AI Ethics Impact Assessment

Procedure Manual
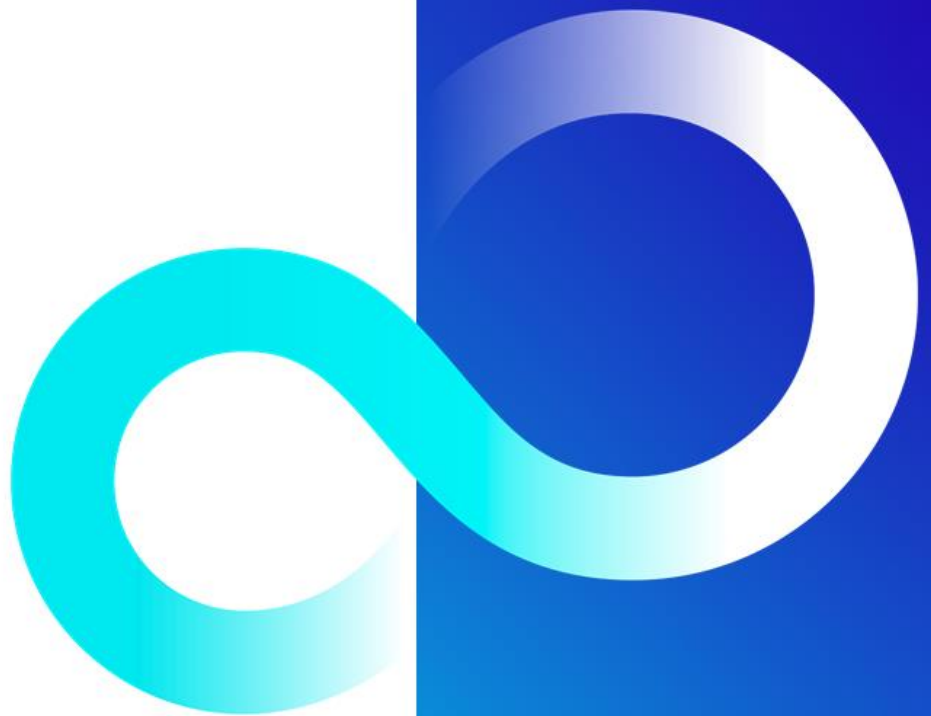
Table of contents

Revision History

Edition Count	Date	Contents
v 1.0	2022-3-31	New

Introduction

The present document presents a procedure manual for the AI Ethics Impact Assessment, an assessment we developed as a method for assessing the ethical risks that may arise in the use of AI systems.

In recent years, society has become aware of the ethical issues that AI raises. Examples include a facial recognition AI which produce racist results[1] and a recruitment AI that was suspended due to sexist results[2]. On the other hand, countries and organizations in Europe and elsewhere are attempting to address ethical issues by developing ethical principles and guidelines for the use of AI[3] [4] [5] [6] [7] [8]. The European Commission has gone even further and proposed a draft AI regulation[9].

Thus, while it has become essential to address ethical risks in the social implementation of AI, the fact that AI systems have multiple stakeholders and the social conditions surrounding them are changing, it is necessary to appropriately address the ethical issues which may arise from the use of AI systems. Recognizing when to do so is a challenge.

With this in mind, we have developed an AI Ethics Impact Assessment to help developers and providers of AI systems assess the ethical implications of their own use cases. We applied this method to a case study build from already known cases of ethical incidents and verified that it is possible to use this method to understand where and how ethical risks can occur in an AI system. The present document presents the procedure manual for the method.

We hope that the present document will help AI developers, AI service providers, business users, and others involved in AI to be aware of the ethical issues that can arise.

PLEASE NOTE THAT THE AI ETHICS IMPACT ASSESSMENT DOES NOT WARRANT THAT NO ETHICAL CONCERN WILL ARISE.

We will continue to improve this methodology through discussions and examinations with various stakeholders based on the present document.

This manual is organized as follows:

Chapter 1: Purpose and structure of AI Ethics Impact Assessment

Chapter 2: An overview of AI Ethics Impact Assessment

Chapter 3: Procedures for AI Ethics Impact Assessment

Chapter 4,5: Application Examples

Expected readers and usage scenes of this book

■ **Intended audience for the present document**

- Sales representatives and engineers considering proposals for AI systems to their customers
- AI algorithm developers, researchers, and data scientists
- Companies and organizations considering applying AI systems to their own business
- AI system users and their stakeholders

■ **Life cycle of the AI system covered in the present document**

- The present document covers the process of planning, designing, developing, and operating AI systems.

■ **Usage cases for the present document**

- Assessing and explaining the possible ethical risks of AI by a sales representative to a customer when planning an AI system
- Evaluating possible ethical risks of AI system components during the design phase of the AI system
- Evaluating possible ethical risks of AI prior to the operation of an AI system
- Evaluating possible ethical risks of AI for AI systems already in operation
- User assessment of an AI system to evaluate the possible ethical risks of AI before implementing the AI system.

Disclaimer

- THE CASES DISCUSSED IN THE PRESENT DOCUMENT ARE FICTITIOUS, ALTHOUGH THEY ARE BASED ON REAL CASES THAT HAVE OCCURRED IN THE PAST (INFORMATION REGISTERED IN THE AI INCIDENT DATABASE[11] PUBLISHED BY THE PARTNERSHIP ON AI[10], AN INTERNATIONAL AI CONSORTIUM), AS WELL AS SIMILAR CASES AND AI USE CASES PUBLISHED BY THE ISO[12]. THE DETAILS ARE NOT THE SAME AS IN THE ACTUAL CASES, HOWEVER.
- THE CONCEPT OF ETHICS VARIES DEPENDING ON INDIVIDUAL VALUES, CULTURE, RELIGION, SOCIAL CONDITIONS, AND TECHNOLOGY IN EACH COUNTRY OR REGION, AS WELL AS THE TIMES. THE CONTENTS OF THE PRESENT DOCUMENT DO NOT REPRESENT THE OPINIONS OF FUJITSU; THEY ARE MERELY ILLUSTRATIVE EXAMPLES.
- FUJITSU DOES NOT WARRANT THAT THERE ARE NO RISKS OTHER THAN THOSE POINTED OUT IN THE PRESENT DOCUMENT; EACH ORGANIZATION OR INDIVIDUAL READER OF THE PRESENT DOCUMENT IS RESPONSIBLE FOR DETERMINING AND IMPLEMENTING MEASURES IN SPECIFIC CASES.
- FUJITSU IS NOT LIABLE FOR ANY DAMAGES INCURRED BY THE READER IN CONNECTION WITH THE PRESENT DOCUMENT OR THE METHOD DESCRIBED THEREIN.

1. Objectives and structure of AI ethics impact assessment

1.1 Objectives

AI ethics impact assessment is a method for systematically analyzing ethical risks that may occur in AI systems based on AI ethics guidelines. The purpose of this project is to enable stakeholders involved in AI systems to recognize the ethical problems that can arise from the use of AI systems and the factors that cause the problems as risks in all phases of the AI system life cycle from planning to development, operation and retirement.

1.2 Scope of AI ethics impact assessment

The AI system targeted for AI ethics impact assessment does not limit the industry in which the AI system is operated, or the applications used by the AI system. In addition, whether the target of AI's judgment is a person or organization is subject to analysis.

The scope of the AI Ethics Impact Assessment is the scope of analysis including AI systems and stakeholders directly or indirectly involved in AI systems.

1.3 Terminology Description

The following terms are used in this manual.

