methods for MLSAs. It summarizes the methods used in each stage of attacks, from reconnaissance and initial access to attack execution and information theft, and can be used as a reference for selecting the events to be detected.

Table I- 4 shows examples of possible attack events that could be detected on the machine learning system side among the five attacks listed in Section I-2.2.

Table I- 4. Examples of attack events

<table>
<tr><th>Attacks</th><th colspan="2">Precursor Detection</th><th>Indicator Detection</th></tr>
<tr><td>Evasion attack</td><td rowspan="5">- Target system survey
✓ Number of classification labels
✓ Query limit
- Creation of attack accounts</td><td>- Activities to create adversarial examples</td><td>- Input of adversarial examples</td></tr>
<tr><td>Poisoning attack</td><td>- Inclusion of contaminated data in the training data
- Inclusion of tainted models</td><td>- Input containing triggers (backdoor attack)</td></tr>
<tr><td>Model extraction attack</td><td>-</td><td>- Input attack query</td></tr>
<tr><td>Model inversion attack</td><td>-</td><td>- Input attack query</td></tr>
<tr><td>Membership inference attack</td><td>-</td><td>- Input attack query</td></tr>
</table>

Common to all attacks, the target of precursor detection includes reconnaissance activities to investigate the target system and preparation for the attack. Investigation of the target system includes surveying the number of classification labels and query limits (size and number) to be used as a reference for attack activities. The preparation includes creating an account to conduct the attack. In terms of the precursor detection targets for each attack, evasion attacks include activities to create adversarial examples, and poisoning attacks include activities to introduce contaminated data or models into the system. Poisoning attacks are also expected to occur during the development process (building AI). Therefore, the detection of poisoning attacks requires precursor detection in the development process as well.

In indicator detection, the detection targets are activities that cause system malfunction and the theft of models or training data. Activities that can cause model or system malfunctions include entering adversarial samples for evasion attacks. Poisoning attacks include input triggers for backdoor attacks. For model and training data theft, model extraction attacks, model inversion attacks, and membership inference attacks involve the input of attack queries to thieve information.

As reference information for designing specific detection methods and logs, “Overview of Detection Techniques for Machine Learning-Specific Attacks” is summarized. In this section, the survey results of technical papers on the detection of MLSAs in terms of “attack events to be detected” and “data to be monitored” are organized.

I-7.3. Response

This section describes the concepts in coping after an attack or damage has occurred, and examples of how to deal with the five attacks described in Section I-2.2. Similar to general cyber-attacks, it is difficult to prevent all attacks specific to machine learning systems. Therefore, functions that enable emergency countermeasures and logs that enable an investigation with reference to Section I-2.2 should be designed.

As with cyber-attack countermeasures, actions should be taken in the order of “emergency countermeasures,” “investigation,” and “permanent countermeasures” when the machine learning system is attacked or damaged. “Emergency countermeasures” are taken to mitigate or halt attacks that are occurring temporarily. The next step, “investigation,” investigates the current attack and checks for similar attacks in the past to recover the system and prevent recurring attacks. The final step takes “permanent countermeasures” to prevent recurring attacks based on the information obtained from the investigation, and the system is restored.

The process is described below.

I-7.3.1. Emergency Countermeasures

To prevent the spread of damage, emergency countermeasures including system restrictions or shutdown are taken (Table I- 5). Emergency countermeasures fall into three major categories. The first is “system restriction/shutdown,” which can respond to any type of attack. The second is “mitigation through pre-processing,” which deals with attacks whose objective is “model or system malfunction.” The third is “output spoofing” for attacks aimed at “model theft,” “training data theft,” or adversarial example generation activities in evasion attacks. However, “output spoofing” requires clarifying the relationship between system transparency and accountability.

Table I- 5. Examples of emergency countermeasures

Classifications	Examples
System restriction/shutdown	- System shutdown (common to all attacks) - Suspension of attacker accounts (common to all attacks) - Limit the number of inputs (adversarial examples generation for evasion attacks, model extraction attacks, model inversion attacks, membership inference attacks) - Rollback to past models (poisoning attacks) ✓ Rollback to past models that could correctly model (before pollution) - Restrict the output of labels that were the attack target (evasion attacks and poisoning attacks)
Mitigation through pre-processing	- Mitigation of perturbation by noise removal, etc. (evasion attacks) - Trigger removal (backdoor attack for poisoning attacks)
Output spoofing	- Spoofing of output labels (adversarial examples generation for evasion attacks, model extraction attacks, model inversion attacks, membership inference attacks) - Spoofing of confidence score (adversarial examples generation for evasion attacks, model extraction attacks, model inversion attacks, membership inference attacks)

I-7.3.2. Investigation

To recover the system and prevent recurring attacks, investigations of previous attacks should be conducted. Basically, three types of the investigations are conducted: the investigation of the purpose, methods, and damage of previous attacks; the investigation of similar previously conducted attacks; and the investigation of combined attacks.

Table I- 6. Examples of attack investigation

Attacks	Investigation of Attacks	Investigation of similar attacks	Combination survey
Evasion attack	- Identification of target labels for attacks - Identification of methods for creating adversarial examples	- Search for previous evasion attacks	- Feasibility investigation of evasion attacks through model extraction attacks - Feasibility Investigation of evasion attacks through poisoning attacks
Poisoning attack	- Identification of target labels for attacks - Trigger identification - Identification of contaminated data/contaminated model - Identification of contamination route - Identification of the method used to create contaminated data	- Search for past misclassification caused by poisoning attacks	- Feasibility investigation of model extraction attacks through poisoning attacks
Model extraction attack	- Identification of the amount of thieved information about the mode (model building from input/output data) - Identification of attack methods	- Search for past attempts of attacks	-
Model inversion attack	- Identification of leaked training data - Identification of attack methods	- Search for past attempts of attacks	- Feasibility investigation of model inversion attacks through model extraction attacks

Membership inference attack	- Identification of leaked information - Identification of attack methods	- Search for past attempts of attacks	- Feasibility investigation of membership inference attacks through model extraction attacks

I-7.3.3. Permanent Countermeasures

Permanent countermeasures involve applying defensive and mitigation measures against attacks and introducing measures to detect attacks. Information obtained from surveys will enable the effective application of defensive and mitigation measures.

- Evasion Attack
  Adversarial training with adversarial examples used in the attack
- Poisoning Attack
  Retraining after removing contaminated data and models

I-8. References

[I-1] Organization for Economic Co-operation and Development (OECD), Principles on Artificial Intelligence.
https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

[I-2] The Integrated Innovation Strategy Promotion Council, Social Principles of Human-Centric AI.
https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf

[I-3] Minister of Economy, Trade and Industry (METI), AI Governance in Japan Ver. 1.1.
https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20210709_8.pdf

[I-4] Ministry of Internal Affairs and Communications (MIC), The Conference toward AI Network Society, AI Utilization Guidelines.
https://www.soumu.go.jp/main_content/000658284.pdf

[I-5] MITRE, ATLAS. https://atlas.mitre.org

[I-6] European Network and Information Security Agency (ENISA), Artificial Intelligence Cybersecurity Challenges.
https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges

[I-7] Microsoft Corporation, Threat Modeling AI/ML Systems and Dependencies.
https://docs.microsoft.com/en-us/security/engineering/threat-modeling-aiml

[I-8] Information Commissioner’s Office (ICO), AI and data protection risk mitigation and management toolkit.
https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ai-and-data-protection-risk-mitigation-and-management-toolkit/

[I-9] National Institute of Standards and Technology (NIST), Draft NIST IR8269: A Taxonomy and Terminology of Adversarial Machine Learning.
https://csrc.nist.gov/publications/detail/nistir/8269/draft

[I-10] National Institute of Advanced Industrial Science and Technology (AIST), Machine Learning Quality Management Guideline, 2nd English Edition.
https://www.digiarc.aist.go.jp/en/publication/aiqm/guideline-rev2.html

[I-11] Ministry of Internal Affairs and Communications (MIC) x Mitsui Bussan Secure Directions, Inc., AI Security Matrix.
https://www.mbsd.jp/aisec_portal/index.html

[I-12] International Business Machines Corporation, Adversarial Robustness Toolbox.
https://github.com/Trusted-AI/adversarial-robustness-toolbox

[I-13] CleverHans Lab, CleverHans.
https://github.com/cleverhans-lab/cleverhans

[I-14] Microsoft Corporation, Counterfit.
https://github.com/Azure/counterfit/

[I-15] National Institute of Standards and Technology (NIST), NIST SP800-61 Rev. 2: Computer Security Incident Handling Guide.
https://csrc.nist.gov/publications/detail/sp/800-61/rev-2/final

[I-16] MITRE, MITRE ATT&CK.
https://attack.mitre.org

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

MLSE 機械学習工学研究会
MACHINE LEARNING SYSTEMS ENGINEERING

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.

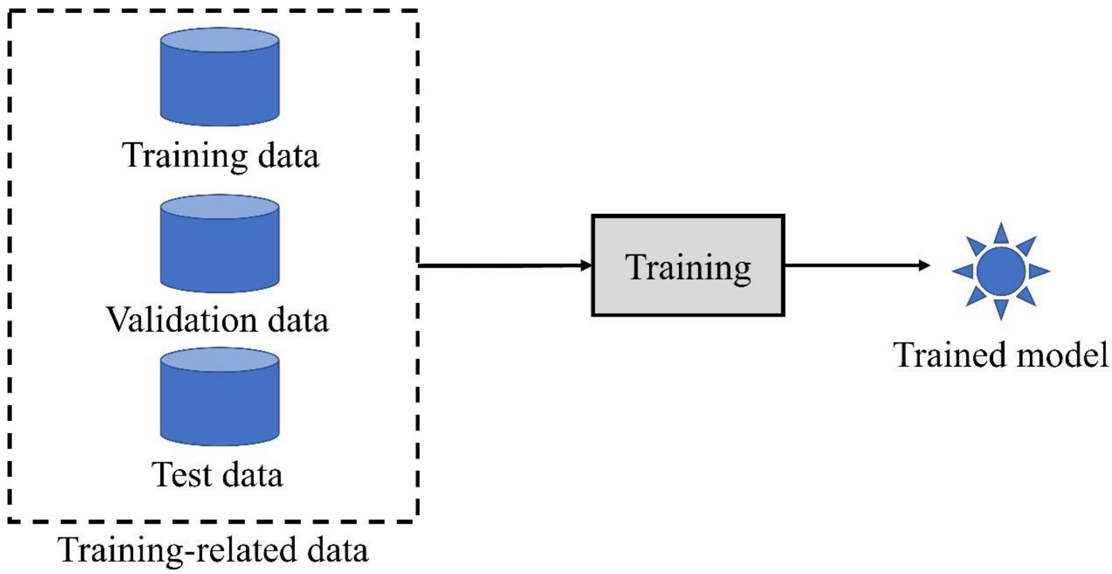

Figure II- 1. Training pipeline in a machine learning system

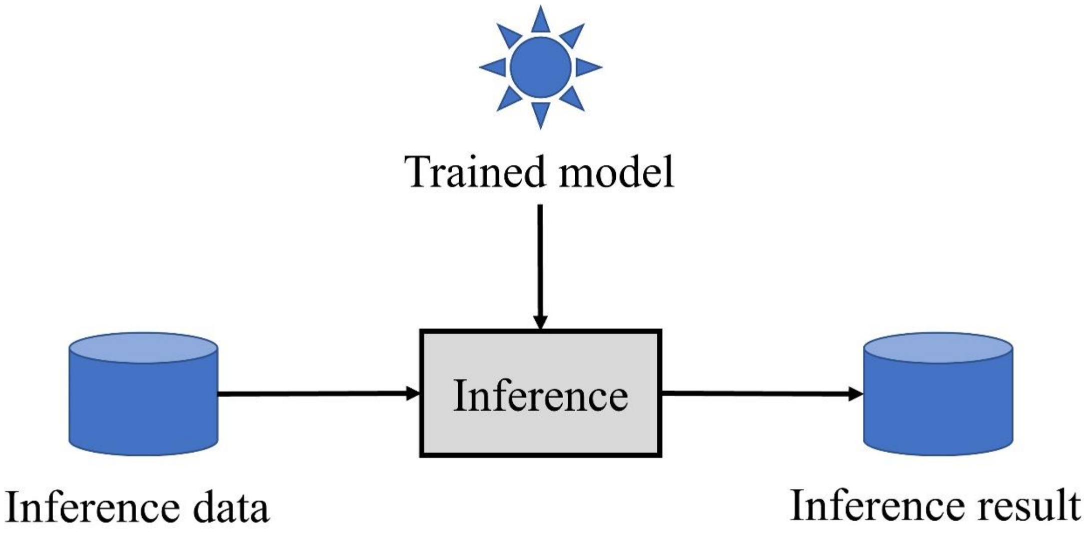

Figure II- 2. Inference pipeline in a machine learning system

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.

