AI Ethics Impact Assessment

～From principles to practice～

White paper
2022

Executive Summary

As the application of AI expands to every aspect of society, the ethical issues that AI creates are becoming more and more evident. Against this backdrop, countries and organizations in Europe and elsewhere are formulating AI ethical principles and AI ethics guidelines, as well as practical technologies to address these ethical issues are developing. However, there is still a gap between principles and practice, and the challenge is to operationalize the first into the latter.

To address this issue, we developed a method to evaluate the ethical impact of AI on people and society based on AI ethics guidelines. First, to find clues to fill this gap, we looked at the ethical issues that AI has caused so far. As a result, we found that the problems arose from interactions such as those between AI and stakeholders, as well as between stakeholders themselves. The AI Ethics Impact Assessment focuses on this point and comprehensively identifies the ethical issues of AI by correlating the requirements for trustworthy AI described in the ethics guidelines with the interactions that appear in AI systems. By conducting this impact assessment at the design and audit stage, proactive measures can be taken to prevent AI from causing serious social problems.

A practice guide showing the procedures for conducting this AI Ethics Impact Assessment and examples of its application to representative cases will be provided free of charge. With this as a first step, we aim to provide trustworthy AI together with countries and organizations that share the same ideas and people with diverse knowledge and perspectives, not only in technology but also in fields such as law and philosophy.

1. Introduction

AI Challenges

AI is widely used in all aspects of society, including financial transactions, healthcare and employment. AI systems can reduce workload by automatically and instantaneously performing work in which the person in charge had corrected each document for tax returns and the like, and enable employees to make decisions that could not have been made easily in the past, such as selecting appropriate drugs according to the type of cancer a patient has based on vast amounts of data. On the other hand, the ethical problems caused by AI have come to light and become widely reported. Facial recognition AI has produced racially discriminatory results, and hiring AI has been suspended due to sexist results. When these problems occur, in addition to the companies and organizations that provided the AI losing public trust, there is also a detrimental impact on the users of AI and on society.

Fujitsu has long cherished the idea of putting people at the center of everything. Based on this idea, we are working to resolve social issues and reform our business. In the social implementation of AI as well, Fujitsu is aiming to enable the usage of AI with peace of mind, that is, to realize trusted AI, through this human-centered approach.

Efforts toward trustworthy AI

Why does AI cause problems? Because machine learning is based on historical data, there are cases in which AI makes biased decisions by learning from biased data, including past cases of discrimination and unfairness, even those that were not clearly identified by society. In addition, there is a growing concern about the use of AI to make decisions without human intervention, without having clarity about the grounds which led to those decisions.

Against this backdrop, countries and organizations in Europe and elsewhere are working to develop AI ethical principles (hereinafter, “ethical principles”) which indicate the fundamental philosophy, and AI ethics guidelines (hereinafter, “ethics guidelines”), which indicate requirements for applying the ethical principles, in order to promote trustworthy AI. Legislation has also begun to prohibit the use of AIs for certain practices, and to impose strict limits on their use in situations where critical decisions for humans are made, including law enforcement and public services. At the same time, the development of, e.g., technologies to mitigate and correct the bias in AI judgment and technologies to explain the rationale behind the decisions of AIs is moving forward.

In this way, efforts are being made to provide trustworthy AIs both from the perspective of clarifying the requirements to be met by trustworthy AI through the formulation of ethics guidelines, and from the perspective of being able to address potential problems that may arise.

Our Efforts toward "Practice": AI Ethics Impact Assessment

The ethical principles and the ethics guidelines that follow them are important in clarifying a trustworthy AI philosophy and the requirements such AIs must meet. Similarly, practical techniques to ensure AI complies with these principles are also essential. However, there is a large gap between principles and practices, and the answer to the question of how to connect them is not always clear.

Hence, we decided to focus on the interaction between AI and its surrounding stakeholders as a way to bridge this gap between principles and practice. This is because, just as ethical issues, including fairness, arise in social relationships, AI incidents are observed between AI and stakeholders, between stakeholders themselves, or in a chain of these relationships.

The AI Ethics Impact Assessment that we supply involves providing information to design and audit your AI so that it would not cause serious social problems, based on the knowledge systematized by analyzing ethics guidelines and by comparing them with past incidents. Specifically, by clarifying the interactions between AI and the people around it, and then analyzing these interactions according to the procedure provided, it is possible to comply with ethics guidelines and obtain information for avoiding known incidents.

1. Introduction

Aims of the AI Ethics Impact Assessment

We applied the AI Ethics Impact Assessment to the cases registered in the AI Incidents Database provided by Partnership on AI[※1] (hereinafter, “PAI”) . PAI is a nonprofit organization that addresses ethical issues caused by AI, and its AI Incidents Database provides an opportunity for AI stakeholders to learn from their incidents. We detected inconsistencies with the requirements in the Ethics Guidelines for Trustworthy AI[※2] (hereinafter, “Trustworthy AI”) provided by the European High-Level Expert Group on AI for the cases in the database, and confirmed that incidents were mapped to interactions and could be detected.