Terminology	Description
AI ethical principles/guidelines	Principles/guidelines published by countries and organizations that AI systems must follow to behave ethically. This manual refers to the Ethics guidelines for Trustworthy AI published by the European HLEG as one of these principles.
Ethical risk of AI	Risks arising from ethical issues raised by AI systems, including those that have a positive impact on the AI system or its stakeholders, as well as those that have a negative impact on the same.
Risk event	Ethical risks of AI that affect stakeholders, as well as risks that stakeholders may be exposed to that result in economic loss, loss of social credibility or, conversely, in increased revenue or social credibility for the stakeholder
Risk factor	Factors that cause ethical risks of AI, specifically risk events. Risk events can be a factor in other risk events.
System diagram	Diagram showing the components of the AI system and the stakeholders, and also describing the interactions
Interaction	A relationship between any two components of an AI system (including its stakeholders). In an AI Ethics Impact Assessment, ethical risks are extracted by linking them to interactions.
AI ethics model	A model that uses use cases to structure the principles of AI ethics and correlates them with interactions of an AI system.
AI ethical characteristics	Characteristics that an ethical AI system should satisfy, which are extracted by embodying AI ethical guidelines. Examples: "Guarantee fundamental human rights," "Group fairness"

1.4   Structure of the practical guide

Procedures, sheets, and drawings for conducting AI ethics impact assessments will be published as a practical guide. The practical guide is organized as follows:

1. Manual: This document. The procedure of AI ethics impact assessment is explained, and a part of application examples are described.
2. AI Ethics Model: A table describing models used for risk extraction in AI Ethics Impact Assessments.
3. Use Case Summary Sheet: This table is used to clarify the items required for analysis of the AI system to be evaluated.
4. AI system pattern sheet: A diagram illustrating the components and stakeholders of an AI system and their relationships. It consists of five drawing patterns.
5. Analysis sheet: This table is used to identify risks using AI ethics models.

2. AI Ethics Impact Assessment Overview

2.1   overall picture

AI Ethics Impact Assessment analyzes where and how ethical risks that could arise from the use of AI systems appear. The analysis results are presented on a system diagram illustrating the AI system, stakeholders, and interactions between them. This method can be implemented procedurally if the user has knowledge of the AI system to be evaluated, and the evaluation can be performed reliably in a realistic time.

This section explains how AI ethics impact assessment is implemented, along with the overall diagram shown in Figure 2-1.

First, create an AI system diagram based on the AI specifications and use case information for the AI to be evaluated. A system diagram (Figure 2-1 on the left) is a diagram in which the AI system components (data, AI algorithms, etc.) and related stakeholders are arranged, and their relationships are shown by arrows. This type of relationship is called interaction. This system diagram can be created using the AI system pattern sheet provided in the Practical Guide.

Next, the requirements for interactions in the system diagram are extracted using the AI ethics model (center of Figure 2-1). The AI ethics model is a mapping between the AI requirements described in the ethics guidelines, which are broken down and organized (Figure 2-1, tree structure of the AI ethics model in the center), and the interaction where ethical problems can occur due to violations of the requirements. The practical guide provides this AI ethics model in a tabular format sheet, enabling the implementation of this method without knowledge of ethics guidelines.

Finally, the extracted requirements are assigned to the interaction of the AI system diagram to be evaluated, and ethical issues are extracted as risk events and risk factors (Figure 2-1 right).
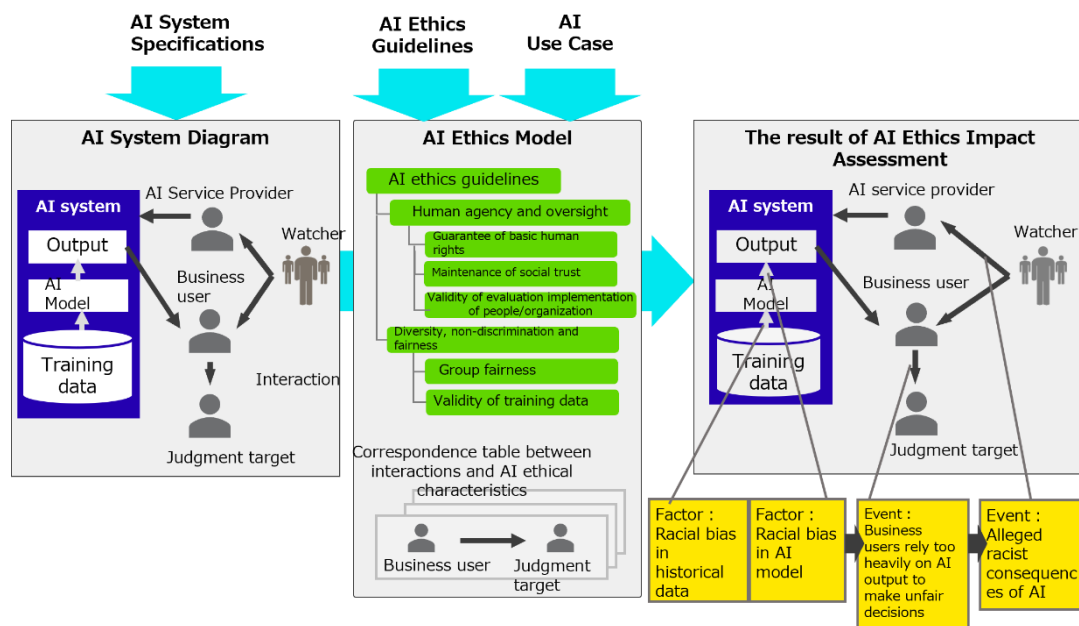
Figure 2-1 Overview of AI ethics impact assessment

2.2 Outline of the Implementation Procedure

The AI Ethics Impact Assessment consists of three steps:

Step 1: Create the System Diagram

Step 2: Extract Interactions

Step 3: Extract Risk

Table 2-1 provides an overview of each step, the sheets to be used, the corresponding table, and the inputs and outputs. The usage sheets and correspondence tables are tabular sheets and drawings provided by the practical guide.

The evaluator of the AI ethics impact assessment shall have the service outline, usage scene, and overall specification information for the AI system to be evaluated. In Step 1, based on this information, the items required for evaluation are clarified and presented as a system diagram. In Step 2, the interaction between the AI system components and stakeholders is identified using the created system diagram, and interaction information is described in an analysis sheet. In Step 3, the AI ethics model is used to examine what risks may occur for each interaction identified in Step 2, and the extracted risks are described in an analysis sheet. Finally, the relationship between the extracted risks is examined, and they are illustrated on the system diagram and completed as an analysis chart.

Table 2-1. Outline of procedures for AI ethics impact assessment

Procedure	Step 1 Creating a System Diagram	Step 2 Extracting Interactions	Step 3 Risk extraction
Task Overview	From the system specification, items necessary for evaluation are clarified, and a system diagram is prepared.	Identify interactions and detail them on an analysis sheet.	Using the AI ethics model, the risk is examined for each interaction, and the extracted risk is illustrated in the analysis chart.
Usage sheet and correspondence table	・Use Case Summary Sheet ・AI system pattern sheet	・Analysis sheet	・AI ethics model
Input	Information such as an overview of AI system services, usage scenes, or specifications	・Use case summary sheet and system diagram created in Step 1	・Analysis sheet created in Step 2 ・System diagram created in Step 1
Output	・Created Use Case Summary Sheet ・System diagram	・Analysis sheet (intermediate)	・Analysis sheet (finished product) ・Analysis chart

3. Procedures for conducting AI Ethics Impact Assessments

This chapter describes the procedure in detail. When conducting an AI Ethics Impact Assessment, the evaluator must have information about the specifications or use cases of the AI system being evaluated.

3.1   Step 1 Create a system diagram

Using the use case summary sheet and the AI system pattern sheet included in the practical guide, clarify the items necessary for evaluation from the specifications of the AI system to be evaluated, and create a system diagram. The AI system handled by this method is based on the following configuration.

- consist of a training department and a prediction department
- The training part generates the AI model by the algorithm of machine learning and statistical analysis using the training data as an input and applies it to the prediction part.
- The prediction part outputs the inference result of the AI model to the input inference data.
- Inference results can be output as is or through the service UI/API or business user.

The AI system pattern sheet provides a drawing of five variations with this basic configuration. Use the Use Case Summary Sheet to clarify the information required to select a pattern.

3.1.1   Clarification of use cases

First, the use case summary sheet contains the information required for evaluation. Table 3-1 shows the use case summary sheet. Required items are indicated by "Major item" or "Medium item". The evaluator writes the information that is clear at the time of evaluation in the Contents column for each item. Refer to the "Description" and "Example" columns for each entry. "Unknown" is indicated for unclear items.

