AI Ethics Impact Assessment based on Requirement Engineering

Izumi Nitta
Fujitsu Limited
Kawasaki, Japan
nitta.izumi@fujitsu.com

Kyoko Ohashi
Fujitsu Limited
Kawasaki, Japan
ohashi.kyoko@fujitsu.com

Satoko Shiga
Fujitsu Limited
Kawasaki, Japan
shiga.satoko@fujitsu.com

Sachiko Onodera
Fujitsu Limited
Kawasaki, Japan
sachiko@fujitsu.com

***Abstract* —This paper proposes a methodology for evaluating the ethical impact of artificial intelligence (AI) systems on people and society based on AI ethics guidelines. The ethical impact of AI has been recognized as a social issue, and countries and organizations have formulated principles and guidelines on AI ethics, and laws and regulations will be enforced in Europe. Because these principles and guidelines are written in terms of philosophy and law, AI service providers, developers, and business users have the challenge of how they should practice the principles and guidelines to their AI systems. To address this challenge, we first analyzed cases of ethical problems caused by AI in the past and assumed that ethical problems could be linked to interactions between components of AI systems and stakeholders related to such systems. On the basis of this assumption , we then developed a methodology to comprehensively extract the ethical risks that an AI system poses. This methodology consists of two approaches. The first approach is to develop an AI ethics model that embodies ethics guidelines as necessary requirements for ethical AI systems and correlates these requirements with interactions. The second approach is an impact assessment process that uses the AI ethics models to extract ethical risks for individual AI systems. In this paper, we discuss the details of this methodology and show the results of an initial validation to verify the above assumption and the ease of the impact assessment process.**

Keywords — AI ethics, AI governance, responsible AI, impact assessment, risk-based approach

I. INTRODUCTION

The ethical impact of artificial intelligence (AI) has become recognized as a social issue, and ethical principles and guidelines that provide the basic requirements for the spread of responsible AI are being developed [1] [2] [3] [4] [5] [6]. In Europe, the European Commission published a draft AI regulation called the Artificial Intelligence Act [7], which categorizes the manipulation of people's subconscious, the use of social scoring, and remote biometrics for law-enforcement purposes in public spaces as prohibited AI systems. It also lists the uses of AI in personal biometrics and classification and its application to critical infrastructure as "high-risk AI" and imposes a number of requirements for its use in these fields with significant fines for violations. In the United States, a bill known as the “Facial Recognition and Biometric Technology Moratorium Act” [8], which prohibits federal officials from using facial-recognition technology, has been proposed. The city of San Francisco has also banned the use of facial-recognition technology by police.

The movement from ethical principles to practice has become active. To comply with ethics guidelines, the development of technology to address ethical issues related to the fairness of AI is moving forward. Several standards have been defined regarding fairness. For example, for sensitive attributes, such as race and gender for which fairness is being considered, there is one approach that equalizes the probability of the potential outputs provided by AI (for example, for recruitment AI, the hiring and rejection of candidates), and another that equalizes the probability of AI output being used in correct data. Machine learning algorithms that follow these various standards are being extensively studied [9]. A technique [10] for taking into account fairness in relation to multiple sensitive attributes and a concept for an interface design for end users to judge and deal with fairness in AI [11] have also been proposed.

Companies are also beginning to implement principles into practice. Many companies have their own AI ethics guidelines. Companies that are making an effort to follow their guidelines have published practical examples and assessment tools, such as open toolkit for fairness and transparency in implementing the guidelines [22] and assessment tools [23].

In this paper, we define "AI ethical impact" as the ethical impact of the use of AI systems on stakeholders or the ethical impact of stakeholders on AI systems and other stakeholders. The purpose of our research was to provide a methodology for AI service providers, AI developers, and business users who are non-AI ethics experts to evaluate the ethical impact of their AI systems on the basis of AI ethics guidelines and recognize where and what risks may occur in AI systems.

Since AI principles and guidelines are written in the language of law and philosophy, for non-AI ethics experts, reading and comprehending these principles and guidelines, and putting them into practice in their AI systems can be burdensome. Even when using the various impact assessment frameworks discussed above, it is likely that knowledge of AI ethics and past case analysis will be required to apply them to AI systems.

To address this issue, we assumed that systematizing ethical problems caused by AI in the past would enable conducting impact assessment procedurally for various AI use cases. From a survey of past AI ethical incident cases, we assumed that ethical risks are mapped to the interactions between the components of an AI system and the stakeholders directly or indirectly involved with the system.

On the basis of this assumption, we propose a methodology called AI Ethics Impact Assessment for comprehensively identifying ethical problems caused by AI by associating the requirements for responsible AI described in the ethics guidelines with the interactions that appear in AI systems. The methodology consists of two approaches based on the requirements engineering approach.

Approach 1: Building an AI ethics model. Embody written ethical guidelines as requirements necessary for AI systems to be ethical, and map these requirements to interactions. Ethical risk can be treated as a situation contrary to the requirements associated with the interaction.

Approach 2: Impact assessment process. The process of extracting ethical risks for individual AI systems using the AI ethics model is presented.

Using the AI Ethics Impact Assessment, AI developers, providers, and users without expertise in ethical guidelines can assess ethical risks at each stage of the AI lifecycle, from planning to development, operation, and retirement. We discuss the details of the two approaches of AI Ethics Impact Assessment, i.e., AI ethics model and ethics impact assessment process, our initial validation on the above assumption, the ease of the impact assessment process, and the effectiveness and issues concerning the AI Ethics Impact Assessment.