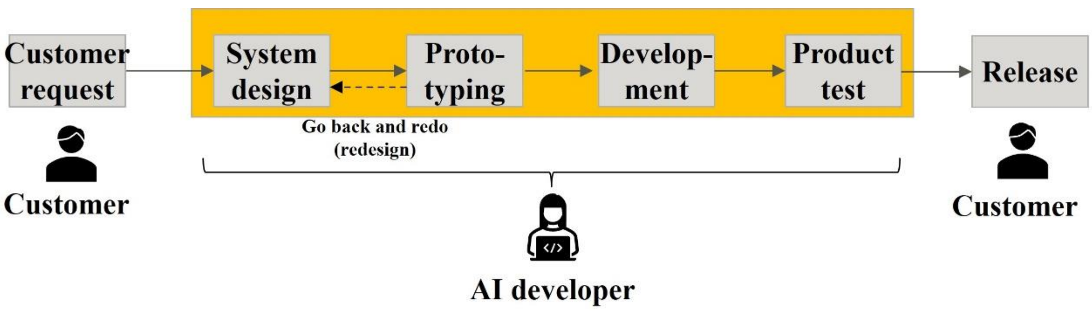

Figure II- 3. General development process of machine learning processing in the machine learning system

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 <u>threat analysis</u> 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.

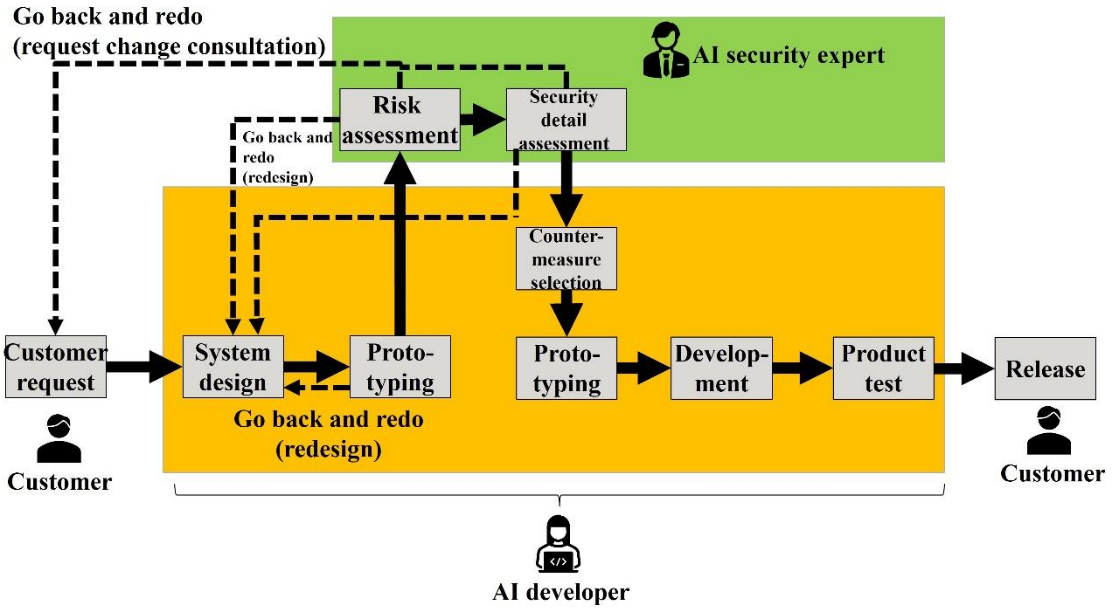

Figure II- 4. Development process of machine learning processing part considered for security against machine learning system-specific attacks

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.

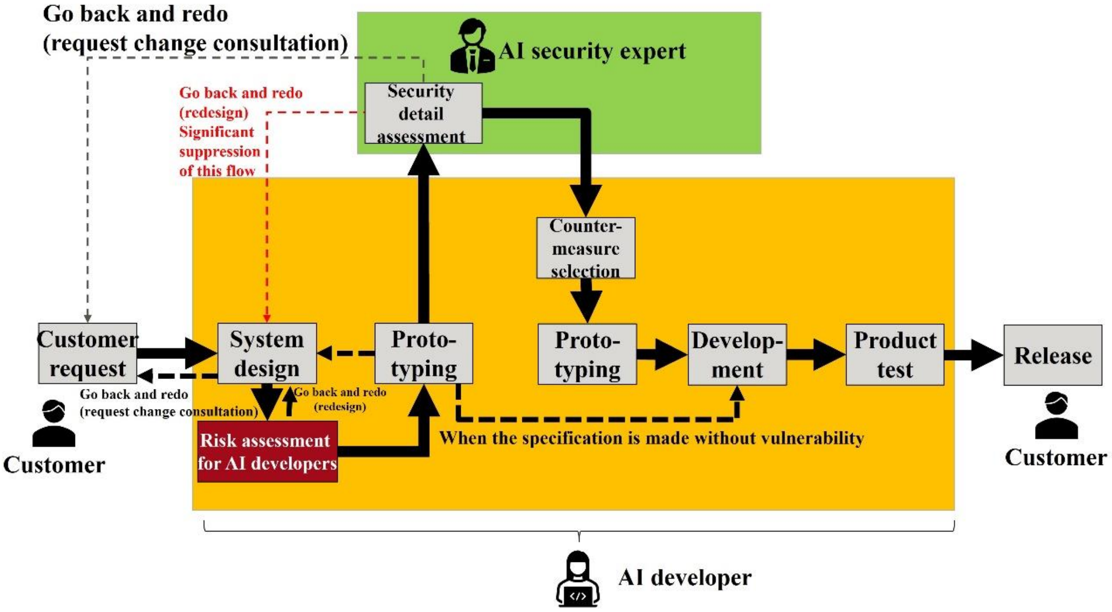

Figure II- 5. Development process of the machine learning processing unit that reduces rework due to security measures (desired model)

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”.

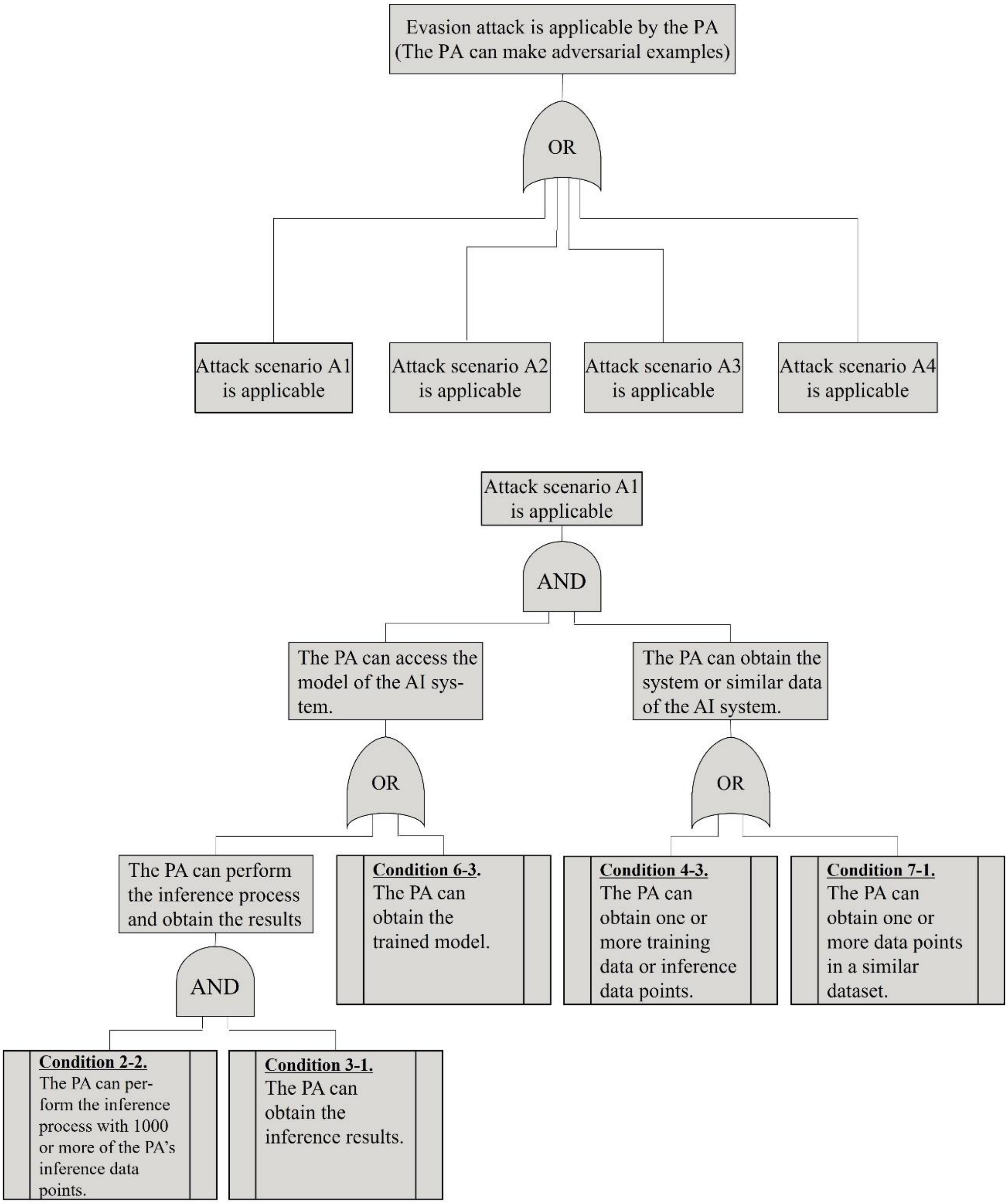

Figure II- 6. Examples of (a part of) constructed attack trees

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.

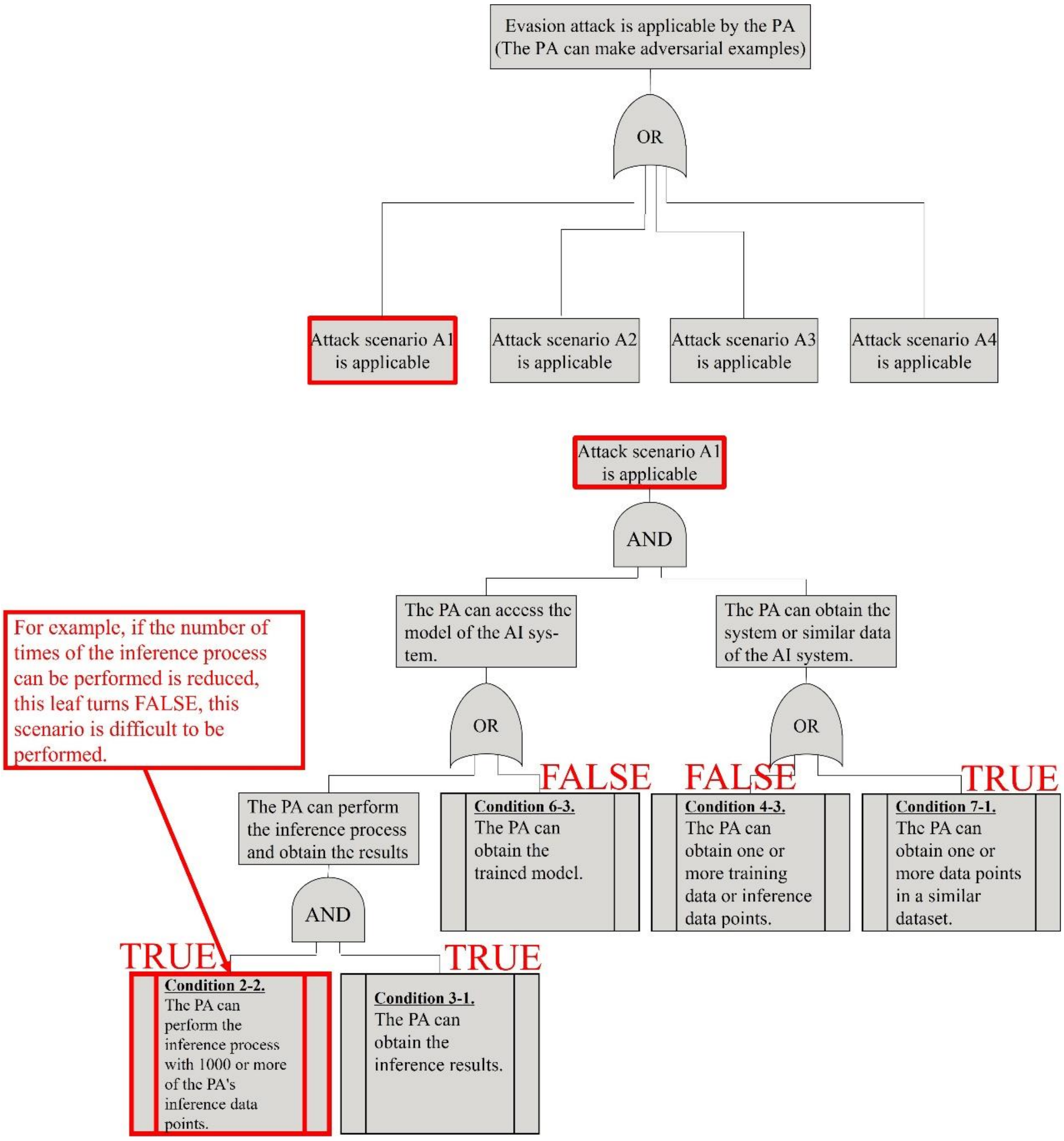

Figure II- 7. Example of the consideration of countermeasures

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.