The main item "Stakeholders and their roles" is to classify and clarify the stakeholders in detail. This method defines the role of stakeholders in detail by extending the target stakeholders to the extent that they are assumed to influence each other directly or indirectly with the AI system. The medium items "AI service provider", "Developer", "Business users", and "Consumer-like users" are based on the definitions in the Ministry of Internal Affairs and Communications AI Utilization Guidelines[7]. Stakeholders involved in training data and provided data are classified into providers, sources of acquisition. The judgment target is the person or organization that the AI is trying to

infer. For example, in an AI that evaluates the personnel of an employee, the employee is a judgment target. "Observers" are organizations that monitor AI, such as media and human rights organizations. Stakeholders whose roles are unclear are classified as "Other stakeholders". Stakeholders classified in this way can pattern how they interact with the components of the AI system and other stakeholders by their roles. These patterned drawings become AI system pattern sheets.

The main item "AI task" clarifies the breakdown and output of the AI model described in the system diagram and clarifies the types of problems such as classification and regression, as well as the characteristics such as additional learning and real-time performance. The characteristics of these AI tasks are in terms of assuming risks in the risk extraction step. For example, by clarifying that the AI task to be evaluated is a classification problem that determines whether the person to be evaluated is acceptable or not, it becomes easier to assume that the rate at which the classification results are acceptable varies greatly depending on gender in the risk extraction step.

The main items "Training data" and "Inference data" are the breakdown of the data to be described in the system diagram, the stakeholders involved are clarified, and the items to be noted for fairness and privacy are described. These items also serve as perspectives for assuming risks in the risk extraction step. For example, if there is a gender in the training data "Protected attributes and attributes used for fairness verification and mitigation", it is easier to assume in the risk extraction step that the training data may be biased toward one gender.

3.1.2   Selecting AI System Patterns

After clarifying the use case, select a pattern like the AI system to be evaluated from the AI system pattern sheet, and enter the names of the AI system components and stakeholders described in the use case summary sheet.

When selecting a pattern, refer to the explanation of stakeholders in the AI system pattern sheet and select a pattern with similar roles. Modify the system diagram if you need to add or remove AI system components or stakeholders. After that, the system diagram shows the required interactions.

Figure 3-1 shows an example of the resulting system diagram. A system diagram consists of an AI system, stakeholders directly or indirectly involved with the AI system, and interactions between them. Each element is described below.

- AI system

The AI system consists of a training part and an inference part. The training part consists of three components: training data, machine learning and statistical analysis, and AI models. Algorithms for learning training data correspond to "Machine learning or statistical analysis". Machine learning or statistical analysis generates an AI model. If the configuration is more complicated, such as when there is pre-processing to generate training data or when multiple AI models are handled, you can add the necessary components.

The inference part consists of three components: inference data, AI model, and inference results. The AI model was generated by the training part. The result inferred by the AI model for the input inference data is the inference result. The name of AI Model corresponds to the item "Breakdown of models required to perform AI tasks" in the use case summary sheet, and the name of Inference Results corresponds to the item "Outputs" in the use case summary sheet. Similarly, to the training part, if the inference part has preprocessing of inference data, postprocessing of inference results, or multiple AI models, the necessary components are added.

- Stakeholders

Stakeholders corresponding to the items in "Stakeholders and Roles," the main item in the use case summary sheet, correspond to human icons on the AI system diagram. A stakeholder has more than one role. In Figure 3-1, "Consumer-like user", "Inference data source" and "Judgment target" are associated with one stakeholder, and the roles of "Business user" and "Inference data provider" are associated with another stakeholder. In some use cases, these roles can be different stakeholders. In such cases, add stakeholders.

- Interaction

Interaction is described in the AI system pattern as an arrow with either a component of the AI system, data, or a stakeholder role at each end.

First, let's look at the interaction that shows the flow of data in Figure 3-1. Training data is data that the training data provider obtains from the source of training data and provides to the AI system. Inference data is data that the inference data provider obtains from the inference data source and inputs to the AI model. The inference result output by the AI model based on the inference data is received by the business user, and the final decision is made and output. The "Consumer-like user" receives this output.

Next, let's look at the interactions involving stakeholders in Figure 3-1. Output interacts with not only the consumer but also the target audience. This interaction

indicates that the "Output" is what determines a person or organization. An AI service provider interacts with an AI system. We set up this interaction because AI service providers are involved from the design to the operation of the AI system. Developers also interact with each component of an AI system because they are involved in analyzing training data to develop appropriate algorithms and validating AI models by making inferences.

Stakeholders who are not directly affected by the output of the AI system, such as "Observers" and "Other stakeholders" are also listed. These stakeholders interact primarily with other stakeholders.

When a stakeholder has multiple roles, each role connects one or more interactions. Figure 3-1 shows a total of 5 interactions for stakeholders who play the roles of "Consumer-like user", "Inference data source", and "Judgment target", including 3 interactions linked to "Consumer-like user", 1 interaction linked to "Inference data source", and 1 interaction linked to "Judgment target".

3.1.3   Adding Additional Elements to the System Diagram

If you have added AI system components and stakeholders to the selected system pattern, enter the interactions that they relate to. When you add a component of an AI system, describe the input and output of the component and its relationship to the preceding and following components as interactions. When stakeholders are added, consider whether there is a relationship with input/output of the AI system or a relationship with other stakeholders, and describe all possible relationships as interactions. You can assign an interaction ID to each interaction and use it in the following steps. There are no specific rules for how interaction IDs are assigned but be careful not to conflict.

Once the possible AI system components, stakeholders, and interactions are described on the system diagram, proceed to Step 2.

Table 3-1 Use Case Summary Sheet

Major Item	Medium Item	Details	Item Description	Example
Industry			The expected type of business or industry in the AI ethics case. Fill in based on the International Standard Industrial Classification[13].	- Finance/Insurance - Public service and national defense/Mandatory social security services
Purpose			Purpose of using an AI system	- To shorten loan approval time
Service	Service Overview		Services provided by AI system providers using AI systems	- The AI responds to loan applicants with loan decisions
	Availability of customization for each customer		Indicates whether the system requires customization for each customer	Write one of: Yes/No/Unknown
	Requirements		Customer-set requirements for the service	Write one of: Yes/No/Unknown
Usage case			Indicates the characteristics of the users of the AI system and their environments, and outlines the tasks to be performed using the AI system	- Loan applicant submits application information via an app, and receives the review results via the app
Stakeholders and their roles	AI Service Provider *		A kind of stakeholder; performs operations using the developed AI system and provides various services	- Loan screening AI service developer
	Developer*		A type of stakeholder that develops AI systems	- AI developer
	Business users*		A type of stakeholder; people who use AI systems or AI services in the course of business	- Bank personnel who make loan decisions based

				on the inference results of the loan screening AI - Police officers who make investigative decisions on people recognized as suspicious by facial recognition AI - Doctors who make treatment decisions based on the inference results of medical image diagnosis AI
	Consumer-like users*		A type of stakeholder; users of AI systems or AI services who are not business users	- Loan applicants using a loan screening AI - Patients undergoing medical imaging diagnosis - Chatbot users
	Training data provider		A kind of stakeholder. A person who provides the original data to create training data	- Credit bureaus, banks - Holders of datasets for natural language processing
	Source of training data		A type of stakeholder. Persons who are directly/indirectly related to the training data provider	- Credit bureaus, banks - People who provide their own face images to the face image dataset
	Parties involved in training data acquisition		A type of stakeholder; people, organizations, or systems directly or indirectly involved in training data acquisition	- People who design credit scores, people to whom credit scores have been provided in the past - Photographers who take facial images -Doctors and technicians who take medical images
Stakeholders and their roles	Inferential data providers		A kind of stakeholder; a person who provides input data to create inferential data	- Credit bureaus, banks - Hospitals that provide medical image data

		Inferential data source		A type of stakeholder; people or organizations involved in some capacity with the content of inferential data	- Credit bureaus, banks - Patients whose images are used in medical image data
		Parties involved in the acquisition of inferential data		A kind of stakeholder; people who are directly/indirectly involved in acquiring inferential data	- Photographers who take facial images - Doctors and technicians who take medical images - The people in the image
		Observers		A type of stakeholder. A person monitoring an AI system or AI service	- Human rights organizations, media
		Service UI/API provider		People, organizations, or systems that build functions for AI system users based on the inference results of the AI system	- Provider of monitors and controls for medical diagnostic imaging - Provider of robot control unit for factory robots - Providers of platforms such as SNS on which chatbots operate
		Judgment target		Persons or organizations to be judged or evaluated by the AI system	- People captured by surveillance cameras, employees evaluated by HR AI
		Service authorizer		Persons or organizations that authorize the development of AI systems or the provision of services using AI systems	- Relevant ministries, regulators
		Other stakeholders		A type of stakeholder; a person or organization indirectly affected by the output from an AI system or AI service.	- Family members and business partners of users - Insurance company with which the user has a contract - The community to which the user belongs