The remainder of this paper is structured as follows. Section II introduces related work, Section III describes the concept for systematizing and evaluating ethical impact, Section IV describes the AI Ethics Impact Assessment, i.e., AI ethical impact assessment using the AI ethics model, Section V describes the initial validation of our methodology, and Section VI presents conclusions and future issues.

II. Related Work

In this section, we introduce AI impact assessment and related work. The Canadian government has issued guidance on Automated Decision-Making [12] and provided a tool to assess the impact of algorithms on decision-making systems. As a method to implementing ethically aligned AI in software engineering, ECCOLA [24] offers 21 cards in 8 themes based on IEEE's (Institute of Electrical and Electronics Engineers) Ethically Aligned Design guidelines [4] and EU AI HLEG's (High-Level Expert Group on Artificial Intelligence) Trustworthy AI guidelines [2]. The Algorithm Impact Assessment toolkit [13] from the Ada Lovelace Institute was developed to assess the impact of AI on medical imaging. Floridi et al. provided guidance on the conformity assessment of AI systems on the basis of the European Artificial Intelligence Act proposal [14]. NIST (National Institute of Standards and Technology) also categorizes AI biases and reports challenges and guidance [15].

In the field of software engineering, Borg et al. [25] demonstrated the applicability of ALTAI (Assessment List for Trustworthy AI) [27] to ongoing development projects and showed that the scope of its evaluation was not only related to developers and providers but also to social and environmental issues.

Johnson et al. surveyed research on ethical software practices in the process of developing, evaluating, and maintaining data-driven software [26].

The risk chain model for assessing ethical risks has also been proposed [16]. This model provides a framework for AI service providers to consider risk assessment and control of their AI services. The relationship between risk scenarios and risk factors in an AI system is visualized, and the examination of risk control is enabled on the basis of this visualization.

III. Concept for Systematizing and Evaluating Ethical Impact

In this section, we first explain the terminology used in this paper then explain the concept for systematizing and evaluating ethical impact.

A. Terminology

- AI ethical risk: A risk arising from ethical issues caused by using AI systems. AI ethical risks include not only negative impact on AI systems and their stakeholders but also positive impact.
- Risk events: An AI ethical risk that affect stakeholders or risks caused by stakeholders.
- Risk factors: AI ethical risks that cause risk events. Risk events may cause other risk events.
- AI ethical impact: the ethical impact of the use of AI systems on stakeholders or the ethical impact of stakeholders on AI systems and other stakeholders.

B. Initial analysis of ethical imapct using past cases

We thought that if we could systematize how ethical issues arise from AI in several patterns, it would be helpful for AI developers, providers, and other stakeholders who understand the specifications and use cases of their AI systems to conduct impact assessments. Accordingly, we analyzed past ethical cases and consider how to systematize these cases.

As an initial analysis, we focused on the fairness issue and examined the ethical incidents on credit card applications, recruitment AI, and recidivism risk prediction from the AI Incident database [20]. In each case, we clarified which stakeholders were affected by the fairness issue then identified the causes of the issue, referring to a survey on fairness-aware machine learning [9]. Next, we identified the stakeholders and components of the AI system in each case and created a diagram to visualize their relationships. We then visualized in the diagram the fairness issue and its causes. In this analysis, we used 2 use cases (loan screening AI and mental healthcare AI) from the ISO (International Organization for Standardization) /IEC (International Electrotechnical Commission) AI use case [19] and 3 use cases from interviews with AI projects related to image recognition to sort out the types of stakeholders and AI system components. Table I shows the types of components of an AI system and stakeholders.

We investigated examples of past ethical problems and whether there were common patterns in the way the ethical problems occurred through the following procedure.

- The use cases in which a past ethical issue occurred are first represented by the components and stakeholders of an AI system and the interactions between them. The visualization

of the components, stakeholders, and interactions is called a system diagram.

- The ethical issue is then mapped on a system diagram.

TABLE I. DEFINITION OF AI SYSTEM COMPONENTS AND STAKEHOLDERS

Component or data	Description
Pre-processing	Processing input data to generate training or inference data
Machine learning and statistical analysis	Algorithms using machine learning or statistical analysis to create AI models from training data
AI model	Inferring based on inference data and outputting of inference results
Post-processing	Processing inference result from an AI model to generate output of an AI system
Service UI/API	UI or API of an AI system
Apparatus	A machine or a device used for input and output to a computer system
AI system	System consisting of a training unit and a prediction unit
Training unit	A processing unit that machine learning or statistical analysis generates an AI model from training data
Inference unit	A processing unit that an AI model infers and outputs an inference result by
Original training data	Training data before pre-processing
Training data	Input to machine learning or statistical analysis
Original inference data	Inference data before pre-processing
Inference data	Input to an AI model
Inference result	Output from an AI model
Final decision	Judgment by a person based on an inference results
Output	Output from an AI system by post-processing based on inference results

Stakeholder	Description
AI service provider	People or organizations that operate and provide an AI system
Developer	Developing the AI system
Business users	People or organizations that use AI for their business
Consumer-like users	Users of an AI system or an AI service who are not business users
Training data provider	A person who provides the original data to create training data
Training data source	A person whose data is provided to a training data provider
Parties involved in training data acquisition	People, organizations, or systems directly or indirectly involved in training data acquisition
Inference data providers	A person who provides input data to create inference data
Inference data source	A person whore data is provided to an inference data provider
Parties involved in the acquisition of inferential data	People, organizations, or systems directly or indirectly involved in inference data acquisition
Observers	People or organizations monitoring AI systems or the AI services
Service UI/API provider	People or organizations that provide input/output API for an AI system
Judgment target	People or organizations to be judged or evaluated by an AI system
Service authorizer	People or organizations that authorize AI services
Other stakeholders	People or organizations who indirectly affected by an AI system or stakeholders, or who indirectly affect on an AI system or stakeholders

Take loan screening AI as an example. The loan screening AI determines whether a loan application is approved on the basis of the attributes, transaction history, and credit score of the loan applicant. The loan officer makes a final decision on the basis of the AI results and responds to the applicant. The teacher label of training data is based on the repayment performance of past loan applicants. It has been pointed out that the results of loan screening AI are biased regarding gender.