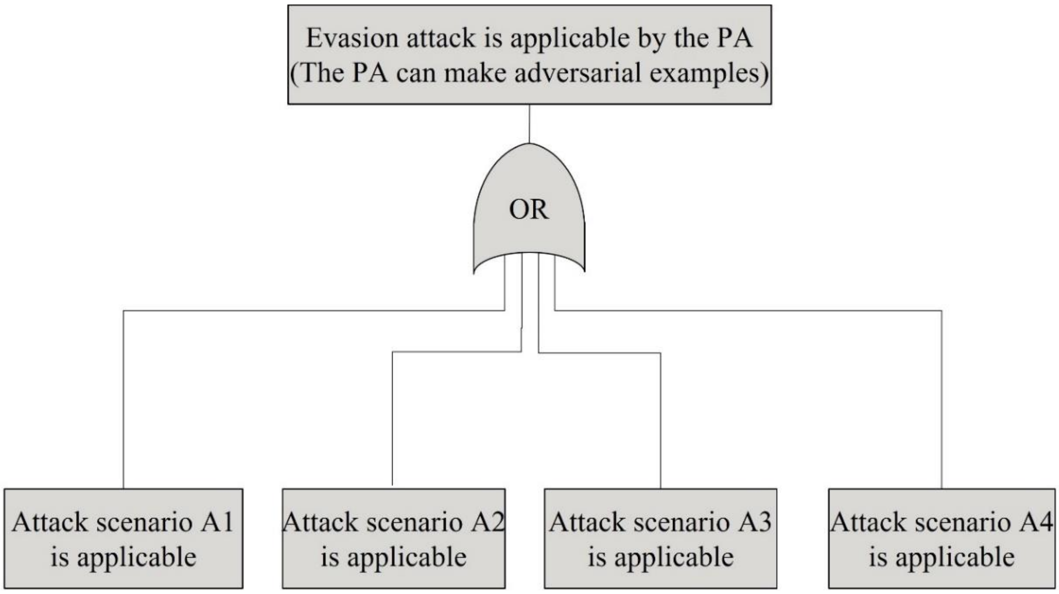

Figure II- 8. Example of the attack tree for evasion attacks (adversarial examples) (upper part)

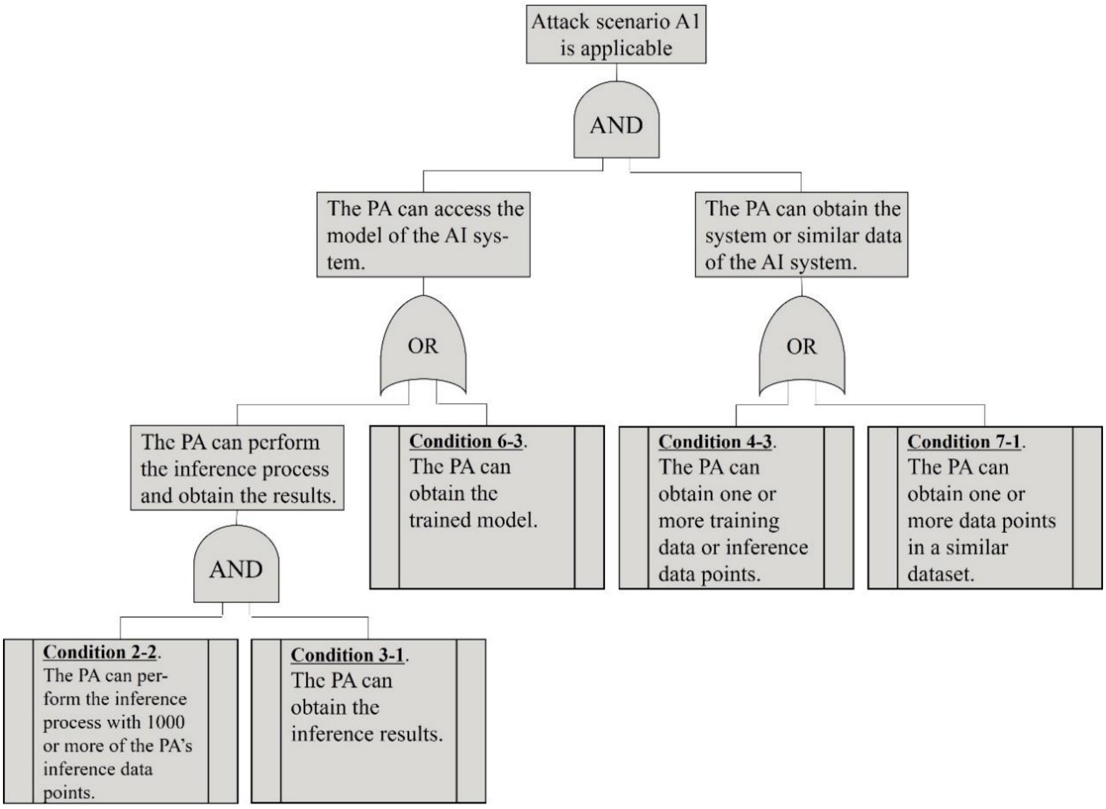

Figure II- 9. Attack tree and attack executable conditions for attack scenario A1 of evasion attacks (adversarial examples)

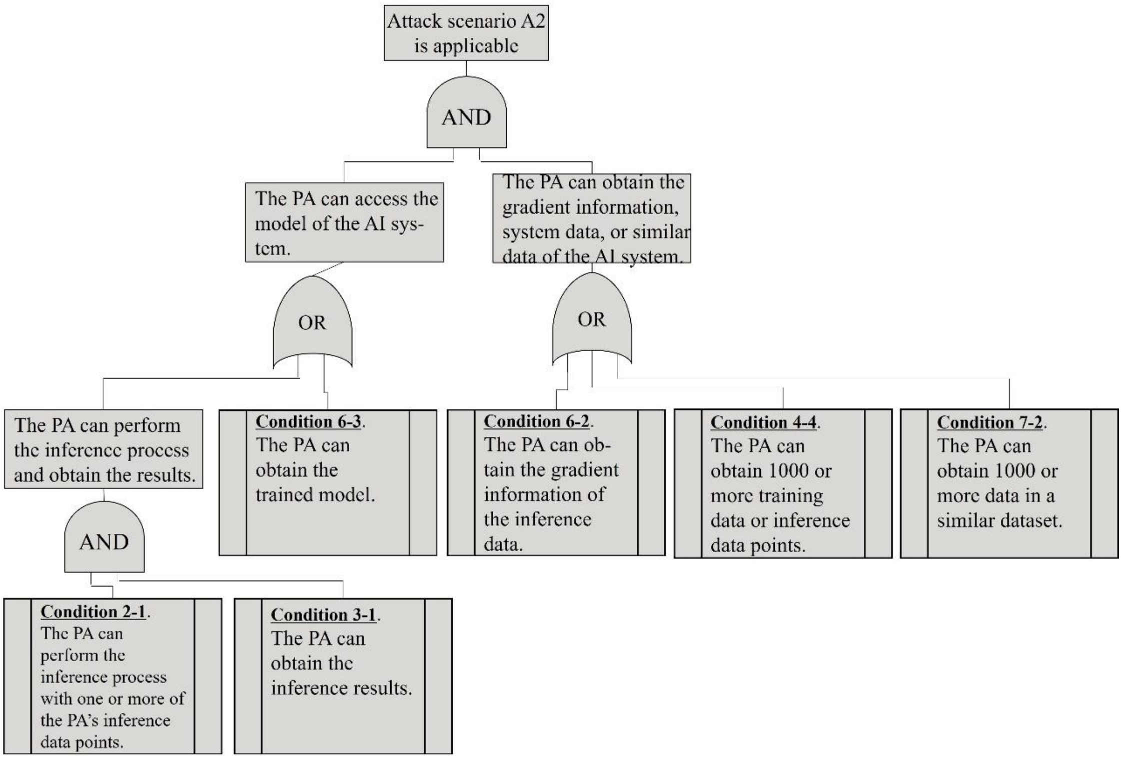

Figure II- 10. Attack tree and attack executable conditions for attack scenario A2 of evasion attacks (adversarial examples)

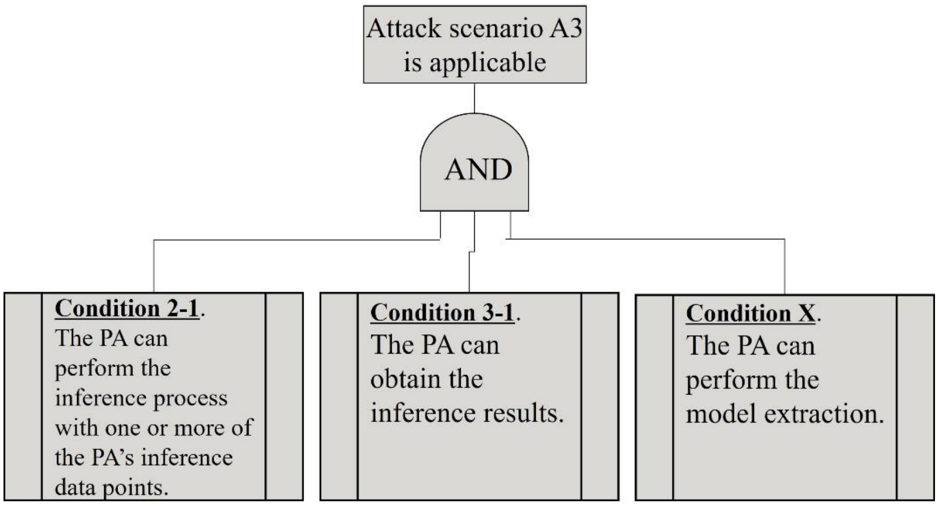

Figure II- 11. Attack tree and attack executable conditions for attack scenario A3 of evasion attacks (adversarial examples)

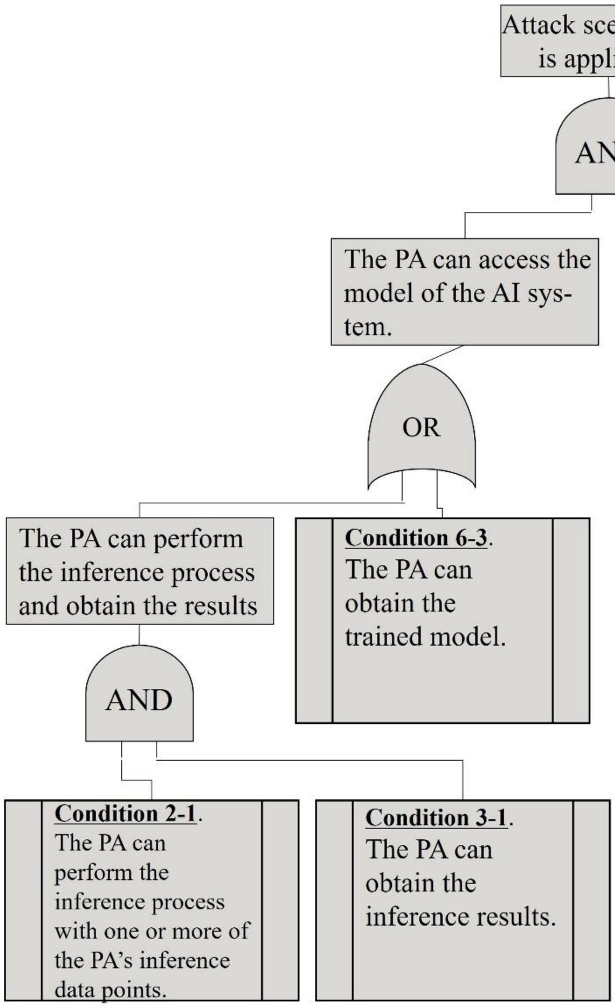

Figure II- 12. Attack tree and attack executable conditions for attack scenario A4 of evasion attacks (adversarial examples)

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.