Major Item	Medium Item	Details	Item Description	Example
Presence/absence of a human-in-the-loop			Indicates whether the AI system will include the judgment of the AI system user with respect to the inference results. Risk events and risk factors to be extracted differ depending on whether human judgment is involved.	- The bank's loan officer makes the final decision on a loan assessment pass/fail with a human-in-the-loop. No human-in-the-loop when AI inference results are communicated directly to the loan applicant
Presence/absence of existing methods			Indicates whether the task to be performed/supported by the AI system can be performed by existing methods. If existing methods can be used, efficiency and accuracy must be compared between the existing methods and the AI system.	
AI Task	Task		An overview of the inferential data, which is the input to the AI model, and the inference results, which is the output	- Input loan applicant's application information, credit score, and transaction data, and output loan decision
	Problem Classification		Types of problems handled by the AI model	Classification, recommendation, regression, natural language processing, speech recognition, image recognition, image generation
	Output		AI Model Output	- Loan assessment pass/fail - Conversational text of a chatbot

Technology		Source of the AI model, indicating whether it was created during the development of the AI system, by others, or a combination	Own, OSS, Own + OSS, other companies' technology
Breakdown of models required to perform AI tasks		Indicate the names of all AI models required to perform the AI task	- Speech recognition model, sentence comprehension model, facial expression recognition model
Update/additional training of training data		Indicate whether additional training is required with updated training data after the start of operation	Write one of: Yes/No/Unknown
Real-time performance		Whether real-time performance is required for processing AI tasks or not	Write one of: Yes/No/Unknown
Verification of AI Tasks (Consistency with decisions made by existing methods)		If similar tasks have been performed by existing means, an indication of the accuracy of the AI task or differences in individual results compared to existing means	- Comparison with decisions made by loan reviewers

Major Item	Medium Item	Details	Item Description	Example
Training Data	Breakdown, provider and source		Breakdown of training data and respective data providers and data sources	- Transaction data of past loan applicants (banks)
	Teacher labels		Details of teacher labels to be set on training data	- Loan availability/non-availability
	Protective attributes, attributes used for fairness verification and mitigation		Protective attributes included in training data (Gender, age, nationality, race)	- Gender, age
	Presence/absence of personal information		Indicates whether personal information is included in the training data	Indicate whether yes/no/unknown
	Notes on data acquisition		Things to keep in mind when acquiring training data	Obtain permission from those involved with the training data
	Data Storage		Whether training data can be stored or not and precautions to be taken when storing training data	Training data may not be stored

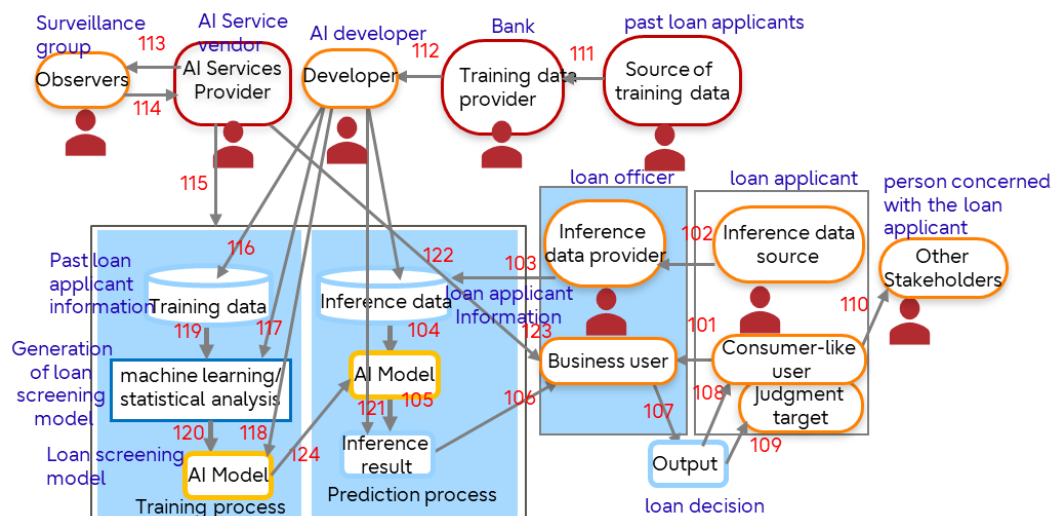| Major Item	Medium Item	Details	Item Description	Example
Inferential data	Breakdown, provider and source		Breakdown of inferential data and respective data providers and data sources	- Credit scores of loan applicants (credit bureaus)
	Protective attributes, attributes used for fairness verification and mitigation		Protected attributes included in inferential data (Gender, age, nationality, race)	- Gender, age
	Presence/absence of personal information		Indicates whether personal information is included in the inferential data	Indicate whether yes/no/unknown
	Notes on data acquisition		Things to keep in mind when acquiring inferential data	Is it necessary to obtain permission from the data source? What are the terms of permission, etc.
AI System Output	Things to keep in mind during output		Things to keep in mind during AI system output	- Provide a point of contact to discuss the results of AI
Citations, reference articles			Incident Database Registration Link to URL or explanatory article	

Figure 3-1 Example of system diagram

3.2   Step2 Extract Interactions

List the interactions you identified in Step 1. Use the "analysis sheet" in the analysis sheet file provided by the practical guide. Table 3-2 shows an analysis sheet that describes interactions based on the system diagram in Figure 3-1. An analysis sheet is equivalent to one row per interaction. In this step, include the Interaction ID, the Type and Name of the Interaction Start and Interaction End, and the Interaction Overview.

The Interaction Start and Interaction End points correspond to the start and end stakeholders, or AI system components as indicated by the arrows in the system diagram. Type corresponds to the stakeholder role or element name of the AI system component described in the AI system pattern. Type is selected from the drop-down list. Provide specific names in the Name column for the start and end interaction points. Interaction overview describes what happens between the start and end points.

Example of an analysis sheet described based on the system diagram of Table 3-2

Interaction ID	Interaction Start Point		Interaction End Point		Inter-action overview	AI ethical characteristics	risk event	risk factor
	Type	Name	Type	Name				
101	Consumer-like user	loan applicant	business user	loan officer	a loan applicant applies to a loan officer			
102	Inference data source	loan applicant	inference data provider	loan officer	loan applicants provide application information to loan officers			
103	inference data provider	loan officer	inference data	loan applicant's information	The loan officer enters loan applicant's information			
104	inference data	Loan applicant's information	AI Model	loan screening model	Input inference data to AI model			
105	AI Model	loan screening model	inference result	loan screening result	AI model outputs results			
106	inference result	loan screening result	business user	loan officer	the loan officer receives the AI results			

3.3   Step 3 Extracting risks

This task extracts risks using an AI ethics model, interaction derived from the AI ethics model, and a table of AI ethics characteristics (Interaction correspondence table below). Enter information in the AI ethical characteristics, risk events, and risk factors columns of the analysis sheet created in Step 2.

3.3.1   AI ethics model

The AI ethics model embodies "Ethics guidelines for trustworthy AI " (Below, Trustworthy AI) [4] from EU AI HLEG, which is one of the AI ethics guidelines, by comparing it with AI use cases, and correlates it with the type of interaction. Trustworthy AI provides an assessment list consisting of 7 requirements and approximately 100 questions, and the AI ethics model embodies them in a tree structure.

The practical guide provides an AI ethics model as a sheet consisting of two tables: the AI Ethics Model and the Interaction Correspondence Table. Table 3 -3 shows a portion of the AI Ethics Model sheet. The columns "Classification 1" through "Classification 4 (AI ethical characteristics)" in the table represent the AI ethical characteristics of each layer of the tree structure. "Category 1 " corresponds to the seven requirements for Trustworthy AI. Categories 2 through 4 are AI ethical characteristics that specify each requirement in a hierarchical manner. Category 1 is followed by Category 4, which corresponds to a tree leaf. An overview of each line describes the category at the lowest level in the line.

The lines in Category 4 (AI ethics), which correspond to leaves in a tree structure, are associated with the type of interaction. The type of interaction is determined by a combination of Interaction Start Type and Interaction End Type.