Figure 1 shows where the ethical risks of loan screening AI appear. The diagram shows a simplified AI system, with loan officers and loan applicants as stakeholders. Arrows connecting the AI system's components and stakeholders represent interactions.

The ethical risks that can be mapped to each of the following four interactions are shown:

1. From training data to AI models: Gender bias is learned from training data and reflected in the AI model.

2. From AI model to output: There is gender bias in the results obtained using the AI model.

3. From loan officer to loan applicant: The loan officer makes final decisions that heavily depend on the AI results.

4. From loan applicant to loan officer: The loan applicant makes an objection to the loan officer.

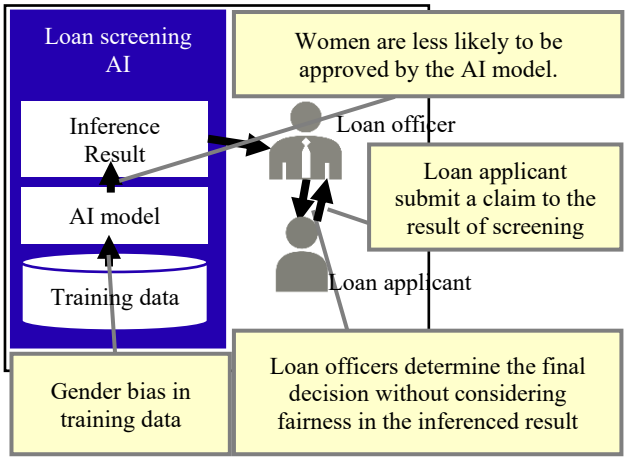

Fig. 1: Example of ethical risks on loan screening AI

C. *Where do AI ethical risks appear in AI system?*

We conducted an analysis of past ethical issues by AI and developed the following assumption.

Assumption: Ethical risks can be related to the interactions between the components of an AI system and the stakeholders that are directly or indirectly involved with the AI system.

The analysis of the cases shows that the components of the AI system and stakeholders can be patterned in accordance with their roles. It is considered that by patterning in this manner, the above assumption can be applied to various AI use cases.

IV. PROPOSED METHODOLOGY

This section describes the AI Ethics Impact Assessment. On the basis of the assumption on ethical risks of AI derived in the previous section, this methodology is used to analyze

where and how the ethical risks may occur by 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 methodology can be carried out procedurally if the analyst of the methodology has knowledge about the AI system to be evaluated and enables reliable evaluations in a realistic time frame.

A. AI ethics model

We first describe an AI ethics model required to extract AI ethical risks. An AI ethics model comprehensively indicates the characteristics that an ethical AI system should have. Individual characteristics included in an AI ethics model are called AI ethical characteristics. An AI ethics model organizes and defines its characteristics hierarchically.

In this context, an AI system is an IT system that uses technologies related to AI. It is assumed to include not only AI models generated using machine learning and statistical analysis but also to be a system in which processing components required for IT systems are combined. Stakeholders directly or indirectly involved in AI systems are also included in the AI ethics model.

The range of AI ethical characteristics of an AI ethics model should be within the range indicated in the ethical guidelines. We believe that ethical guidelines can be embodied in requirements that AI systems must have. This is based on the concept of requirements engineering. The user experience (UX) quality model [17] combines a top-down approach with a bottom-up approach to comprehensively collect and organize UX quality characteristics. Specifically, the UX quality model is expressed in four levels: the upper two levels are expressed using the definition of SQuaRE (Systems and software Quality Requirements and Evaluation) [18], and the lower two levels are embodied using the results of UX evaluation by the user. Through this approach, a practical model can be constructed by comprehensively expressing quality characteristics in the upper levels and preventing excessive embodiment in the lower levels. We use this concept to construct an AI ethics model.

An AI ethics model is also expressed in a four-level structure. In the top-down approach, the text of the ethical guidelines is structured in the upper two levels. In the bottom-up approach, regarding ethical problems obtained from the analysis of AI system use cases [19] and AI system cases that caused problems in the past from the AI Incident database [20] published by Partnership on AI (a non-profit organization that promotes AI), the characteristics that can be obtained by solving the problems are extracted as AI ethical characteristics. These characteristics are embodied in the lower two levels, and guidelines and AI ethical characteristics are made to correspond to each other.

The type of interaction (for example, AI ethical characteristics related to AI fairness include handling interactions from AI outputs to business users, etc.) to satisfy each AI ethical characteristic is extracted and correlated. In this case, several interactions may correspond to one AI ethical characteristic, and one interaction may correspond to several AI ethical characteristics. This prevents interactions and AI ethical characteristics from becoming incompatible. The above approaches provide the following advantages:

- By basing the approaches on ethical guidelines, it is possible to ensure completeness in the sense of compliance with these guidelines.
- Because interactions in AI systems correspond to AI ethical characteristics and guidelines correspond to the upper levels of AI ethical characteristics, it is possible to clarify to which part of the guidelines an interaction should correspond. This makes it easier to consider measures such as technical solutions or operational measures.