The following materials, being the products of the above study, will be released free of charge. One is a practice guide consisting of an AI Ethics Model which organizes requirements from Trustworthy AI by substantiating them and mapping them to interactions; and a manual to apply the AI Ethics Impact Assessment using the AI Ethics Model. The other is an AI Ethics Impact Assessment application casebook consisting of application results to those cases. By applying this impact assessment to various AI cases, we aim to prevent ethical problems caused by AI before they can take place and make it possible to provide AI with peace of mind.

※1） https://partnershiponai.org/

※2） https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

2. Trends in AI ethics

Developments in AI ethical principles and AI ethics guidelines

In response to growing awareness of the ethical issues posed by AI, ethical principles and ethics guidelines that present basic principles are being formulated, aiming to drive the spread of trustworthy AI. In Japan, the Cabinet Office has issued “Social Principles of Human-Centric AI[※3]” . The OECD has published the first international policy guideline on AI, “OECD Principles on Artificial Intelligence[※4]” , which includes these ethical principles. Based on these OECD principles, as more practical guidelines, “AI Utilization Guidelines[※5]” has been published by the Ministry of Internal Affairs and Communications, and “Governance Guidelines for Implementation of AI Principles[※6]” has been published by the Ministry of Economy, Trade and Industry.

In the European Union, the "Ethics Guidelines for Trustworthy AI," created by the High Level Expert Group on AI, has been published by European Commission, aiming to act as a guideline for encouraging the development of trustworthy AI.

IEEE has issued “Ethically Aligned Design[※7]” as a guideline for promoting ethical AI development.

Ethics guidelines are established in accordance with the values of each region and organization, so each has its own characteristics. However, common points can be recognized across these different guidelines.

Legal and regulatory trends for AI

In order to avoid social turmoil caused by AI, the regulatory landscape is advancing with proposals for new laws. In the European Union, proposed Artificial Intelligence Act[※8] from the European Commission has been published. This proposed Artificial Intelligence Act categorizes the manipulation of people's subconscious minds, the use of social scoring, and remote biometrics for law enforcement purposes in public spaces as "AI that create an unacceptable risk" and prohibits this type of AI. They also list the use of AI in personal biometrics and classification, and its application to critical infrastructure as "high-risk AI," and impose a number of obligations 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[※9]" , which prohibits federal officials from using facial recognition technology, has been proposed. In addition, the city of San Francisco has also banned the use of facial recognition technology by police.

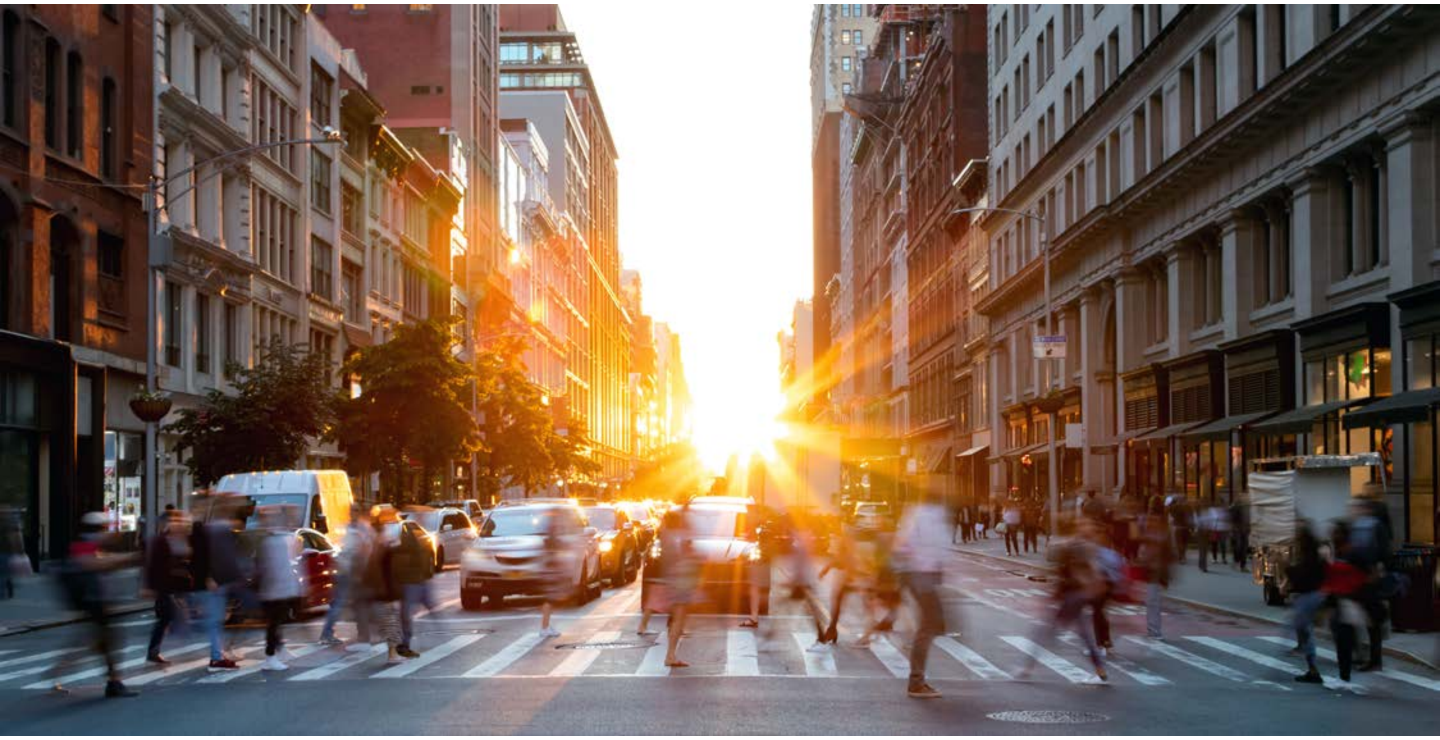
Trends in formulation of AI development guidelines