AI ethics models are used to determine from which requirements in the guidelines individual AI ethics characteristics are derived. The following interaction correspondence table is used to identify risks.

Table 3-3 AI ethics model (excerpt)

Type of starting point of interaction	Type of end point of interaction	Whether to consider when extracting risk events	Whether to consider when extracting risk factors	Category 1	Category 2	Category 3	Category 4(AI Ethics Characteristics)	Summary	Risk event or risk factor example
				Human agency and oversight					
				Human agency and oversight	Fundamental rights				
				Human agency and oversight	Fundamental rights	Guarantee of basic human rights			
Outputs	Target person of judgment	Yes		Human agency and oversight	Fundamental rights	Guarantee of basic human rights	Maintenance of social trust	Trust of stakeholders is not compromised by using the AI system	Employment candidates determined by human resources employment AI shows an extreme bias towards males
Business users	Outputs		Yes	Human agency and oversight	Fundamental rights	Guarantee of basic human rights	Validity of evaluation implementation of people/organizations	Validity of deciding capabilities of people and organizations based on AI system outputs has been confirmed	Person in charge of employment decides the employment candidates without verifying the validity of AI results
				Diversity, non-discrimination and fairness					
				Diversity, non-discrimination and fairness	Unfair bias avoidance				
				Diversity, non-discrimination and fairness	Unfair bias avoidance	Validity of final decisions	–		
AI models	Inference results	Yes		Diversity, non-discrimination and fairness	Unfair bias avoidance	Validity of final decisions	Group fairness	Differences of inference results by AI are within the tolerance range between groups of protection attributes	In loan screening AI, female or black AI models are less likely to be approved in the decisions
AI models	Inference results	Yes		Diversity, non-discrimination and fairness	Unfair bias avoidance	Validity of final decisions	Individual fairness	A pair of individuals with different sensitive information but identical non-sensitive information receives the same treatment	In loan screening AI, two people with the same AI models with the exception of gender receive different screening results

3.3.2 Interaction correspondence table

The Interaction Correspondence Table is a table in which the leaf portion of the tree structure (the line described in Category 4 (AI ethical characteristic)) is extracted from the table of the AI ethics model. Table 3-4 shows a portion of the Interaction Equivalents table. The Interaction Mapping Table maps each AI ethical characteristic to one of the interaction types. For example, the first line in Table 3 -4 indicates that the AI ethical characteristic of "maintaining social trust" corresponds to the interaction from "output" to "judgment target person". An explanation of this AI ethical trait of "maintaining social trust" is provided in the "Overview" column. "Examples of AI ethical characteristics that are not satisfied" describes examples extracted by applying this method to specific use cases, which correspond to ethical risks. Table 3 shows an example in which "maintaining social credibility" in row 1 of Table -4 is not satisfied, stating that "gender bias exists in recruitment AI, making it difficult for women to become candidates for recruitment.". The risk extraction work assumes such an example specifically for the use case to be evaluated.

Each row in the Interaction Correspondence Table has a "Yes" in either the "Whether to consider when extracting risk events" or "Whether to consider when extracting risk factors" column. A line with a "Yes" in "Whether to consider when extracting risk events"

assumes a risk event that may occur in the interaction if the AI ethical characteristics are not satisfied. Interactions that assume risk events involve stakeholders in principle. On the other hand, "Whether to consider when extracting risk factors" assumes risk factors that cause risk events. Interactions that assume risk factors are mainly related to interactions involving components within AI systems and interactions involving training data.

3.3.3　Risk extraction procedure

Each interaction with the analysis sheet created in Step 2 will be considered for risk. First, extract the rows whose interaction start and end types of matches from the interaction mapping table. The AI ethical characteristic of the extracted row is the AI ethical characteristic that the interaction should satisfy. Depending on the type of interaction, there might be no corresponding row in the interaction correspondence table. In this case, the interaction is not subject to risk consideration. In addition, multiple AI ethical characteristics may correspond to one interaction type. In this case, the risk is considered for each AI ethical characteristic. For risk assessment, assume possible risks by referring to "Examples of risk events or risk factors" in the Interaction Response Table, and describe them in "AI ethical characteristics" and "Risk events" or "Risk factors" in the analysis sheet.

The results of the risk statement for the analysis sheet in Table 3-2 are shown in Table 3-5. The process of deriving Table 3-5 from Table 3-2 is described. Interaction 105 in Table 3-2 is the interaction where the AI model is the starting point, and the inference result is the ending point. From the interaction correspondence table in Table 3-4, we can see that there are two AI ethical characteristics corresponding to this interaction, "Group fairness" and "Individual fairness, " and that "Yes" is attached to "Whether to consider when extracting risk events " in both cases. Therefore, "Group fairness" and "Individual fairness" are listed under "AI ethical characteristics " in the relevant row of the analysis sheet, and conditions that are contrary to "Group fairness" and "Individual fairness" are considered as risk events and listed under "Risk events".

The evaluator assumes specific risk events and risk factors on a case-by-case basis based on the knowledge of the AI system to be evaluated. To support this hypothetical task, an example of a risk event or risk factor is provided in the Interaction Response Table. Also refer to the risk events and risk factors described in the application examples provided separately.

Examples of risk events or risk factors are derived from the use cases used to build the AI ethics model, and there may be risks other than those listed here. In addition, AI

ethical characteristics that have already been described or are not described in "Examples of risk events or risk factors " may be changed in future editions of the Practical Guide.

3.3.4　Risk management

Once the risk assessment is complete for all interactions in the analysis sheet, the identified risks are sorted out and the risk events are associated with the risk factors that cause them. A system diagram that visualizes the associated risk events and risk factors is called an analysis diagram. Figure 3-2 shows an example of an analytic diagram.

Relationships are easier to consider when you start with a risk event and follow the interaction from the outside to the inside of the AI system. For example, in Figure 3-2, for risk events that occur in the interactions involving the judgment target person, if each risk factor extracted in the interactions involving the components of the AI system could cause the risk event, the risk factor is associated with the risk event.

Risk events and risk factors do not always correspond one-to-one. A risk event may be the cause of another risk event, or a risk factor may be the cause of multiple risk events.

In these three steps, we can visualize how the ethical risks that can occur in AI systems are related to the relationships between AI system components and stakeholders.

Table 3-4. Part of the interaction mapping table

Interaction Start Type	Interaction End Type	Whether to consider when extracting events	Whether to consider when extracting factors	Category 4 (AI ethical characteristics)	Overview	Examples of risk events or risk factors
Output	Person subject to judgment	Yes		maintenance of social credibility	The use of AI systems does not damage the trust of stakeholders.	Hiring AI ' s AI-judged candidates are heavily male-dominated
business user	Output		Yes	Adequacy of conducting a person/organization assessment	The validity of judging the capabilities of people and organizations based on the output of AI systems has been confirmed.	Recruiters determine candidates without validating AI results
AI Model	inference result	Yes		group fairness	The difference in AI inference results between groups with protected attributes must be within an acceptable range.	In loan screening AI, AI makes decisions that are difficult for women and blacks to pass.
AI Model	inference result	Yes		individual fairness	Sensitive information is different, but non-sensitive information is handled similarly by a pair of identical individuals.	In the loan examination AI, two persons with the same attributes except for gender have different examination results.