Figure 2 shows an overview of an AI ethics model. The first layer of the AI ethics model corresponds to the seven requirements (“Human agency and oversight”, “Technical robustness and safety”, “Privacy and data governance”, “Transparency”, “Diversity, non-discrimination, and fairness”, “Societal and environmental well-being”, “Accountability”) from the EU AI HLEG’s ethics guidelines for Trustworthy AI (Trustworthy AI) [2]. The second and lower layers are structured from Assessment Lists of the Trustworthy AI (ALTAI) [27]. Each node of the fourth layer is associated with an interaction defined in the previous section III

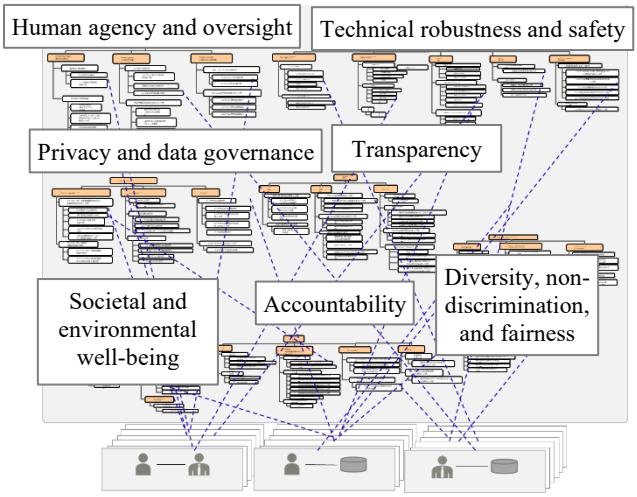

Fig. 2: Overview of the AI ethics model

We represent AI ethics model in table format shown in TABLE II. “Category 1” to “Category 4” columns of the table correspond to the AI ethics model in which Trustworthy AI and ALTAI are structured and embodied in four layers. “Category 4” is the AI ethical characteristic.

Each AI ethical characteristic is associated with the type of interaction (combination of the “type of start point of interaction” and “type of end point of interaction”). Multiple types of interactions may correspond to one AI ethical characteristic. “Summary” is a description of AI ethical characteristics. “Risk event or risk factor example” describes examples of risks extracted from analysis of past incident cases.

For example, an AI ethical characteristic called “Maintenance of social trust” is associated with an interaction between a business user and judgment target. A situation in which the target system violates this requirement constitutes an ethical risk.

TABLE II. EXCERPT OF THE AI ETHICS MODEL

Type of start point of interaction	Type of end point of interaction	Category 1	Category 2	Category 3	Category 4 (AI Ethics Characteristics)	Summary	Risk event or risk factor example
Business user	Judgment target	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	Loan officer makes final decision to the loan applicant without validating fairness of the AI result
Business user	Judgment target	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	Recruitment staff decides the final decision to the job applicant without verifying the validity of the AI result
AI model	Inference result	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 recruiting AI, female or black are less likely to be adopted
AI model	Inference result	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 persons with same attribues but gender had different results from AI

In the example of the loan screening AI mentioned in the previous section, this requirement allows the loan officer to make final decisions that heavily depend on the AI results to be extracted as an ethical risk. With this approach, it is possible to conduct impact assessments for various use cases by creating an AI ethics model once per guideline.

B. *Risk-extraction procedure*

We describe how the AI Ethical Impact Assessment is conducted along the overall diagram shown in Figure 3 by using the constructed AI ethics model. The procedure consists of three steps.

1) Step 1: We first create an AI system diagram based on AI specifications and use case information. A system diagram shows the arrangement of stakeholders related to the components (data, AI model, etc.) of an AI system and their interactions with arrows.

2) Step 2: Next, AI ethical characteristics corresponding to all interactions in the system diagram are extracted using the AI ethics model. This step can be mechanically carried out using the interaction extracted in Step 1 as an input.

3) Step 3: Finally, a situation contrary to each extracted AI ethical characteristic is extracted as a risk. This is done manually by the analyst for each use case. To facilitate this task, the AI ethics model has a description of for each AI ethical characteristic and an example risk extracted from the analysis of past cases.

The following is an example of loan screening AI. Figure 4 is the system diagram created in step 1. The arrows in the system diagram are the interactions. The number attached to the arrow is the ID of the interaction. Table III lists the AI ethical characteristics corresponding to interactions and their descriptions generated in step 2. Consider an example of extracting risks for interaction ID 105 in Table III. AI ethics characteristics linked to ID 105 are "Group fairness". The “summary” of “Group fairness” is that the "Difference of interference results by AI are within the tolerance range between groups of protection attributes". The statement describes an ethical situation, and the risk would be a situation that violates the statement. ID 105 is an interaction from the AI model to the inference result, and the risk that can occur in this interaction is extracted as "the result is unfair according to gender or race.”

A practical guide for conducting this impact assessment process consisting of procedures, AI ethics models, and analysis sheets has been published [21]. Case studies using the impact assessment process for certain use cases have also been published.

V. VERIFICATION OF PROPOSED METHODOLOGY

We conducted an initial validation by means of a questionnaire survey to determine whether it is possible to extract the risk of AI ethics in accordance with the AI Ethics Impact Assessment. The details of the questionnaire are as follows:

1) Eight participants with experience in case studies of AI ethics participated. Three of them were researchers involved in AI fairness.

2) Participants downloaded the questionnaire file and responded to the items. The response period was set at one week.