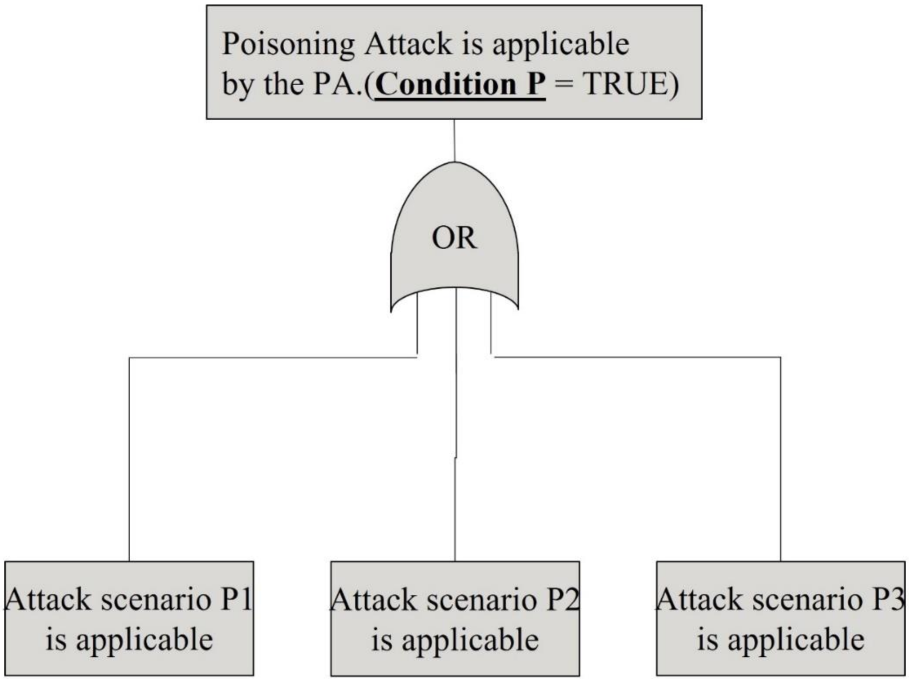

Figure II- 13. Example of the attack tree for poisoning attacks (upper part)

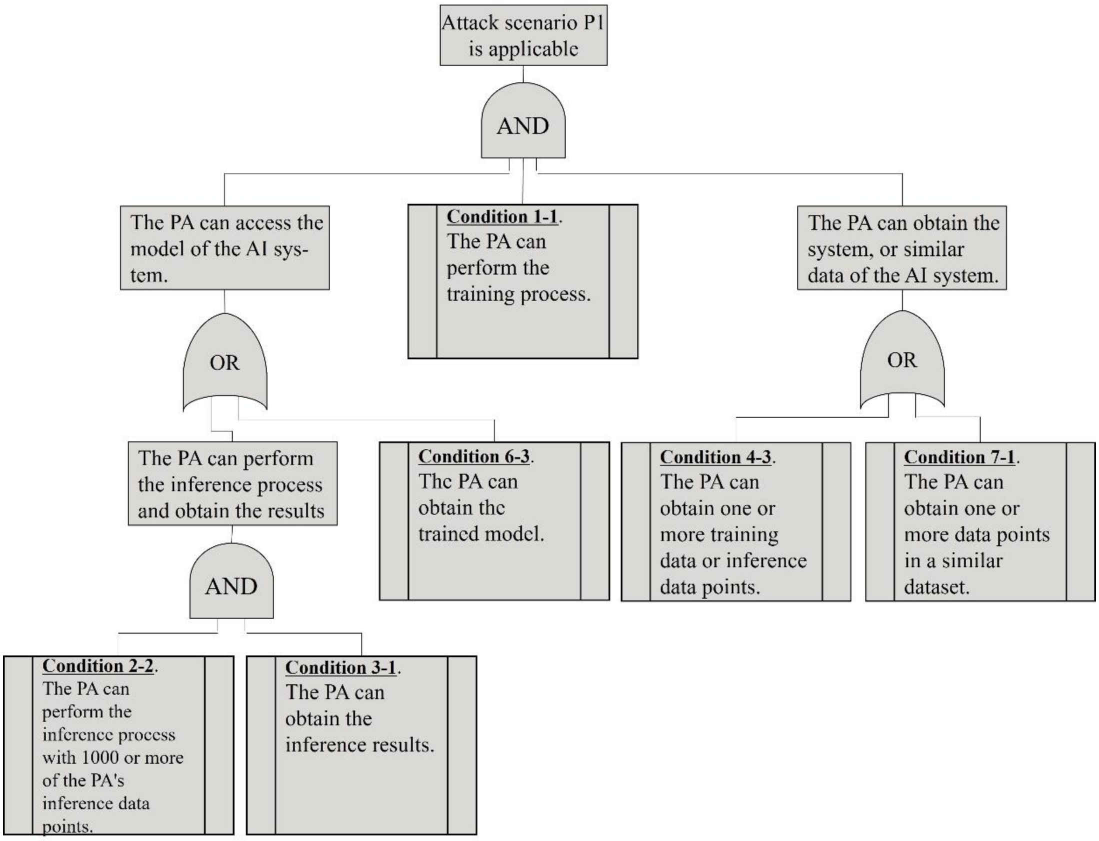

Figure II- 14. Attack tree and attack executable conditions for attack scenario P1 of poisoning attacks

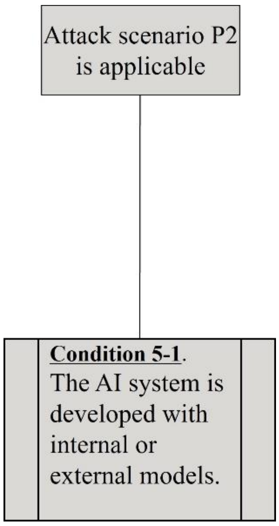

Figure II- 15. Attack tree and attack executable conditions for attack scenario P2 of poisoning attacks

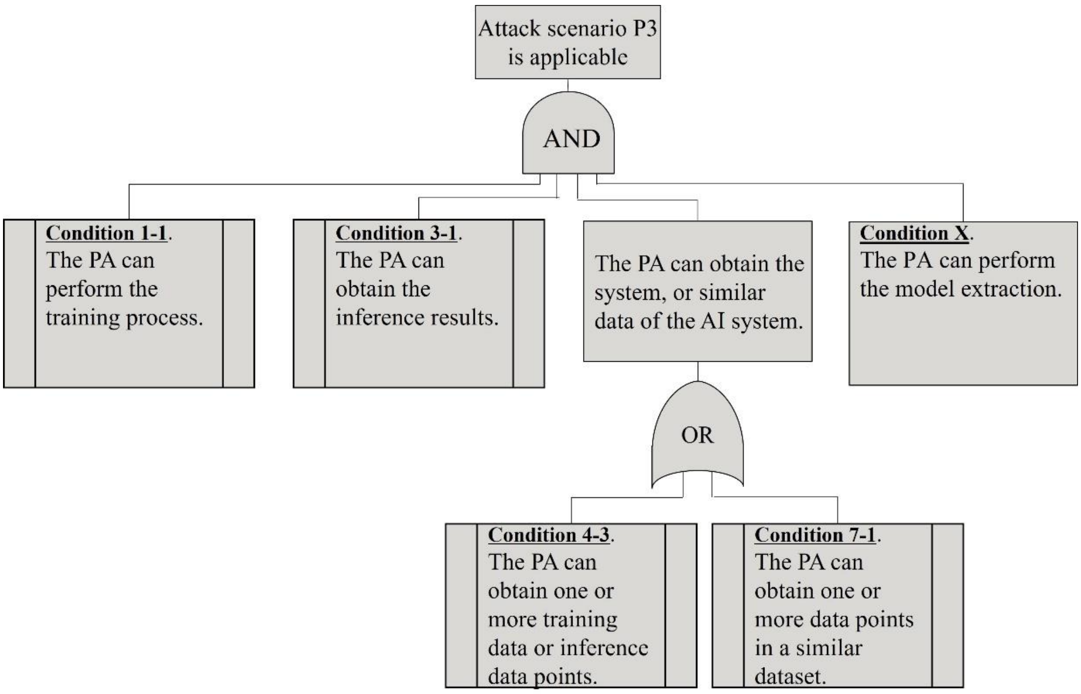

Figure II- 16. Attack tree and attack executable conditions for attack scenario P3 of poisoning attacks

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.

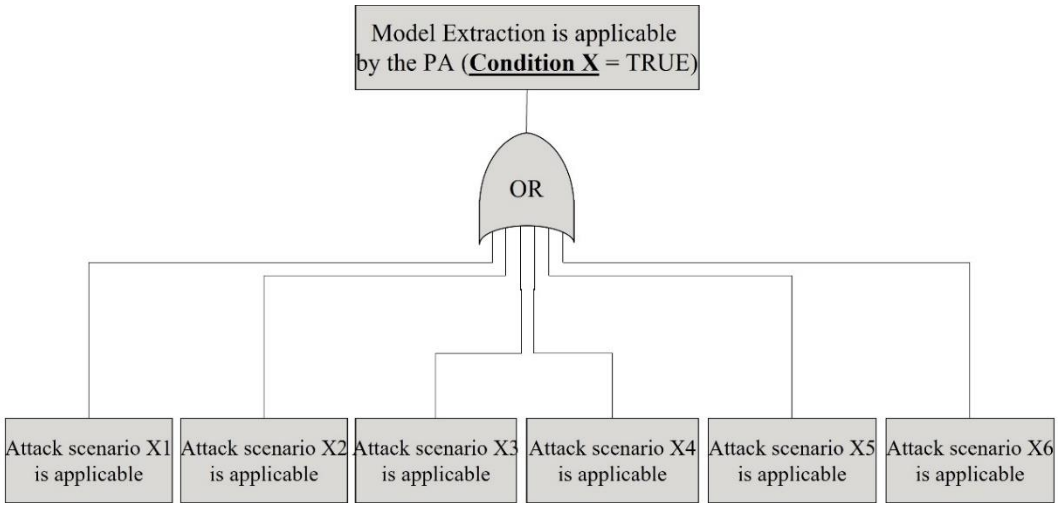

Figure II- 17. Example of the attack tree for model extraction (upper part)

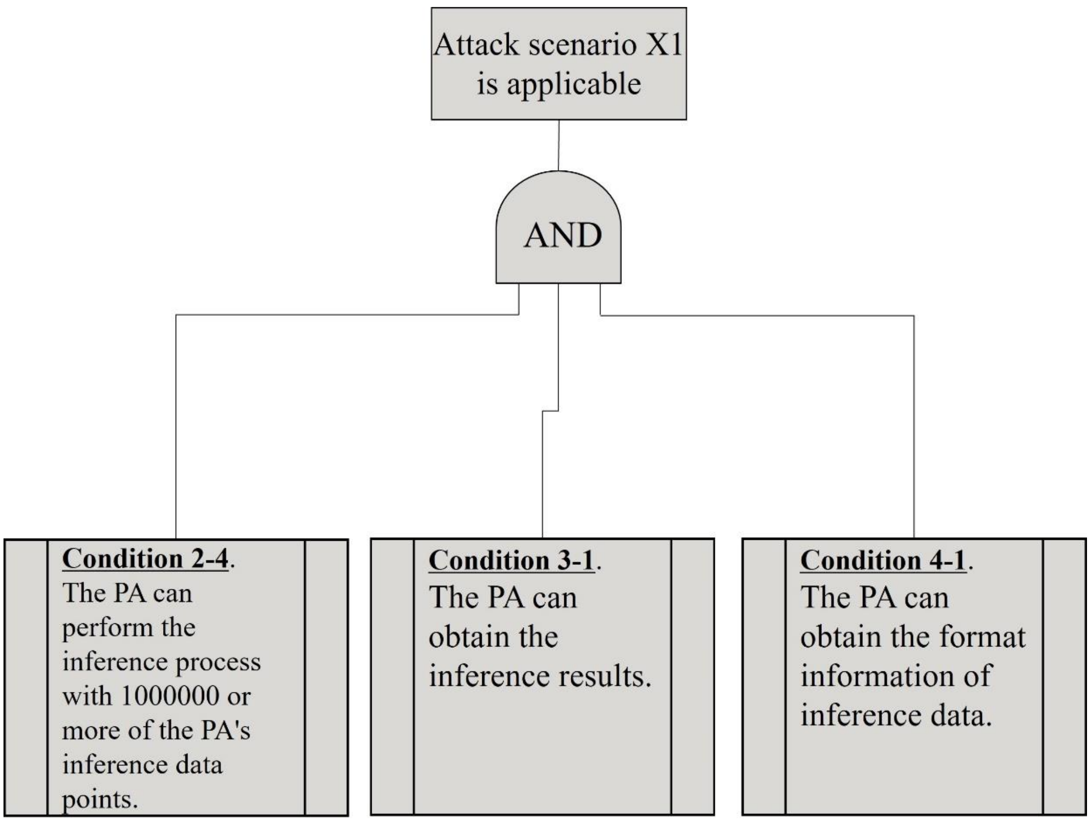

Figure II- 18. Attack tree and attack executable conditions for attack scenario X1 of model extraction