Table 3-5 Analysis sheet after risk extraction

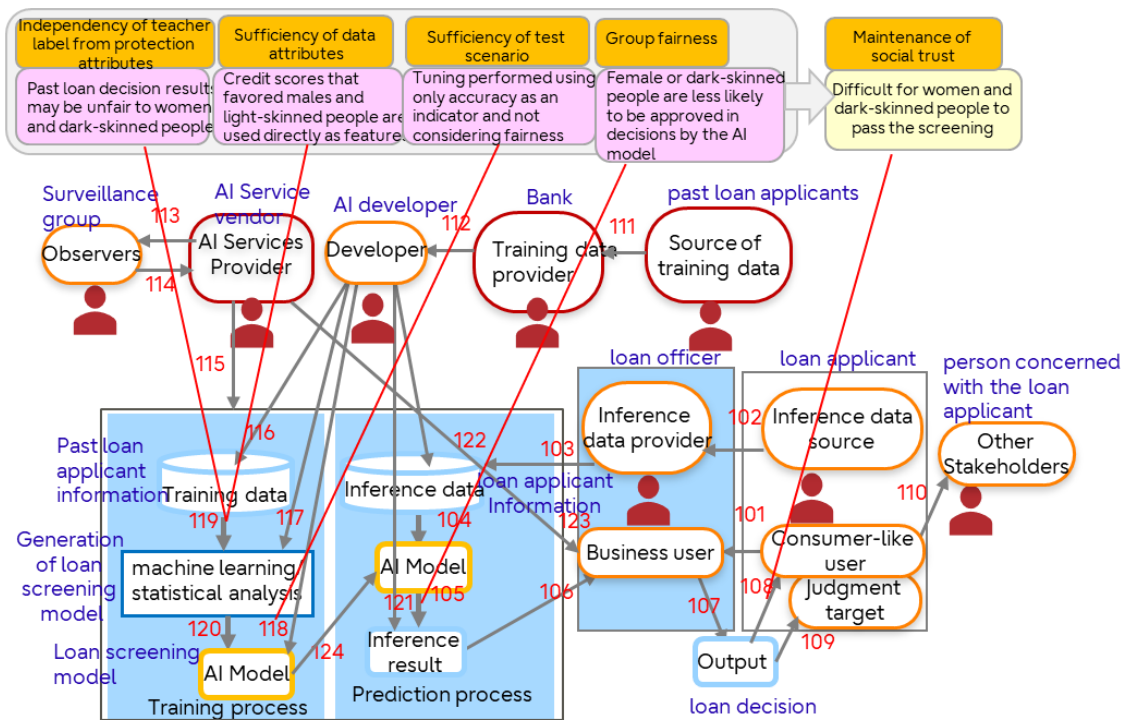| Interactio nID	Interaction Start Point		Interaction End Point		Interaction overview	AI ethical characteristics	Risk event	Risk factor
	Type	Name	Type	Name				
105	AI Model	loan examination model	inference result	result of examination	AI model outputs audit results	Group fairness	Female or dark-skinned people are less likely to be approved in decisions by the AI model.	
108	Output	loan decision	Target judgment	loan applicant	The applicant receives the decision	Maintenance of social trust	Difficult for women and dark-skinned people to pass the screening	
118	Developer	AI development vendor	AI Model	loan examination model	Develop a loan screening model	Sufficiency of test scenario		Tuning performed using only accuracy as an indicator and not considering fairness
119	training data	historical loan applicant information	machine learning and statistical analysis	loan examination model generation algorithm	Generate a loan award model	i) independence of teacher labels from protected attributes ii) Sufficiency of Data Attributes		i)Past loan decision results may be unfair to women and dark-skinned people ii) Credit scores that favored males and light-skinned people are used directly as features

Figure 3-2 Analysis Chart

4. Application example recidivism risk prediction

The result of applying AI ethics impact assessment to use cases of AI for recidivism risk prediction is shown below. This application example is a fictitious example constructed with reference to past cases and similar use cases.

4.1 AI System Overview

The recidivism risk prediction AI uses information about the defendant as input data to determine the defendant's recidivism risk on a 10-pont scale. Judges use the recidivism risk score to determine whether the defendant has been released on parole or not and to determine the sentence. Information about the accused includes criminal history, drug use, education, and employment level, but not race. The AI model learns information about past defendants, including whether they actually repeated offenses.

A possible ethical problem is that AI for predicting recidivism risk may make judgments biased by race. Legal systems such as criminal and judicial systems in various countries are not subject to evaluation under this application.

4.2 Use case overview

The use case summary sheet is shown in Table 4-1. The AI system assumes the configuration of a standard AI system that performs classification tasks. We also assume that other stakeholders are citizens with the same attributes as the defendant.

Table 4-1 Use case summary sheet for recidivism risk prediction AI

Major Item	Medium Item	Details
Industry		Public service/National defense and mandatory social security services
Purpose		Reference for use by judges to make decisions regarding parole and sentencing
Service	Service Overview	Determine the defendant's risk of recidivism on a 10-point scale.
	Availability of customization for each customer	N/A
	Requirements	Unknown
Usage case		Judges will refer to the recidivism risk score for decisions regarding pre-trial

		parole and determining whether a defendant will reoffend .
Stakeholders and their roles	AI service provider	Recidivism risk prediction AI developer vendor
	Developer	Recidivism risk prediction AI developer vendor
	Business users	Court judges
	Training data provider	Organizations that manage defendant data
	Source of training data	Past defendants
	Parties involved in training data acquisition	Unknown
	Inferential data providers	Court
	Inferential data source	Defendant
	Parties involved in the acquisition of inferential data	Unknown
	Consumer-like users	N/A
	Observers	Media, general public (e.g., people who read news in the media, and people who are interested in analyzing the data)
Stakeholders and their roles	Service UI/API provider	N/A
	Judgment target	Defendant
	Service authorizer	N/A
	Other stakeholders	People with the same attributes as the defendant
Presence/absence of a human-in-the-loop		N/A
Presence/absence of existing methods		Judgment passed down by the judge
AI Task	Task	Estimate the defendant's risk of recidivism on a 10-point scale
	Problem Classification	Classification issues
	Output	10-point recidivism risk score
	Technology	Unknown
	Breakdown of models required to perform AI tasks	Recidivism risk prediction model
	Update/additional training of training data	Unknown

	Real-time performance	Unknown
	Verification of AI tasks (consistency with decisions made by existing means)	Prior cases
Training Data	Breakdown, provider and source	Information about the defendant's past (includes criminal history, drug use, education and employment level; race is not included) (Provider and source: the authorities)
	Teacher labels	10-point recidivism risk score
	Protective attributes, attributes used for fairness verification and mitigation	Age, sex
	Presence/absence of personal information	Unknown
	Notes on data acquisition	Unknown
	Data Storage	Unknown
Inferential data	Breakdown, provider and source	Information about the defendant (not including race) (Provider: the court, source: the defendant)
	Protective attributes, attributes used for fairness verification and mitigation	Age, sex
	Presence/absence of personal information	Yes
	Notes on data acquisition	Unknown
AI System Output	Things to keep in mind during output	Unknown
Citations, reference articles		PAI Artificial Intelligence Incident Database #11 https://incidentdatabase.ai/cite/11/

4.3　System diagram

Select a pattern from the AI system pattern sheet based on the Use Case Summary sheet. Select Pattern 2 because there is an interaction in which business users (judges) receive inference results (recidivism risk scores), make judgments, and the judgment target (defendant) receives the output. The consumer-like user is also listed in the pattern 2, but since this use case uses "none" for the consumer-like user, we remove the consumer user and its associated interaction from the pattern drawing.

Enter the names "Judge", "Defendant", etc. in the person type (stakeholder) in the system diagram. It also lists the names of the AI system components. The name of machine learning and statistical analysis is "recidivism risk prediction model generation processing". The inference result will be the repeat offense risk score and the output will be the judge's decision. The resulting system diagram is shown in Figure 4-1.
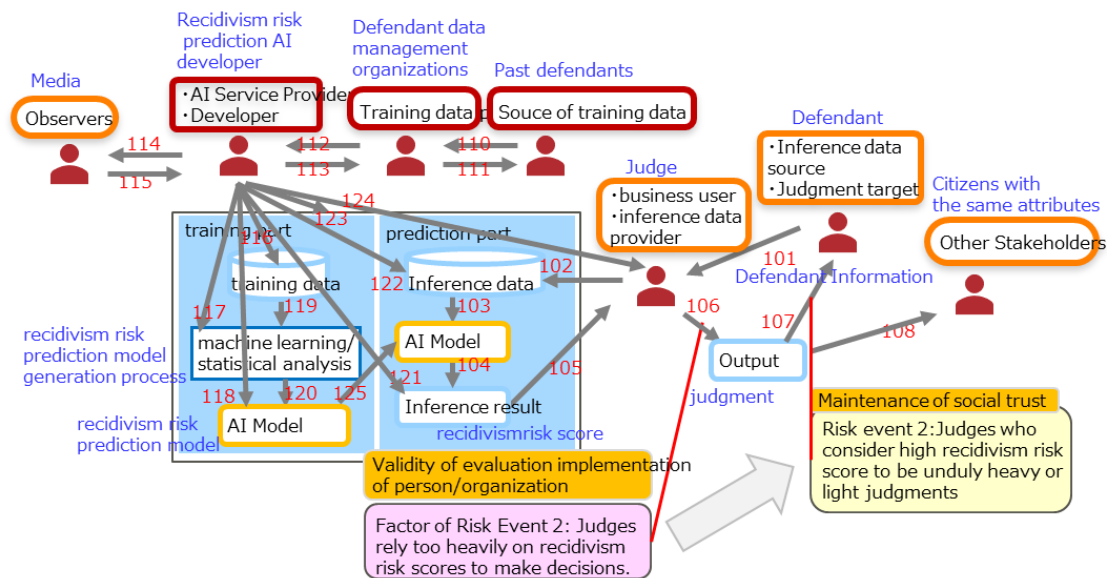
Figure 4-1 System diagram of recidivism risk prediction AI

4.4　Analysis result

Figure 4-2 shows an analysis chart that identifies risks and describes them in the system chart. Table 4-2 shows the locations where risk events and risk factors are described in the analysis sheet.