3) The AI Ethics Impact Assessment was conducted on nine cases, and the results of the analysis were visualized in an analysis chart on the basis of the system diagram. In each case, one of the risk events was left blank, and the participants had to respond with a statement that describes the risk extracted using the related AI ethical characteristic and its description.

The participants were presented with nine use cases of the AI Ethics Impact Assessment and asked to respond to Questions 1 and 2, which are described later. The use cases were selected from the AI incident database [20]. Table IV presents the details of these use cases. The column "ID" in the table indicates the index of the AI Incident database.

Participants read the explanatory text on the objectives and usage scene of the AI service and the configuration of the AI system for each use case and responded to the following steps.

A. Preliminary prepaction

Participants read instructions on how to define AI system components and stakeholders and how to view system diagrams, which are covered in the AI Ethics Impact Assessment.

B. Question 1: Validity of our assumption

Objectives: For each risk presented in the analysis results, examine the validity of the interactions associated with that risk and test the validity of our assumption.

Questions: Participants responded to the following questions for each of the risks listed in the system diagram of the analysis results. In the example shown in Figure 5, three ethical risks are described, each of which is to be responded to.

- Q1-1: Are the interactions linking the risk valid? (Valid/Valid but there are other relevant interactions/Not valid and there are other relevant interactions/No relevant interactions)
- Q1-2: If you select “Valid, but there are other relevant interactions”, provide the ID of any other interactions that may apply. (free format)
- Q1-3: Reasons for selecting the response from Q1-2. (free format)
- Q1-4: Are there any interactions that are not shown in the system diagram that may pose risks? (free format)

C. Question 2: Ease of risk extraction

Objectives: To examine the ease of risk extraction from the AI ethical characteristics and corresponding guideline text linked to interactions. The text is from assessment list of Trustworthy AI that was used to derive corresponding AI ethical characteristics.

Questions: In the system diagram shown in Question 1, one of the ethical risks is blank. Q2-1 asks to fill in this blank from the explanatory text of the corresponding guideline.

- Q2-1: On the basis of the text of the guideline, assume the concrete risks and describe them in text (free format).
- Q2-2: Was it easy to respond to Q2-1? (Yes/No)

Table V shows an example of the questions and answers about the recruitment AI case.

D. Results of Question 1

For each use case, as an indicator of the validity of the association between risk and interaction, we defined the scores for each option in Q1-1 as follows:

- (Valid) = (Total number of responses with “Valid”)/(Number of risks) * (Number of participants)
- (Valid but other interactions) = (Total number of responses with “Valid, but there are other relevant interactions”)/(Number of risks) * (Number of participants)
- (Not Valid and other interactions) = (Total number of responses with “Not valid, and there are other relevant interactions”)/(Number of risks) * (Number of participants)
- (No interactions) = (Total number of responses with “No interactions”)/(Number of risks) * (Number of participants)

Figure 6 shows the scores for each use cases. The combined scores of "Valid" and "Valid but other interactions" exceeded 0.7. We consider this as validating the association between risk and interaction.

E. Results of Question 2

We defined the following score on Q2-2:

- Q2-2: (Ease of risk assumption) = (Number of responses with “Yes”)/(Number of responses)

Figure 7 shows "Ease of risk assumption" for each use case. Chatbot, recruitment AI, recidivism risk prediction, facial recognition by police, and photo tagging scored over 0.8. In these use cases, participants could extract the risk “AI makes discriminatory decision” from the explanatory text. For teacher evaluation, manufacturing robot, and video interview screening cases, however, participants were more likely to find it difficult to assume risks. In these use cases, it is important to investigate the factors that make risk identification difficult and to consider improvement measures.

F. Example of questions using use cases of recruitment AI

The results of the responses to the recruitment AI shown in Figure 5 and Table V are discussed. In Table V, with regard to the risk of "The word "women" in a resume lowers the score" associated with the interaction ID 111 which is from the Machine learning to the AI model in Figure 5, a participant who selected "Valid but other interactions" responded to Q1-2 regarding which other interactions were associated with the risk. The participant answered that interaction ID 108 which is from user company to training data in Figure 5 were associated with the same risk as interaction ID 111. This response suggest that the participant considered the risk is associated not only interactions between components of an AI system, but also interactions involving training data provider.

G. Summary of initial validation

From the result of the questionnaire, we confirmed that participants with knowledge of AI ethics generally agreed on the nature of the risks they identified and the interactions that occur. We also found that the task of assuming the risk from the ethical characteristics of AI associated with the interaction and its explanation differs depending on the use case.

However, these results are not sufficient for validation because the number of participants was too small, and the results are biased toward researchers with knowledge of AI ethics. Therefore, it will be necessary to validate our methodology and improve it by involving a wider range of participants.

VI. SUMMARY AND FUTURE WORK

We proposed the AI Ethics Impact Assessment, a methodology for comprehensively extracting potential ethical risks in AI systems in accordance with AI ethics guidelines. On the basis of an analysis of past ethical issues, we assumed that ethical risks are associated with interactions between AI systems and stakeholders. From this assumption, we developed the AI Ethics Impact Assessment, which involves constructing an AI ethics model that embodies ethical guidelines and associating it with interactions using requirement engineering, and an impact assessment process. A questionnaire survey of participants was conducted as an initial validation of our methodology. The results of the questionnaire suggested the validity of our assumption that ethical risks can be associated to interactions between AI systems and stakeholders. However, problems remain with the small number of participants, the bias toward those who have knowledge of AI ethics, and the setting of questions. In the future, it is important to improve the verification method and show reliable results.