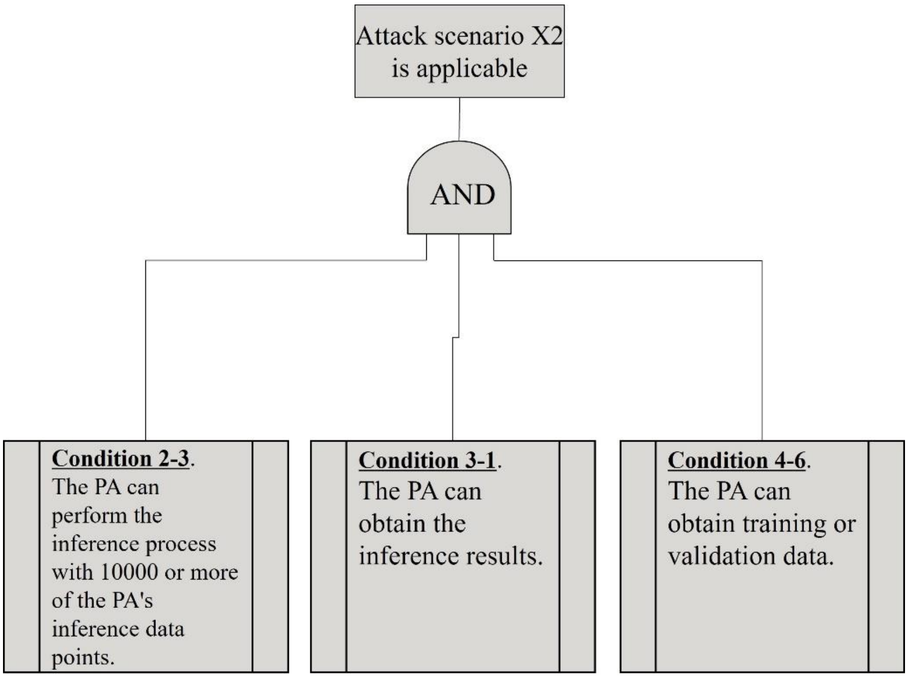

Figure II- 19. Attack tree and attack executable conditions for attack scenario X2 of model extraction

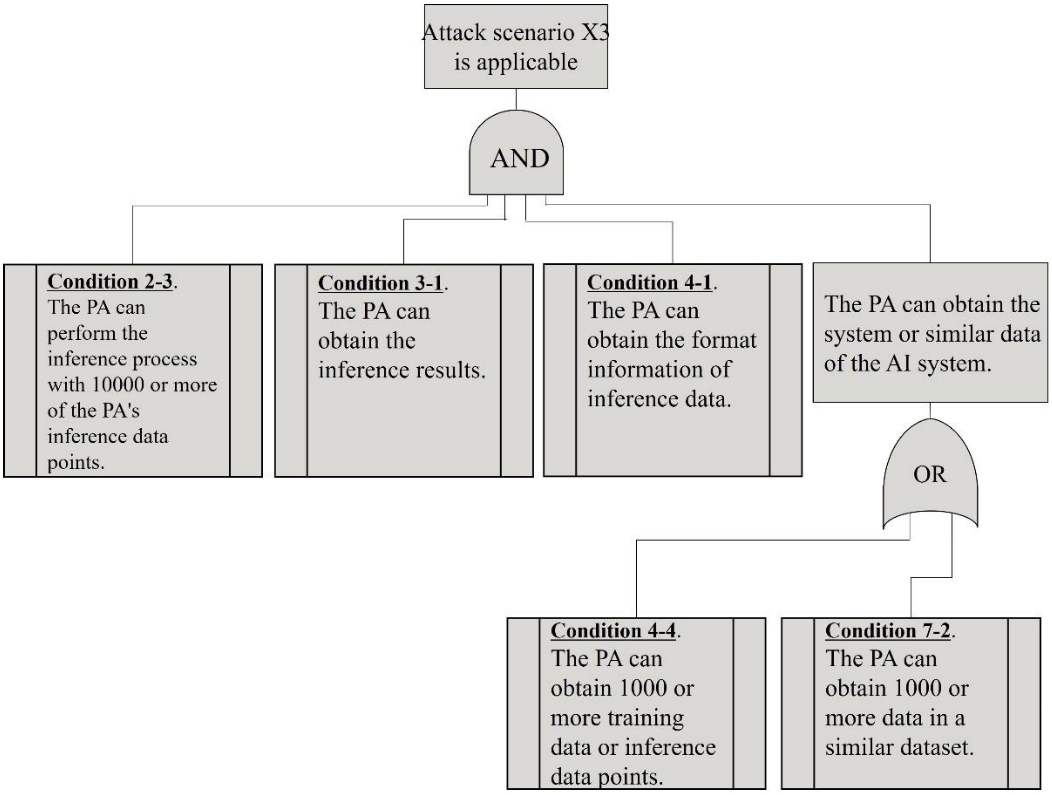

Figure II- 20. Attack tree and attack executable conditions for attack scenario X3 of model extraction

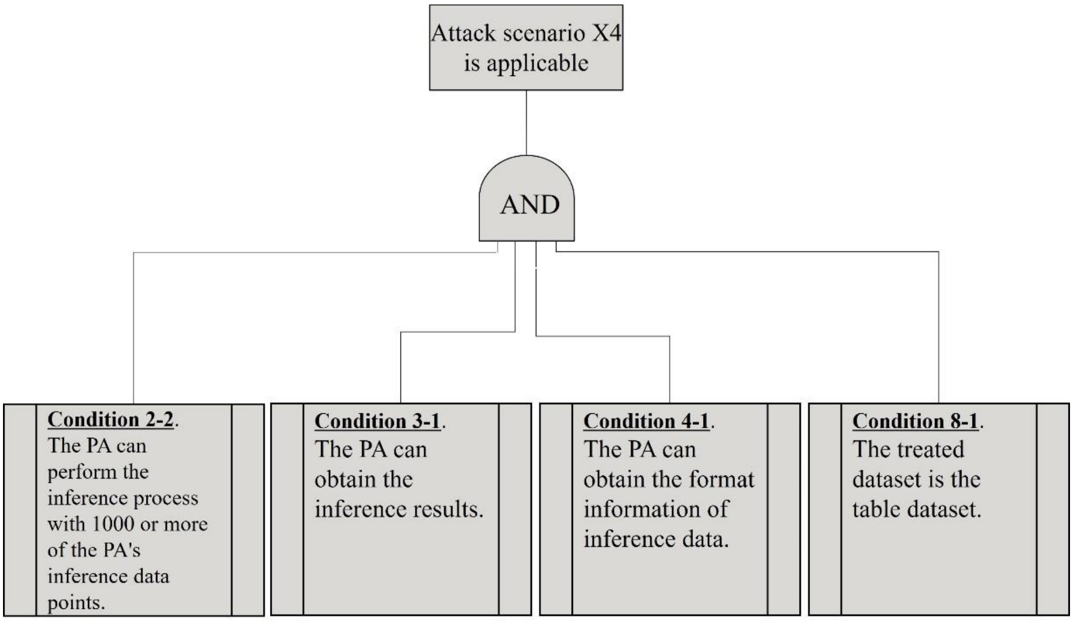

Figure II- 21. Attack tree and attack executable conditions for attack scenario X4 of model extraction

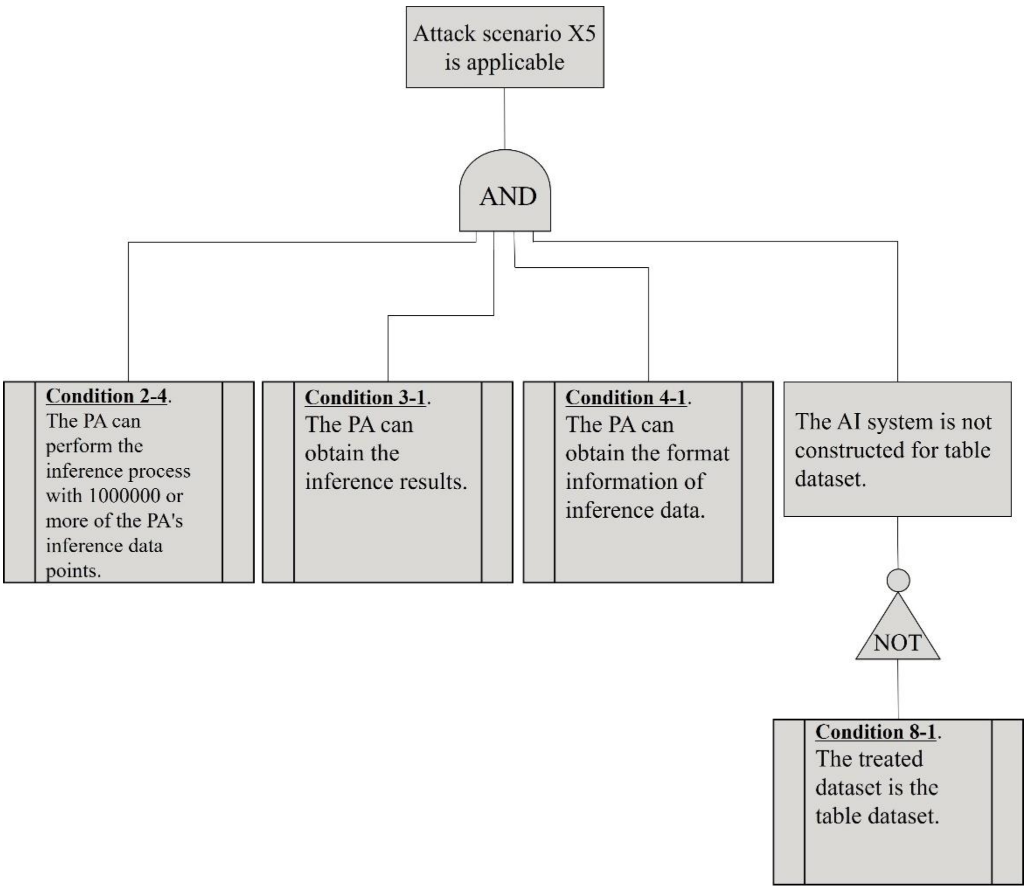

Figure II- 22. Attack tree and attack executable conditions for attack scenario X5 of model extraction

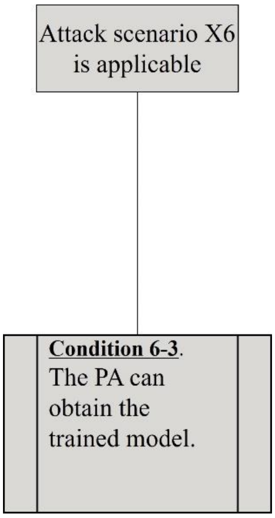

Figure II- 23. Attack tree and attack executable conditions for attack scenario X6 of model extraction

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.

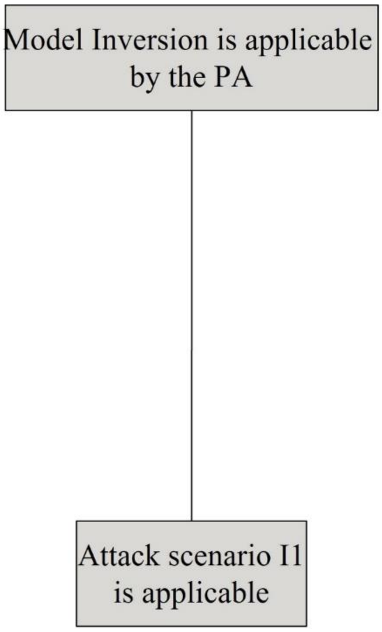

Figure II- 24. Example of the attack tree for model inversion (upper part)

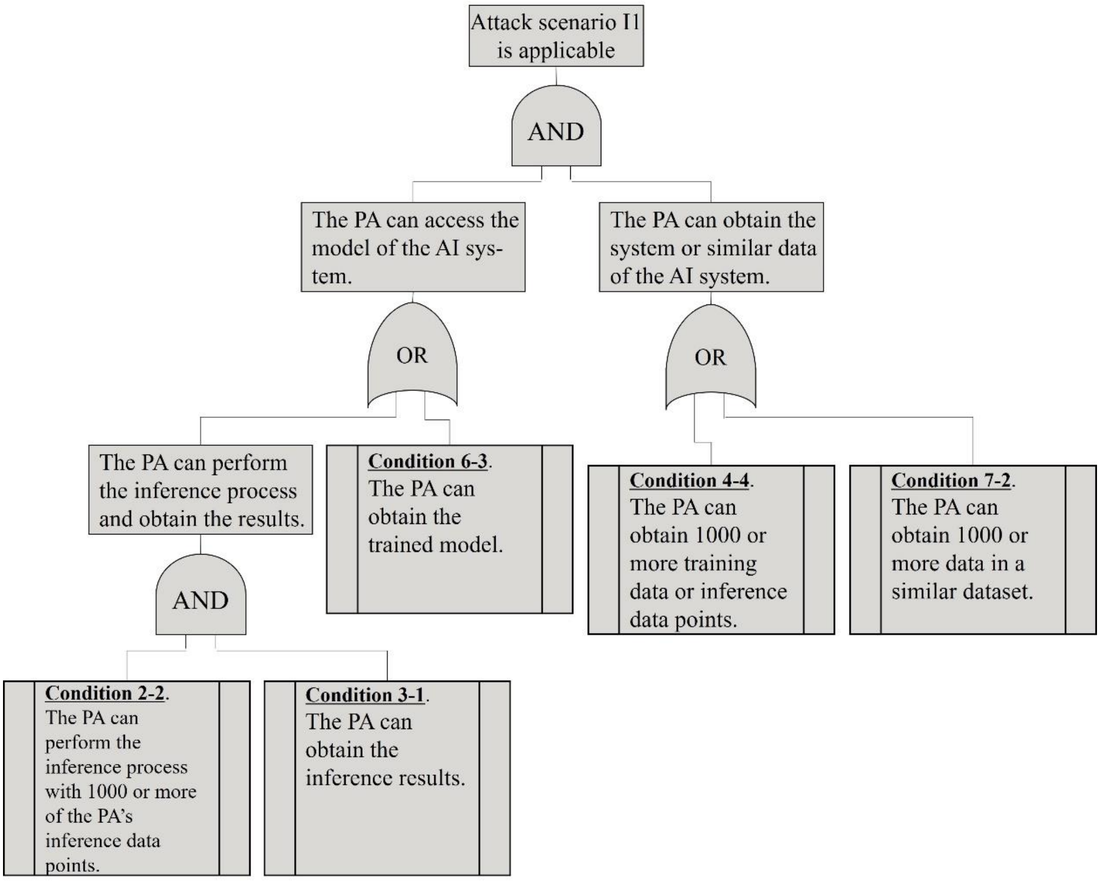

Figure II- 25. Attack tree and attack executable conditions for attack scenario I1 of model inversion

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.