The arrows in Figure 4-2 represent interactions, and the boxes that extend from the arrows represent risk events or factors in each interaction. In the figure, risk factors such as "past defendant data may be biased toward black people " and" black residential areas are factors that increase the repeat offense risk score " are shown in the two interactions of the training department of the AI for predicting repeat offense risk, and the risk events that they cause are shown as "In spite of the fact that they did not actually repeat offenses, there is a possibility that blacks are misjudged to have a higher risk of repeat offenses than whites.". This risk event also leads to potential risk events in the interaction between the media and the vendor. In the interaction in which judges make judgments, risk factors such as judges overly rely on the recidivism risk score are assumed, leading to possible risk events in the interaction between judgments and the accused.
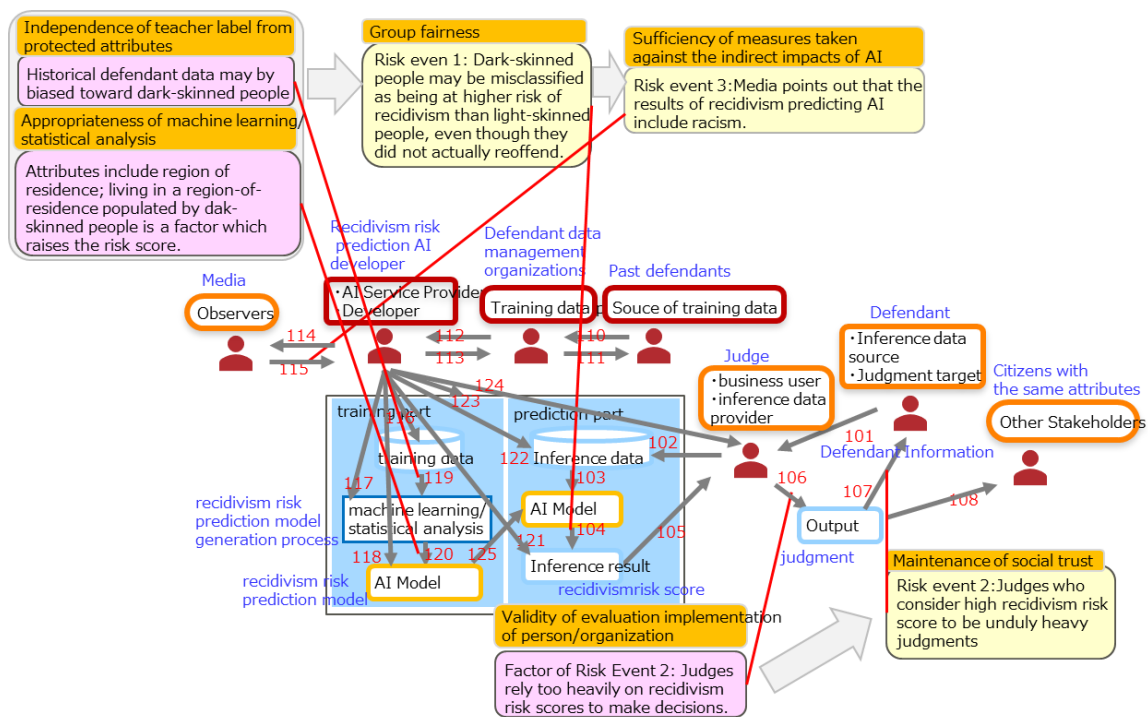
Figure 4-2 Analysis chart of repeat offense risk prediction AI

Table 4-2 Analysis sheet of AI for predicting repeat offense risk

Interaction ID	Interaction Start Point		Interaction End Point		Interaction overview	AI ethical characteristics	risk event	risk factor
	Type	Name	Type	Name				
104	AI model	repeat offense risk prediction model	Inference result	Recidivism risk score	AI model outputs recidivism risk score	Group fairness	Dark-skinned people may be misclassified as being at higher risk of recidivism than light-skinned people, even though they did not actually reoffend	
106	Business user	Judge	Output	judgment	The judge decides the decision.	Validity of evaluation implementation of person/organization		Judges rely too heavily on recidivism risk scores to make decisions.

107	Output	judgment	Person subject to judgment	defendant	be subject to judgment	Maintenance of social trust	Judges who consider the high recidivism risk score to be unduly heavy judgments.	
115	Observer	Media	AI Service Provider	AI development vendor	report problems with AI systems	Sufficiency of measures against the indirect effects of AI	Media points out that the results of recidivism predicting AI include racism.	
119	Training data	Past defendant information	Machine learning or statistical analysis	Recidivism risk prediction model generation Process	input training data into machine learning and statistical analysis	Independence of teacher label from protection attribute		Historical defendant data may by biased toward dark-skinned people.
120	machine learning or statistical analysis	Recidivism prediction model generation process	AI Model	Recidivism risk prediction model	Generate an AI model	Appropriateness of machine learning or statistical analysis		Attributes include region of residence; living in a region-of-residence populated by dark-skinned people is a factor which raises the risk score.

5. Application Example Recruitment AI

This section presents the results of applying the AI Ethics Impact Assessment to AI use cases for human resource recruitment. This application example is a fictitious example constructed with reference to past cases and similar use cases.

5.1 AI System Overview

In the recruitment process, we consider an AI system that screens candidates for job interviews from the text of job applicants' resumes. The AI model is assumed to generate the resume and employment results of past job seekers as training data. Résumé does not indicate race or gender.

5.2 Use case overview

Table 5-1 Use Case Summary Sheet for Recruitment AI

Major Item	Medium Item	Details
Industry		Management and support services industry
Purpose		Narrowing down the number of candidates for employment interviews
Service	Service Overview	AI tool that automatically categorizes the resumes of your job candidates and selects the most promising candidates
	Availability of customization for each customer	N/A
	Requirements	N/A
Usage case		Automatically categorizes the resumes of job candidates and selects the most promising candidates
Stakeholders and their roles	AI service provider	Development vendors for a recruitment AI
	Developer	Development vendors for a recruitment AI
	Business Users	Recruiter for a company that uses the recruitment AI
	Training data provider	Companies that use the recruitment AI
	Source of training data	Companies that use the recruitment AI
	Parties involved in training data acquisition	Past Job applicants
	Inferential data providers	Job applicants

	Inferential data source	Job applicants
	Parties involved in the acquisition of inferential data	Unknown
	Consumer-like users	Job applicants
	Observers	Unknown
	Service UI/API provider	Job applicants
	Judgment target	Job applicants
	Service Authorizer	Unknown
	Other Stakeholders	Unknown
Presence/absence of a human-in-the-loop		N/A
Presence/absence of existing methods		Scoring job applicants based on resume content
AI Task	Task	Determining the level of a potential candidacy of an applicant based on resume data
	Problem Classification	Classification issues
	Output label	Score on a 5-point scale
	Technology	N/A
	Breakdown of models required to perform AI tasks	N/A
	Update/additional training of training data	N/A
	Real-time performance	N/A
	Verification of AI tasks (consistency with decisions made by existing means)	Decision by the hiring manager
Training Data	Breakdown, provider and source	Past resumes (Provider: company using the recruitment AI, Source: Past job applicants)
	Teacher labels	Score on a 5-point scale
	Protective attributes, attributes used for fairness verification and mitigation	Sex
	Presence/absence of personal information	Unknown
	Notes on data acquisition	Unknown
	Data Storage	Unknown
Inferential data	Breakdown, provider and source	Resume (Provider/source: job applicant)

	Protective attributes, attributes used for fairness verification and mitigation	Sex
	Presence/absence of personal information	Yes
	Notes on data acquisition	Unknown
AI System Output	Things to keep in mind during output	Unknown
Citations, reference articles		PAI incident database #37 https://incidentdatabase.ai/ cite/37

5.3   System diagram

Select Pattern 2 from the AI System Pattern Sheet based on the Use Case Summary Sheet. The reason for choosing Pattern 2 is that there is an interaction in which business users (recruiters) receive inference results (recruitment scores), make judgments, and the target audience (job seekers) receives the output. The consumer user is also listed in the pattern 2, but since the consumer user is "none" in this use case, we remove the consumer user and its associated interaction from the pattern drawing. Similarly, since we have no Observer, we remove the Observer and its associated interaction from the drawing.