ACKNOWLEDGMENTS

We are grateful to mediaX at Stanford University for holding a valuable workshop to advance this research.

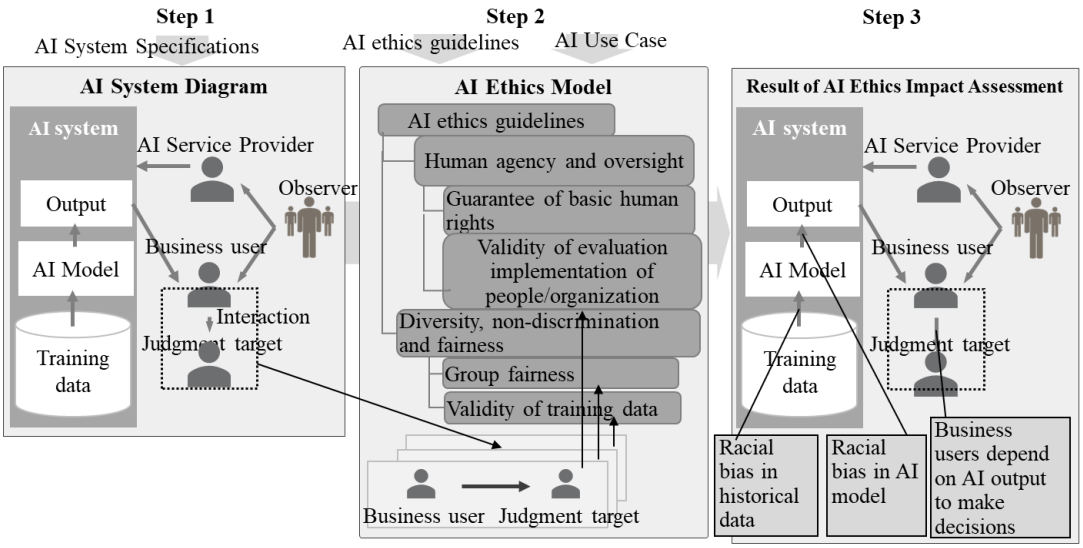

Fig. 3: Overview of the AI Ethics Impact Assessment

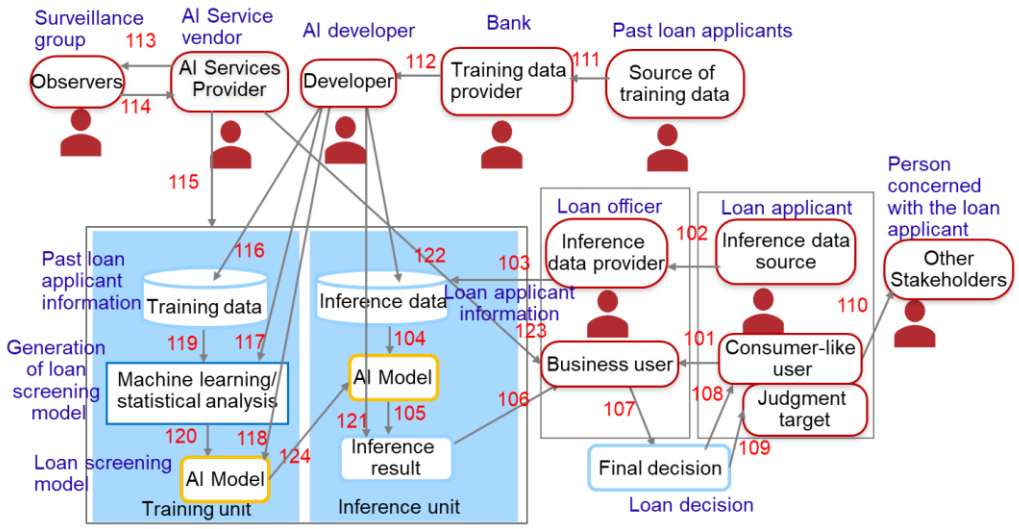

Fig. 4: System diagram of the loan screening AI case

TABLE III. AI ETHICAL CHARACTERISTICS CORRESPONDING TO INTERACTIONS IN LOAN SCREENING USE CASE

InteractionID	Risk type	Category 4(AI Ethics Characteristics)	Summary	Risk event or risk factor example	Risk Event or Risk Factor
105	Factor	Group fairness	Differences of inference results by AI are within the tolerance range between groups of protection attributes such as gender, race, age	In recruiting AI, female or black are less likely to be adopted	
107	Factor	Validity of evaluation implementation of people/organizations	Validity of deciding capabilities of people and organizations based on AI system outputs has been confirmed	Recruitment staff decides the final decision to the job applicant without verifying the validity of the AI result	
109	Event	Maintenance of social trust	Trust of stakeholders is not compromised by using the AI system	Loan officer makes final decision to the loan applicant without validating fairness of the AI result	
118	Factor	Sufficiency of test scenarios	AI model tests are designed by assuming specific groups or cases with likeliness of issues	Risks of learning conversations of malicious users are not assumed when designing a chatbot	