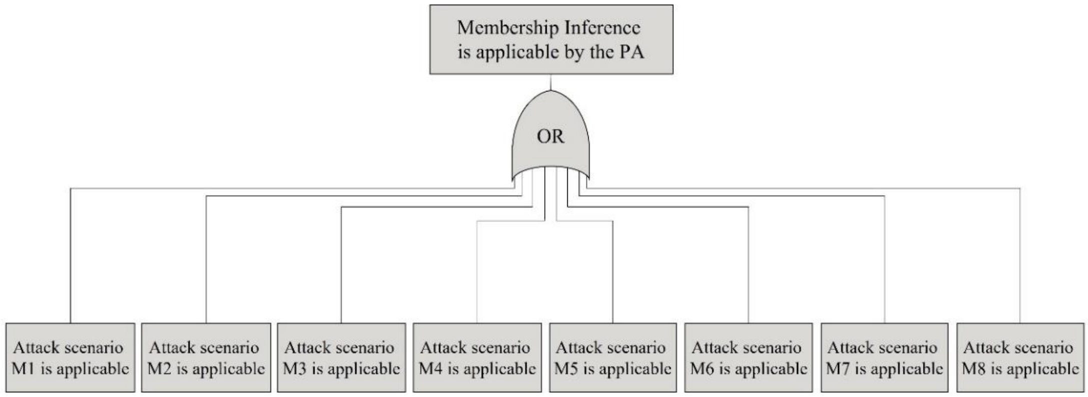

Figure II- 26. Example of the attack tree for membership inference (upper part)

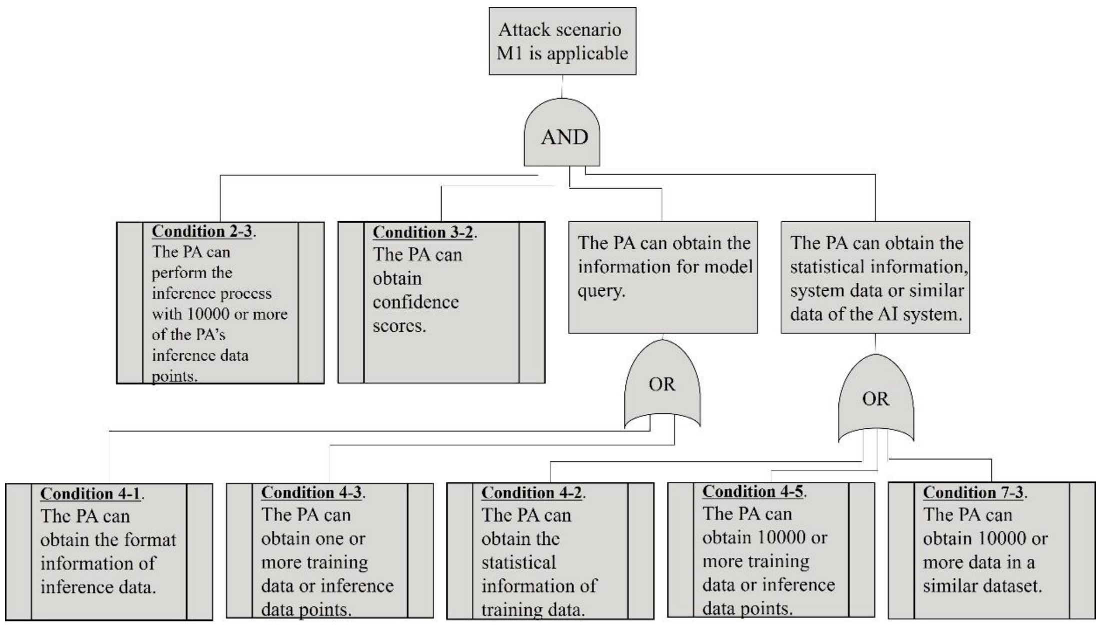

Figure II- 27. Attack tree and attack executable conditions for attack scenario M1 of membership inference

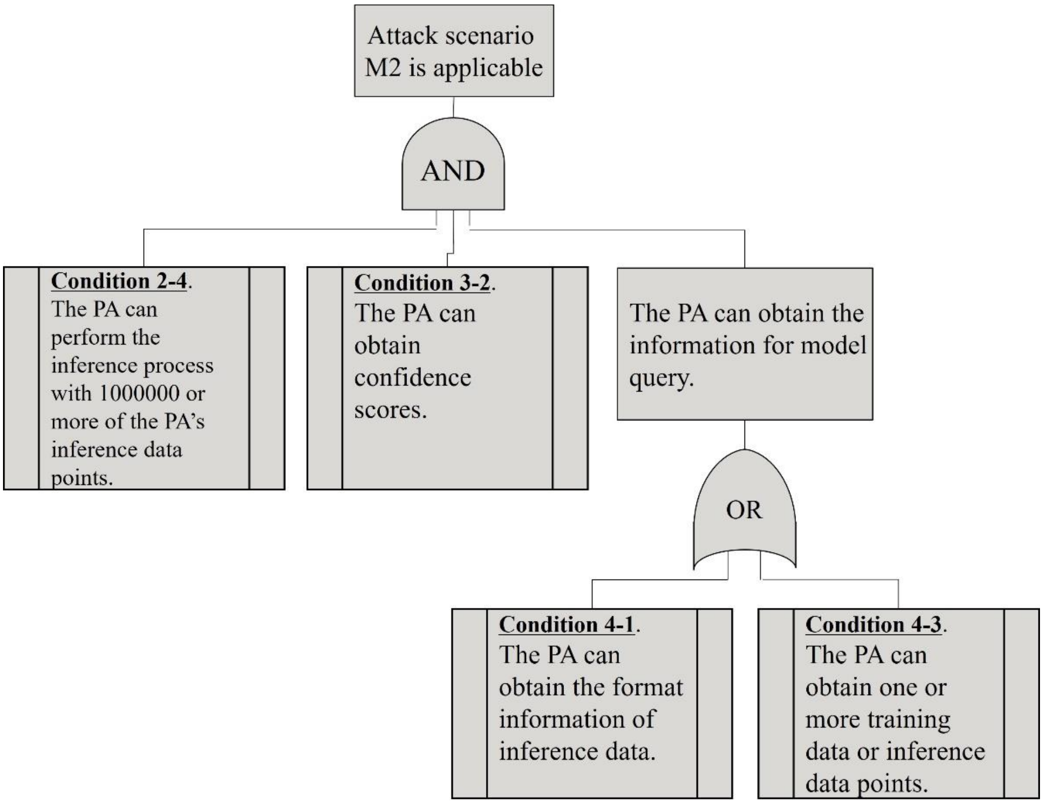

Figure II- 28. Attack tree and attack executable conditions for attack scenario M2 of membership inference

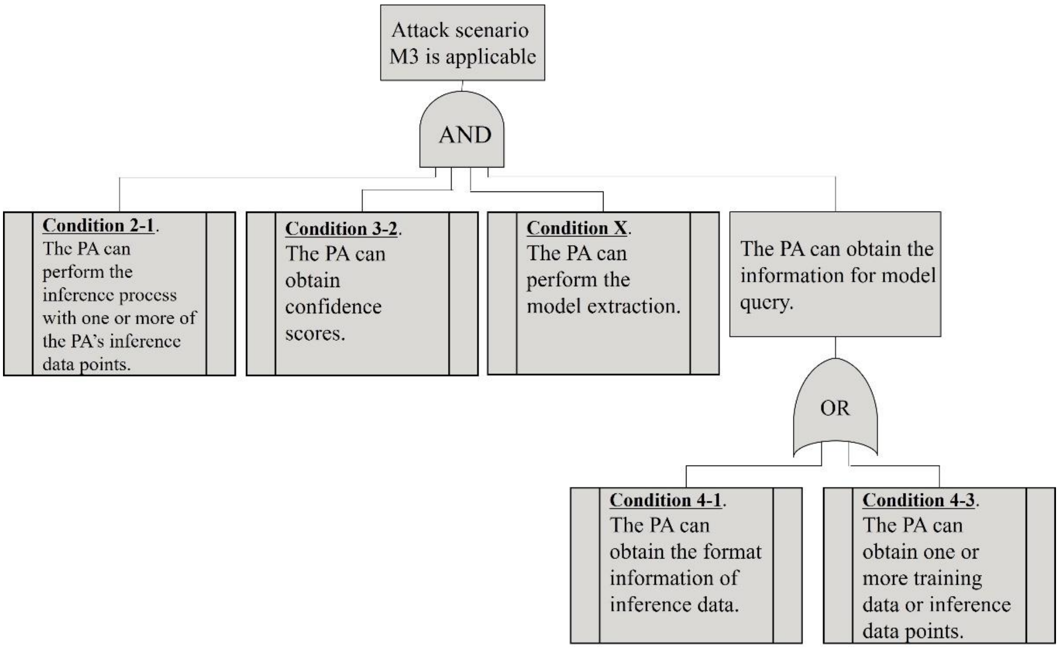

Figure II- 29. Attack tree and attack executable conditions for attack scenario M3 of membership inference

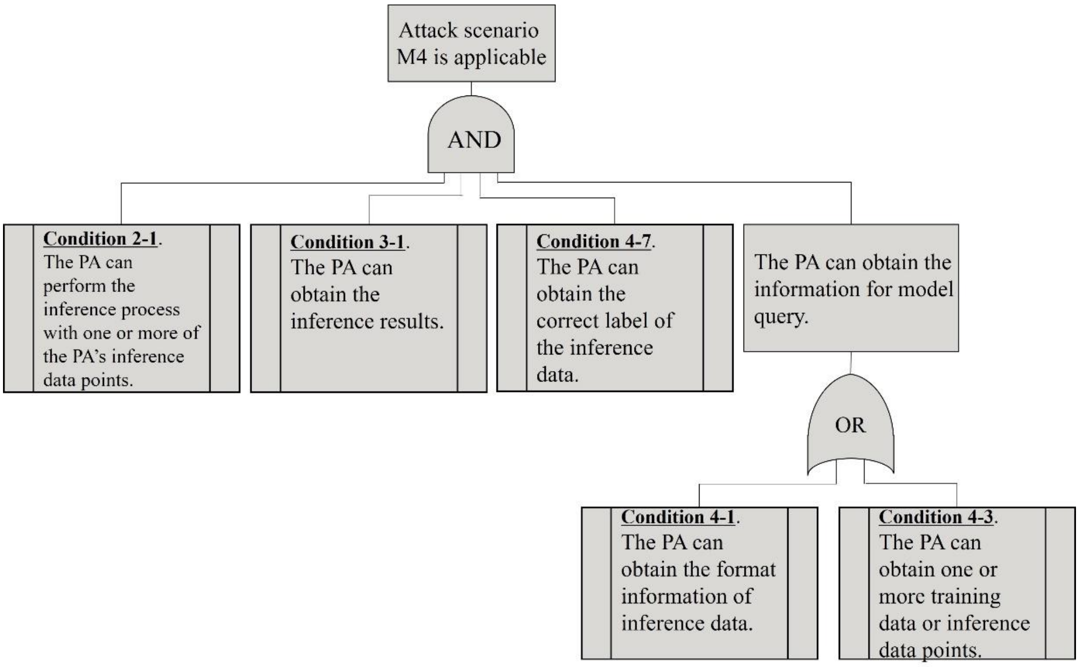

Figure II- 30. Attack tree and attack executable conditions for attack scenario M4 of membership inference