In the system diagram, indicate the name of the human type (stakeholder), such as "Recruiter" or "Recruiter AI Development Vendor.". It also lists the names of the AI system components. The generated system diagram is shown in Figure 5-1.
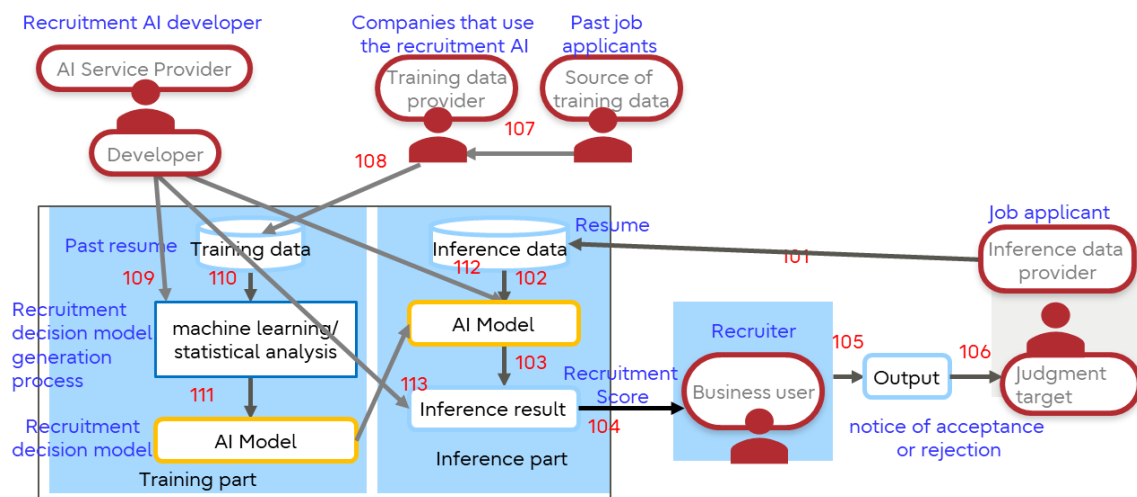

Figure 5-1 System Diagram of Adopted AI

5.4   Analysis result

Figure 5-2 shows the analysis chart after risk extraction, and Table 5-2 shows the locations where risk events or risk factors are listed from the analysis sheet. In the figure, we can see that the risk factors that can occur in the interaction between the training data of the AI system and the AI model lead to a risk event in the interaction from the AI model to the inference result (1) "Candidates for employment judged by the AI model are biased toward men, " which in turn leads to a risk event in the interaction from the acceptance/rejection notification to job seekers (2)" Candidates for employment interviews are biased toward men. ".
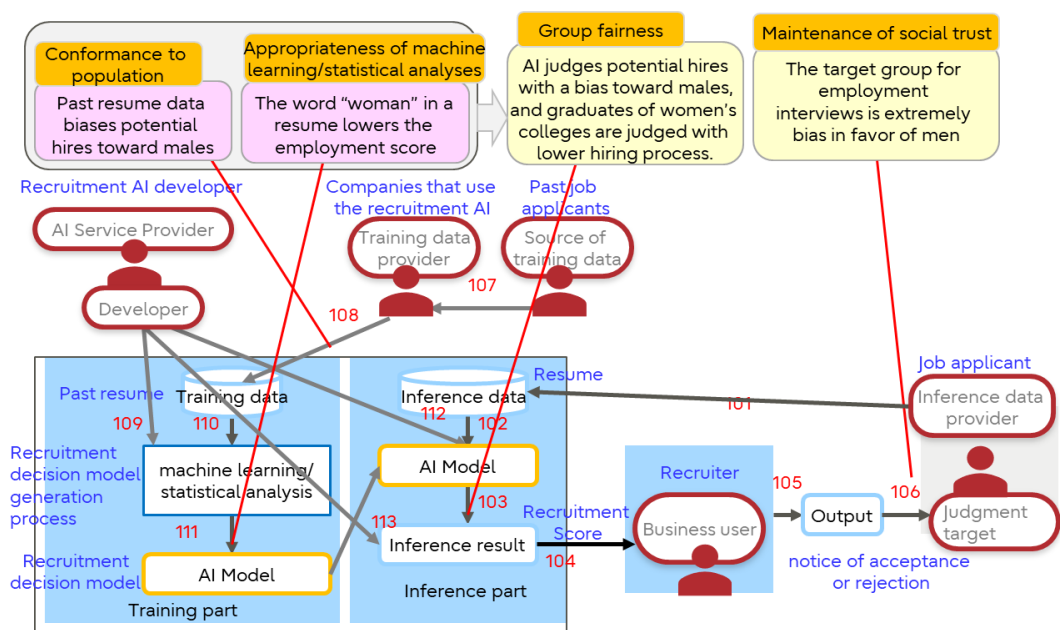

Figure 5-2 Analysis chart of AI employed

Table 5-2 Some analysis sheets for recruitment AI

Interac tion ID	Interaction Start Point		Interaction End point		Interaction overview	AI ethical characteristics	risk event	risk factor
	Type	Name	Type	Name				
103	AI Mod el	Recruit ment decisio n model	inferen ce result	Recruitm ent Score	AI models determine recruitment scores	Group fairness	AI judges potential hires with a bias toward males, and graduates of women's colleges are judged with lower hiring process.	
106	Outp ut	notice of accept ance or rejectio n	Judgme nt target	job applicant	Job applicant receives a notice of acceptance or rejection	Maintenance of social trust	The target group for employment interviews is extremely bias in favor of men.	
108	Traini ng data provi der	Recruit ment Inform ation Manag er	Trainin g data	Curriculu m vitae and employm ent results of past job seekers	provide training data	Conformance to population		Past resume data biases potential hires toward males.
111	mach ine learni ng/ statis tical analy sis	Recruit ment decisio n model genera tion proces s	AI Model	adoption decision model	Generate a recruitment decision model	Appropriatenes s of machine learning /statistical analysis		The word "woman" in a resume lowers the employment score.

Conclusion

This manual describes the practical procedures of the AI Ethics Impact Assessment Method, which extracts ethical risks that may occur in AI systems in accordance with the AI Ethics Guidelines. In this manual, we introduce an application example and show that this method can extract where and how ethical risks can occur in AI systems.

The AI Ethics Impact Assessment Practice Guide provides this manual and the AI Ethics Model on which this method is based, as well as sheets and drawings for analysis. It also provides examples for use cases not covered in this book to help readers implement this approach.

In the future, this method will be improved by incorporating the knowledge and viewpoints of various stakeholders, aiming to provide trusted AI.

References

[1] https://incidentdatabase.ai/cite/16

[2] https://incidentdatabase.ai/cite/37

[3] https://www.eismd.eu/wp-content/uploads/2019/11/AI4People%E2%80%99s-Ethical-Framework-for-a-Good-AI-Society_compressed.pdf

[4] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[5] https://oecd.ai/en/ai-principles

[6] https://ethicsinaction.ieee.org/

[7] https://www.soumu.go.jp/main_content/000637097.pdf

[8] https://www8.cao.go.jp/cstp/aigensoku.pdf

[9] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN

[10] https://partnershiponai.org/

[11] https://incidentdatabase.ai/

[12] https://www.iso.org/standard/77610.html

[13] https://www.jil.go.jp/kokunai/statistics/databook/2019/03/d2019_T3-01A.pdf