TABLE IV. ETHICAL ISSUE CASES USED IN QUESTIONNAIRE

No	ID	Name	AI task	Ethical issues
1	6	Chatbot	Text generation	Chatbot replies in a discriminatory chat
2	9	Teacher evaluation	Classification	The teachers' union filed a lawsuit claiming that the AI's assessment was unwarranted
3	11	Recidivism risk prediction	Classification	Black people are more likely than white people to be falsely predicted by AI as having a higher risk of recidivism
4	16	Photo tagging	Classification	Photos posted on social networks are racially tagged
5	24	Manufac-turing robot	Image recognition, environment sensing	The robot could not recognize the approaching worker and caused a contact accident
6	36	Traffic violator detection	Classification	AI wrongly detected an irrelevant person
7	37	Recruit-ment AI	Classification	Discrimination against women in screening results
8	74	Facial recognition by police	Classification	An irrelevant citizen was wrongly arrested
9	95	Video interview screening	Classification	Gender and race bias in results from video interviews

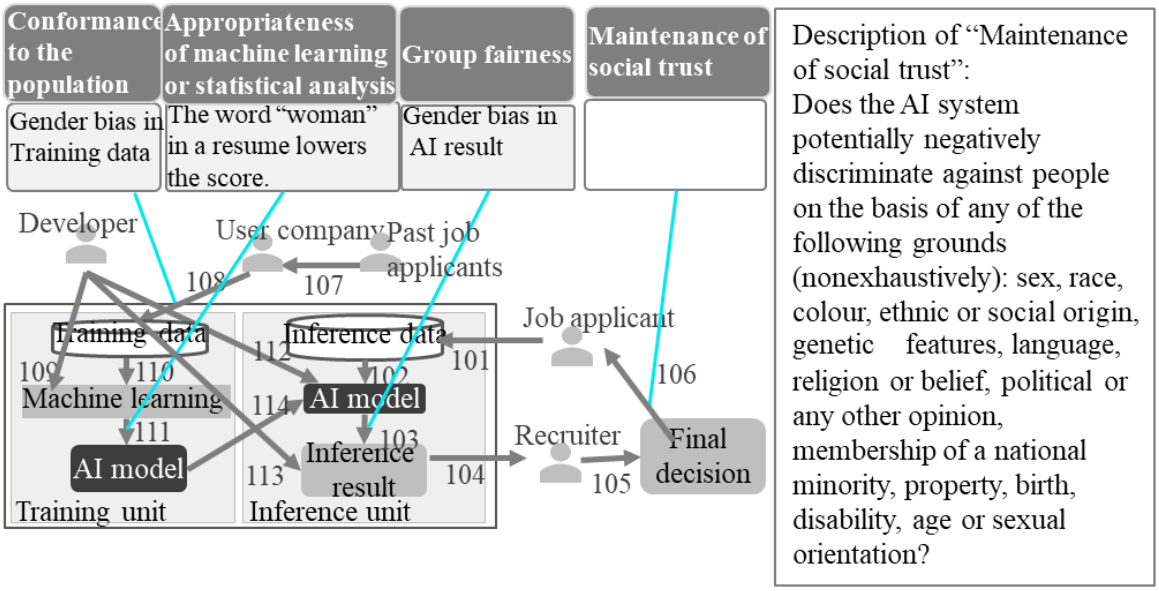

Fig. 5 A result of application example of recruitment AI case

TABLE V. AN EXAMPLE OF THE QUESTIONNARE ABUOT THE RECRUITMENT AI

Interaction ID: Risk	Q1-1.Are the interactions linking the risk valid?	Q1-2.If you select "Valid, but there are other relevant interactions", provide the ID of any other interactions that may apply	Q1-3 (Reasons for selecting the response from Q1-2	Q1-4.Are there any interactions that are not shown in the system diagram that are considered to pose risks?
103: Gender bias in AI result	Valid	N/A	N/A	N/A
111: The work "woman" in a resume lowers the score	Valid but there are other relevant interactions	108	The training data provided by the training data provider reduced the employment score of resumes containing "women".	N/A
108: Gender bias in training data	Valid	N/A	N/A	N/A

Interaction ID	Q2-1. Based on the text of the guideline, assume the concrete risks and describe them in text	Q2-2. Is that easy to answer Q2-1?
106	Recruitment AI uses attributes other than the gender or race of the job applicants, giving society the impression that the company doing business respects fundamental human rights. However, in reality, the results of AI are biased by gender and race, making it impossible to maintain the social credibility of job applicants.	Yes