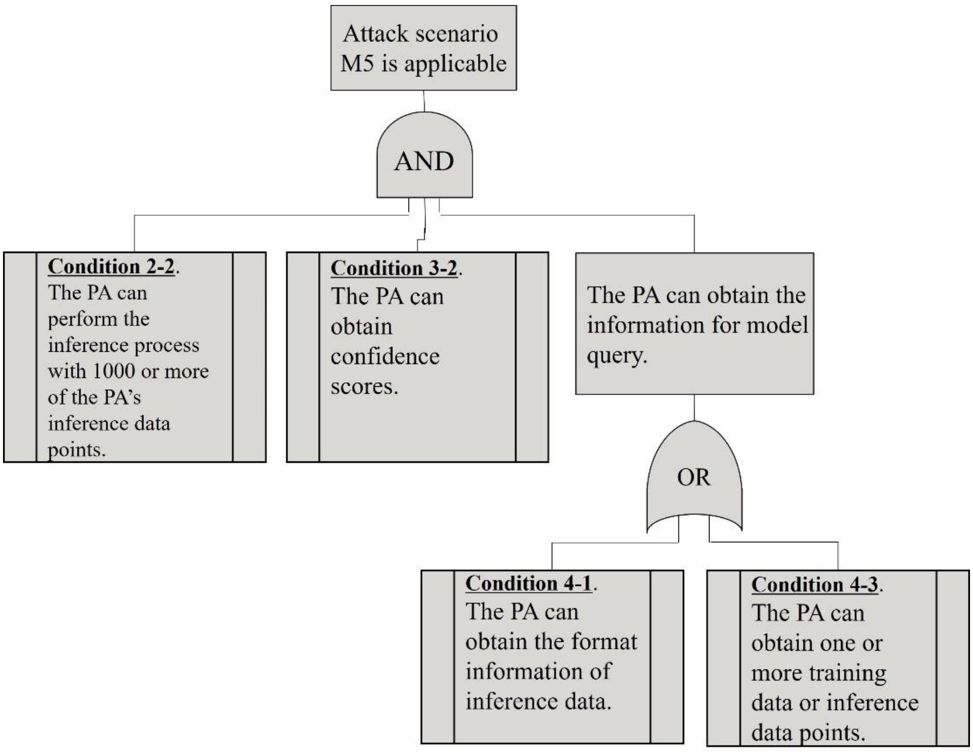

Figure II- 31. Attack tree and attack executable conditions for attack scenario M5 of membership inference

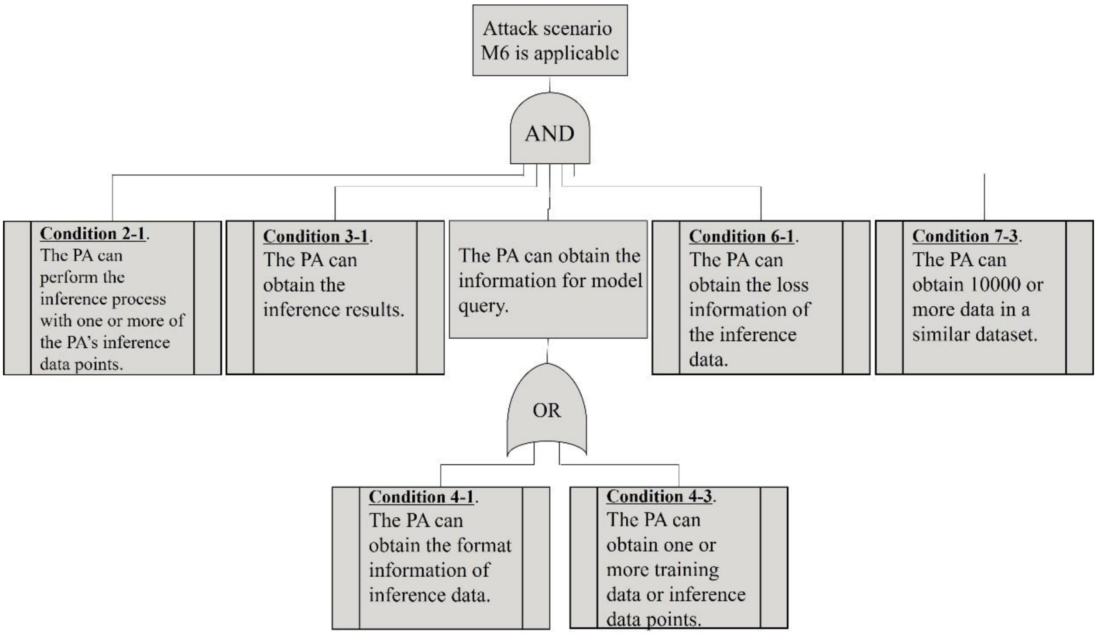

Figure II- 32. Attack tree and attack executable conditions for attack scenario M6 of membership inference

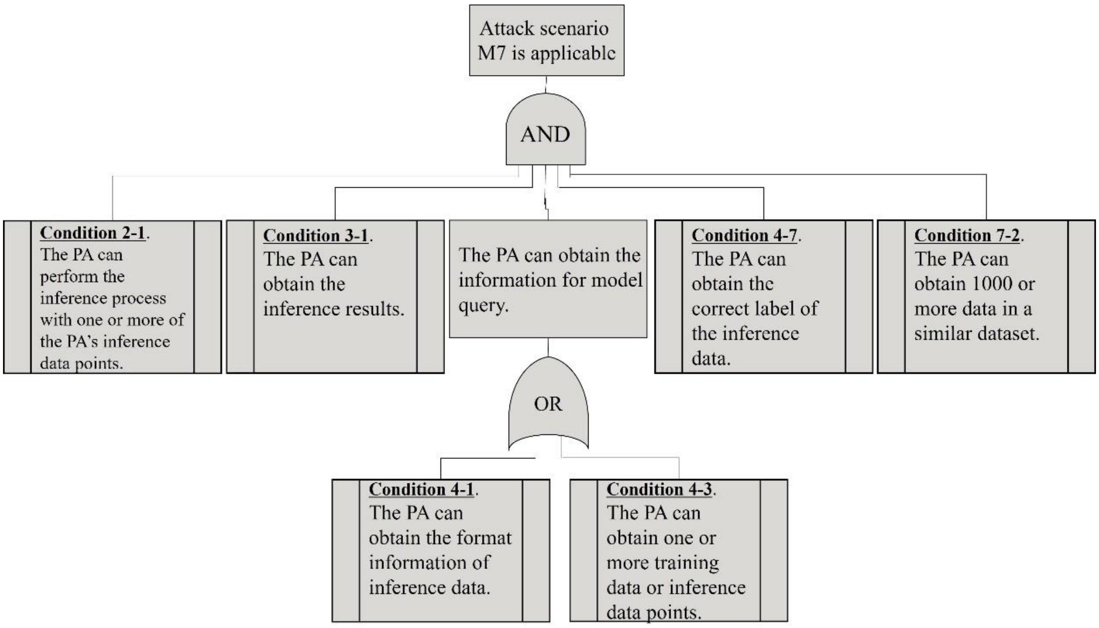

Figure II- 33. Attack tree and attack executable conditions for attack scenario M7 of membership inference

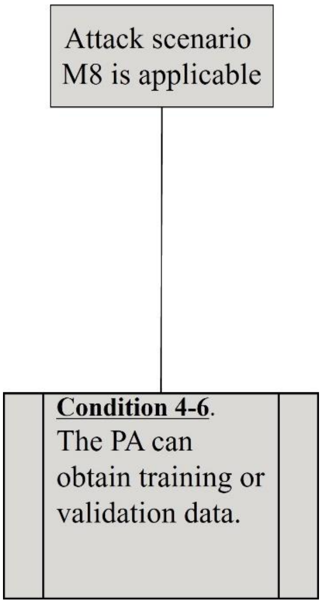

Figure II- 34. Attack tree and attack executable conditions for attack scenario M8 of membership inference

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	The answer to question 2-3A	TRUE	Prevent the PA from performing the inference

	inference process with 10000 or more of the PA's inference data points.	or question 2-3B is "Yes."		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	The answer to question 4-3 is "Yes."	TRUE	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 4-4	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 answer to question 4-7 is “Yes.”	TRUE	Do not expose detailed system specifications and prevent the PA from

	the inference data.			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	The answer to question 2-3A	FALSE	Prevent the PA from performing the inference

	inference process with 10000 or more of the PA’s inference data points.	or question 2-3B is “Yes.”		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.”	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	The answer to question 4-3 is “Yes.”	TRUE	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 4-4	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 answer to question 4-7 is “Yes.”	TRUE	Do not expose detailed system specifications and prevent the PA from

	the inference data.			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	The answer to question 2-3A	FALSE	Prevent the PA from performing the inference

	inference process with 10000 or more of the PA’s inference data points.	or question 2-3B is “Yes.”		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	The answer to question 4-3 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 4-4	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 answer to question 4-7 is “Yes.”	FALSE	Do not expose detailed system specifications and prevent the PA from

	the inference data.			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 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?	-
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,	Yes

	validation, or inference data points) of the AI system?	
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 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.”	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	The answer to question 4-1 is “Yes.”	FALSE	Prevent the PA from obtaining and estimating the format information of inference data.

	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).

Machine Learning System Security Guidelines, Appendix. “Overview of Detection Techniques for Machine Learning-Specific Attacks”

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

A-1. Introduction

Machine learning systems are affected by machine learning-specific attacks (MLSA) that can mislead decisions or infer information about a target model. Therefore, not only general security measures but also measures against MLSA must be taken. In addition to building systems that are robust against attacks, it is also important to detect attacks.

Many techniques for detecting MLSA have been proposed. However, they are often effective only for specific data or tasks, and a definitive detection technique has not yet been established. Carlini et al. evaluated several existing techniques for detecting adversarial examples and demonstrated that none of them could withstand adaptive attacks [A-1]. Kumar et al. cited attack detection as a challenge of machine learning security and proposed that detection techniques should be easily shared among security analysts [A-2]. Therefore, appropriate detection techniques are not easily implemented without expertise in machine learning security.

In general cyber security, MITRE ATT & CK frameworks [A-3], which systematize adversary tactics and techniques, can be used to select detection techniques [A-4]. In machine learning systems, MITRE ATLAS [A-5], which provides an attack strategy specific to machine learning (Reconnaissance, ML Attack Staging, Impact, etc.), can be used. However, as far as we know, no document systematizes techniques for detecting MLSA in relation to attack strategy. Therefore, in this appendix, the attack strategy is divided into two stages: precursor (reconnaissance or preparation for attacks) and indicator (attacks that mislead decisions or infer information about a target model), and the technique for detecting each stage was classified into precursor detection and indicator detection. Furthermore, whether the data used for detection can be acquired by the machine learning system is also important information when selecting detection techniques, so the detection techniques were also arranged in terms of the data used (training data, trained models, etc.).

This appendix summarizes the literature on attack detection in terms of adversary tactics and data used to assist developers and security analysts in selecting detection techniques. The target system is an image classification system, and the target attacks are evasion attacks, poisoning attacks, and model extraction attacks (the scope will be expanded in the future).

A-2. Classification of Detection Techniques Based on Adversary Tactics

In this appendix, adversary tactics are divided into two stages: precursor (reconnaissance or preparation for attacks) and indicator (attacks that mislead decisions or infer information about a target model), and the technique for detecting each stage was classified into precursor detection and indicator detection. In this section, the detection techniques for evasion attacks, poisoning attacks, and model extraction attacks are classified into two detection purposes, precursor detection and indicator detection (a summarized list is shown in Table A-1).

A-2.1. Precursor Detection

A-2.1.1. Evasion Attack Detection

In image classification, an attacker can mislead a machine learning model by adding small noise (perturbation) to the input image [A-6]. Such intentionally perturbed data are called adversarial examples (Nicholas Carlini assembled a list of papers on adversarial examples (arXiv) [A-7]).

The method for creating an adversarial example also depends on the knowledge of the attacker. For example, an attacker who does not have knowledge of the target system may input a series of queries against the system to create an adversarial example [A-8]. On the other hand, an attacker with knowledge about the target system can prepare a substitute model offline and create adversarial examples without inputting a query to the target system [A-9]. An attack that attempts to create an adversarial example through query input, as in the former example, should be detected as an attack precursor before the adversarial example is created. The detection techniques are described below.

- Detect attacks to create an adversarial example

Since an attacker is likely to input multiple similar data to create an adversarial example, detection techniques have been proposed that exploit the fact that the series of queries inputted by the attacker are distributed differently than those of legitimate users [A-10], [A-11]. In addition, since such an attack often results in more queries than those of legitimate users or in a biased output label distribution, it might be detected using a simple method such as monitoring the query frequency per unit time or the output label distribution.