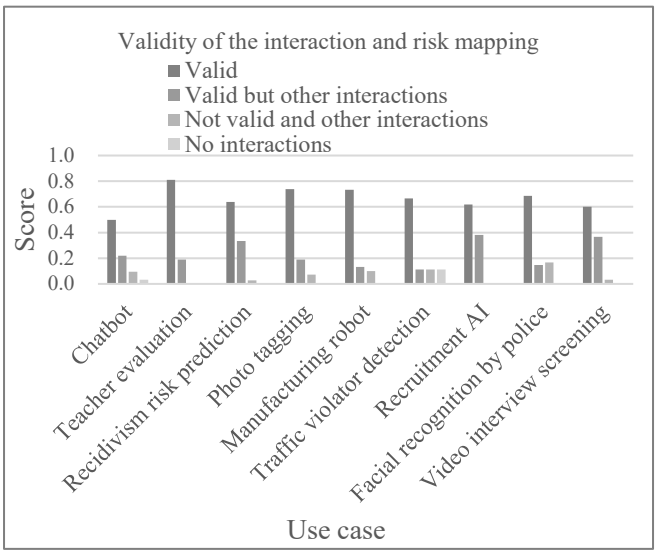

Fig. 6. Results of Question 1

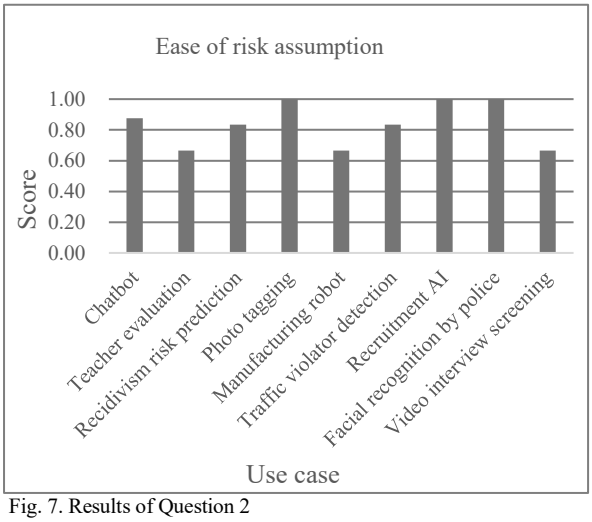

Fig. 7. Results of Question 2

REFERENCES

[1] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F, Rossi, B. Schafer, P. Valcke, and E. Vayena, "AI4 People-An Ethical framework for a good AI society: options, risks, principles, and recommendations," Minds and Machines, 28, 689 -707, 2018.

[2] The EU High-Level Expert Group on AI, "Ethics guidelines for trustworthy AI ",https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai? April, 2019.

[3] OECD, "OECD AI Principles overview ",https://oecd.ai/en/ai-principlesSay, May, 2019.

[4] IEEE, "Ethically aligned in design First Edition", https://ethicsinaction.ieee.org/wp-content/uploads/ead1e.pdf, 2019.

[5] The conference toward AI network society, "AI Utilization Guidelines ", https://www.soumu.go.jp/main_content/000658284.pdf, August, 2019.

[6] The cabinet office of Japan, "Social Principles of Human-centric AI (Draft)", https://www8.cao.go.jp/cstp/stmain/aisocialprinciples.pdf, 2019.

[7] European Commission, "Proposal for a Regulation of the European pariliament and of the council – Laying down harmonized rules on artificial intelligence (artistic intelligence act) and amending certain union legitimative acts", https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=ENG, April, 2021.

[8] Congress.gov, "H.R.3907 - Facial Recognition and Biometric Technology Moratorium Act of 2021", https://www.congress.gov/bill/117th-congress/house-bill/3907, June, 2021.

[9] T. Kamishima, "Fairness-Aware Machine Learning and Data Mining", https://www.kamishima.net/archive/faml.pdf, April, 2022.

[10] K. Kobayashi, Y. Nakao, "One-vs-One Mitigation of Intersectional Bias: A General Method to Extend Fairness-Aware Binary Classification", DiTTE 2021, pp. 43 -54, Octover, 2021.

[11] Y. Nakao, S. Stumpf, S. Ahmed, A. Naseer, L. Strappelli, "Towards Involving End-users in Interactive Human-in-the-loop AI Fairness", https://arxiv.org/abs/2204.10464, April, 2022.

[12] Government of Canada, "Algorithmic Impact Assessment tool ", https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html, April, 2022.

[13] Ada Lovelace Institute, "Algorithmic impact assessment: a case study in healthcare", https://www.adalovelaceinstitute.org/report/algorithmic-impact-assessment-case-study-healthcare/, February, 2022.

[14] L. Floridi, M. Holweg, M. Taddeo, J.A. Silva, J. Mökander, Y. Wen, "capAI – A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act", https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064091, March, 2022.

[15] R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, and P. Hall, "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence", NIST Special Publication 1270, March 2022.

[16] T. Matsumoto and A. Ema, "RCModel, a Risk Chain Model for Risk Reduction in AI Services", https://arxiv.org/abs/2007.0321, 2020.

[17] K. Ohashi, A. Katayama, N. Hasegawa, H. Kurihara, R. Yamamoto, J. Doerr, and D.P. Magin, "Focusing Requirements Elicitation by Using a UX Measurement Method", 2018 IEEE 26 th International Requirements Engineering Conference, October, 2018.

[18] ISO/IEC 25010: 2011, "Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- System and software quality models", 2011.

[19] ISO/IEC 24030: 2021, "Information technology — Artificial intelligence (AI) — Use cases", https://www.iso.org/standard/77610.html, May, 2021.

[20] S. Mcgregor, "When AI Systems Fail: Introducing the AI Incident Database", https://partnershiponai.org/aiincidentdatabase/, November, 2020.

[21] https://www.fujitsu.com/global/about/research/technology/aiethics/, April, 2022.

[22] https://www.ibm.com/artificial-intelligence/ethics, June, 2022.

[23] https://www.microsoft.com/en-us/ai/responsible-ai, June, 2022.

[24] V. Vakkuri, K-K. Kernell, M. Jantunen, E. Halme, and P. Abrahamsson, "ECCOLA – A method for implementing ethically aligned AI systems", Journal of Systems and Software, 182, December, 2021.

[25] M. Borg, J. Bronson, L. Christensson, F. Olsson, O. Lennartsson, E. Sonnsjö, H. Ebabi, and M. Karsberg, "Exploring the Assessment List for Trustwrthy AI in the Context of Advanced Driver-Assistance Systems", IEE/AMC 2nd International Workshop on Ethics in Software Engineering Research and Practice (SEThics), June, 2021.

[26] B. Johnson, J. Smith, "Towards Ethical Data-Driven Software: Filling the Gaps in Ethics Research & Practice", IEE/AMC 2nd International Workshop on Ethics in Software Engineering Research and Practice (SEThics), June, 2021.

[27] European Comission, "Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment", https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment, July, 2020.