A-2.1.2. Poisoning Attack Detection

There are two types of poisoning attacks: those that intentionally degrade the model inference accuracy by injecting malicious data into the training data [A-12], and those that inject backdoor data into the training data to misclassify specific input data into the target label (backdoor attack) [A-13]. There is a risk of poisoning attacks by outsourcing data preparation or model training, data collection from untrusted websites, federated learning, and transfer learning [A-14], [A-15], [A-16], [A-17].

Such poisoning attacks should be detected at the testing stage or earlier (as precursor detection). If the operational input data are retrained, whether the data is poisoned must be checked. The detection method is described below.

- Detect poisoned datasets that degrade the model performance

Several techniques are used to detect whether training datasets contain malicious data that would degrade performance [A-18], [A-19].

- Detect backdoors

Several techniques are used to detect backdoor data in training data sets [A-20], [A-21]. For a systematic and comprehensive review of backdoor attacks and countermeasures, see [A-17].

- Detect poisoned models

Several techniques are available for detecting whether a trained model obtained from untrusted sources is poisoned [A-22], [A-23].

A-2.2. Indicator Detection

A-2.2.1. Evasion Attack Detection

- Detect inputs of adversarial examples

As mentioned above, if attackers have knowledge of the target system, they may create an adversarial example in some way and input it into the system. To detect such attacks, input data, model inference, and intermediate layer output often must be analyzed [A-24], [A-25], [A-26], [A-27], [A-28], [A-29], [A-30], [A-31], [A-32]. However, as generally known, detecting adversarial examples is not easy [A-1]. Since there is no effective detection method for every system, multiple detection methods should be applied.

For a comprehensive review of the detection method for adversarial examples, see [A-33].

A-2.2.2. Poisoning Attack Detection

- Detect backdoor triggers injected into input data

Several techniques are used to detect a backdoor trigger injected into input data during operation [A-34].

A-2.2.3. Detecting Model Extraction Attacks

In model extraction attacks, an adversary prepares input–output pairs by inputting a series of intelligent queries into the system and generates a substitute model that behaves similarly to the victim model. The model extraction attack may infer information about a target model or be used for other attacks [A-35] (the latter attack is a precursor). Techniques for detecting a model extraction attack are shown below.

- Detect malicious queries for model extraction attacks

Since the log of a model extraction attack is likely to differ from that of a legitimate user, detecting techniques using this difference have been proposed [A-36], [A-37], [A-38], [A-39]. Alternatively, such attacks may be detected using simple techniques, such as monitoring query frequency per unit time, since the number of queries may be greater than that of normal users.

A-3. Data Used for Detection

This chapter summarizes the existing detection techniques in terms of the data used in their implementation. The data referred to here are the training data, the trained model, the input data in operation, and the output data of the model. These data can be divided into “Data to be analyzed” and "Data used for detection (not to be analyzed)." The former data needs to be managed in association with a date and time or an account (an example of the former data: the input image that may have traces of an evasion attack), and the latter data needs to be managed appropriately because it is required when implementing detection techniques (an example of the latter data: training data used to calculate detection thresholds). Table A-1 summarizes the detection techniques from the above viewpoints. Here, the output data refer to the confidence score or the intermediate layer output. They vary depending on the detection method, so see the paper on each detection method for details.

Table A-1. Data Used for Detection

X：Data to be analyzed

Y：Data used for detection (not to be analyzed)

Purpose	Attack	Detection	Development		Operation	
			Training data	Trained model	Input data	Output data
Precursor Detection	**Creation of Adversarial Examples**	Chen et al. [A-10]	Y		X	
		Li et al. [A-11]	Y		X	
	Data Poisoning (degrades the model performance)	Müller et al. [A-18]	X	Y		
		Tavallali et al. [A-19]	X			
	Backdoor	Chen et al. [A-20]	X	Y		
		Hayase et al. [A-21]	X	Y		
		Dong et al. [A-22]		X		
		Huster et al. [A-23]		X		
Indicator Detection	**Input of Adversarial Examples**	Hendrycks et al. [A-24]	Y		X	
		Meng et al. [A-25]	Y		X	
		Grosse et al. [A-26]	Y		X	
		Gong et al. [A-27]	Y	Y	X	
		Lu et al. [A-28]	Y	Y		X
		Feinman et al. [A-29]	Y	Y		X
		Aigrain et al. [A-30]	Y	Y		X

		Xu et al. [A-31]	Y	Y	X	X
		Monteiro et al. [A-32]	Y	Y	X	X
	Backdoor Triggers	Kiourti et al. [A-34]	Y	Y	X	X
	Model Extraction	Juuti et al. [A-36]			X	
		Pal et al. [A-37]	Y		X	
		Atli et al. [A-38]	Y		X	
		Sadeghzadeh et al.[A-39]	Y		X	

A-4. Summary

This appendix summarizes techniques for detecting evasion attacks, poisoning attacks, and model extraction attacks in terms of attack strategies and data used to assist developers and security analysts in selecting detection methods. By associating detection methods with attack strategies, this appendix can be used as a reference to detect precursor and indicator attacks at multiple levels or to detect attacks as early as possible. Further, by arranging the detection method in terms of the data, even when data use is restricted (e.g., input data cannot be acquired), a detection method that use only available data can be selected. The existing review papers [A-17], [A-33] are also helpful in selecting detection methods.

A-5. References

[A-1] N. Carlini , D. Wagner, “Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods,” In Workshop on Artificial Intelligence and Security, 2017.

[A-2] R. S. S. Kumar, M. Nyström, J. Lambert, A. Marshall, M. Goertzel, A. Comissoneru, M. Swann , S. Xia, “Adversarial Machine Learning -- Industry Perspectives,” In IEEE Security and Privacy Workshops, 2020.

[A-3] “ATT&CK,” MITRE, [online]. Available: https://attack.mitre.org/.

[A-4] B. E. Strom, J. A. Battaglia, M. S. Kemmerer, W. Kupersanin, D. P. Miller, C. Wampler, S. M. Whitley , a. R. D. Wolf, “Finding Cyber Threats with ATT&CK™-Based Analytics,” The MITRE Corporation, Bedford, MA, Technical Report No. MTR170202, 2017.

[A-5] “ATLAS,” MITRE, [online]. Available: https://atlas.mitre.org/.

[A-6] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow , R. Fergus, “Intriguing properties of neural networks,” arXiv preprint arXiv:1312.6199, 2013.

[A-7] N. Carlini, “A Complete List of All (arXiv) Adversarial Example Papers,” https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html, 2019.

[A-8] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi , C.-J. Hsieh, “ZOO: Zeroth Order Optimization based Black-box Attacks to Deep Neural Networks without Training Substitute Models,” In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, pp. 15–26, 2017.

[A-9] I. J. Goodfellow, J. Shlens , C. Szegedy, “Explaining and Harnessing Adversarial Examples,” In International Conference on Learning Representations, 2015.

[A-10] S. Chen, N. Carlini , D. Wagner, “Stateful Detection of Black-Box Adversarial Attacks,” In Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence, pp.30-39, 2019.

[A-11] H. Li, S. Shan, E. Wenger, J. Zhang, H. Zheng , B. Y. Zhao, “Blacklight: Defending Black-Box Adversarial Attacks on Deep Neural Networks,” arXiv preprint arXiv:2006.14042, 2020.

[A-12] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru , B. Li, “Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning,” In IEEE Security and Privacy, 2018.

[A-13] T. Gu, B. Dolan-Gavitt , S. Garg, “BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain,” In Proceedings of Machine Learning and Computer Security Workshop, 2017.

[A-14] Y. Chen, X. Gong, Q. Wang, X. Di , H. Huang, “Backdoor Attacks and Defenses for Deep Neural Networks in Outsourced Cloud Environments,” IEEE Network, vol. 34, no. 5, pp. 141–147, 2020.

[A-15] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin , V. Shmatikov, “How To Backdoor Federated Learning,” In International Conference on Artificial Intelligence and Statistics, 2020.

[A-16] Y. Ji, Z. Liu, X. Hu, P. Wang , Y. Zhang, “Programmable Neural Network Trojan for Pre-Trained Feature Extractor,” arXiv preprint arXiv:1901.07766, 2019.

[A-17] Y. Gao, B. G. Doan, Z. Zhang, S. Ma, J. Zhang, A. Fu, S. Nepal , H. Kim, “Backdoor Attacks and Countermeasures on Deep Learning: A Comprehensive Review,” arXiv preprint arXiv:2007.10760, 2020.

[A-18] N. M. Müller, S. Roschmann , K. Böttinger, “Defending Against Adversarial Denial-of-Service Data Poisoning Attacks,” arXiv preprint arXiv:2104.06744, 2021.

[A-19] P. Tavallali, V. Behzadan, P. Tavallali , M. Singhal, “Adversarial Poisoning Attacks and Defense for General Multi-Class Models Based On Synthetic Reduced Nearest Neighbors,” arXiv preprint arXiv:2102.05867, 2021.

[A-20] J. Chen, X. Zhang, R. Zhang, C. Wang , L. Liu, “De-Pois: An Attack-Agnostic Defense against Data Poisoning Attacks,” In IEEE Transactions on Information Forensics and Security, 2021.

[A-21] J. Hayase, W. Kong, R. Somani , S. Oh, “SPECTRE: Defending Against Backdoor Attacks Using Robust Statistics,” In Proceedings of the 38th International Conference on Machine Learning, 2021.

[A-22] Y. Dong, X. Yang, Z. Deng, T. Pang, Z. Xiao, H. Su , J. Zhu, “Black-box Detection of Backdoor Attacks with Limited Information and Data,” In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[A-23] T. Huster , E. Ekwedike, “TOP: Backdoor Detection in Neural Networks via Transferability of Perturbation,” arXiv preprint arXiv:2103.10274, 2021.

[A-24] D. Hendrycks , K. Gimpel, “Early Methods for Detecting Adversarial Images,” In Proceedings of the International Conference on Learning Representations, 2017.

[A-25] D. Meng , H. Chen, “MagNet: a Two-Pronged Defense against Adversarial Examples,” In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 135–147, 2017.

[A-26] K. Grosse, P. Manoharan, N. Papernot, M. Backes , P. McDaniel, “On the (Statistical) Detection of Adversarial Examples,” arXiv preprint arXiv:1702.06280, 2017.

[A-27] Z. Gong, W. Wang , W.-S. Ku, “Adversarial and Clean Data Are Not Twins,” arXiv preprint arXiv:1704.04960, 2017.

[A-28] J. Lu, T. Issaranon , D. Forsyth, “SafetyNet: Detecting and Rejecting Adversarial Examples Robustly,” In Proceedings of the IEEE International Conference on Computer Vision, pp. 446–454, 2017.

[A-29] R. Feinman, R. R. Curtin, S. Shintre , A. B. Gardner, “Detecting Adversarial Samples from Artifacts,” arXiv preprint arXiv:1703.00410, 2017.

[A-30] J. Aigrain , M. Detyniecki, “Detecting Adversarial Examples and Other Misclassifications in Neural Networks by Introspection,” In Proceedings of the 35th International Conference on Machine Learning, pp. 7167–7177, 2019.

[A-31] W. Xu , Y. Q. David Evans, “Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks,” In Proceedings of Network and Distributed System Security Symposium, 2018.

[A-32] J. Monteiro, I. Albuquerque, Z. Akhtar , T. H. Falk, “Generalizable Adversarial Examples Detection Based on Bi-model Decision Mismatch,” In 2019 IEEE International Conference on Systems, Man and Cybernetics, pp. 2839–2844, 2019.

[A-33] A. Aldahdooh, W. Hamidouche, S. A. Fezza , O. Deforges, “Adversarial Example Detection for DNN Models: A Review and Experimental Comparison,” arXiv preprint arXiv:2105.00203, 2021.

[A-34] P. Kiourti, W. Li, A. Roy, K. Sikka , S. Jha, “MISA: Online Defense of Trojaned Models using Misattributions,” arXiv preprint arXiv:2103.15918, 2021.

[A-35] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik , A. Swami, “Practical Black-Box Attacks against Machine Learning,” In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, pp. 506–519, 2017.

[A-36] M. Juuti, S. Szyller, S. Marchal , N. Asokan, “PRADA: Protecting against DNN Model Stealing Attacks,” In IEEE European Symposium on Security & Privacy, pp. 512–527, 2019.

[A-37] S. Pal, Y. Gupta, A. Kanade , S. Shevade, “Stateful Detection of Model Extraction Attacks,” arXiv preprint arXiv:2107.05166, 2021.

[A-38] B. G. Atli, S. Szyller, M. Juuti, S. Marchal , N. Asokan, “Extraction of Complex DNN Models: Real Threat or Boogeyman?,” In International Workshop on Engineering Dependable and Secure Machine Learning Systems. Springer, pp. 42–57, 2020.

[A-39] A. M. Sadeghzadeh, F. Dehghan, A. M. Sobhanian , R. Jalili, “HODA: Hardness-Oriented Detection of Model Extraction Attacks,” arXiv preprint arXiv:2106.11424, 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)