Based on the ethics guidelines presented in Japan and overseas, development guidelines and development processes that are more suited to sites of development are being created, and systems to prevent ethical problems from the AI development stage are being built. As part of "Next-Generation Artificial Intelligence that Evolves with People," led by the National Institute of Advanced Industrial Science and Technology (AIST) as a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO), "Machine Learning Quality Management Guideline[※10]" has been published, placing a focus on AI quality. This is a development guideline concerning AI quality, and the second edition has been enhanced and expanded to also include fairness. Moreover, AI developers and vendors are also developing their own development guidelines.

Trends in technical development to address AI ethical issues

Aiming to comply with ethics guidelines, the development of technology to address issues related to the fairness of AI is moving forward. Several standards have been defined regarding fairness. For example, for the attribute (e.g., gender) for which fairness is being considered, there is one approach that equalizes the probability of the potential outputs provided by the AI (e.g., in the case of recruitment AI, the hiring and rejection of candidates), and another that equalizes the probability of AI output being adopted in correct data. Machine learning algorithms that follow these various standards are being studied extensively.

"Explainable AI" technologies that offer insight into how the AIs make decisions are also being developed. For example, techniques are being developed to show which areas of an image the AI has focused on during image recognition tasks, and to quantitatively show what attributes of the inference object contributed to the AI's judgment.



Desirable future developments

In this way, in countries and regions where the AI industry is at the center of policy, many achievements concerning the attitudes and realization of AI ethics across academic disciplines, as well as across industry, government, and academia, have become known. These will undoubtedly affect various industries not only in the countries and regions concerned, but also globally. In order for AI, which has a major impact on society, to continue to be an industry that conforms to human-centered values, it is highly desirable that AI researchers continue to conduct cutting-edge research in each area of specialization and engage in constructive communication with society. At the same time, we believe that a new perspective is needed in order to connect the required expertise across different areas of specialization.

※3） https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/aigensoku.pdf
※4） https://www.oecd.org/tokyo/newsroom/forty-two-countries-adopt-new-oecd-principles-on-artificial-intelligence-japanese-version.htm
※5） https://www.soumu.go.jp/main_content/000637097.pdf
※6） https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20210709_6.pdf
※7） https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf
※8） https://ec.europa.eu/commission/presscorner/detail/en/IP_21_1682
※9） https://www.congress.gov/bill/117th-congress/house-bill/3907
※10) https://www.digiarc.aist.go.jp/publication/aiqm/

3. AI Ethics Impact Assessment

Ideas for AI Ethics Impact Assessment

Following years of fundamental discussions about the principles of AI ethics at the philosophical, technical, and policy levels, the discussion is now focusing on the transition from these principles to practice. In order to obtain AIs that make these principles operational, there are at least two approaches: one is to ensure that each AI develops to be reliable, the other is to strengthen technologies able to handle the ethical aspects. It is conceivable to adopt a realistic approach that combines both.

As a complementary approach, we considered a perspective common between various cases to design AI to behave ethically and to audit that it functions in that way. In observing past incidents, it has become apparent that incidents occur in interactions and chains of interactions between AI and its surrounding stakeholders. These observations suggest the possibility of an objective assessment of AI's compliance with ethical guidelines by focusing on the interactions between AI components and stakeholders. Based on software requirements engineering, we developed an AI Ethics Impact Assessment through detailed analysis of different ethics guidelines and specific AI use-cases. The following section contains an overview of this AI Ethics Impact Assessment.

What is the AI Ethics Impact Assessment?

The AI Ethics Impact Assessment identifies the ethical issues resulting from AI and the factors that cause them as risks based on ethics guidelines by the following process:

First, an AI ethics model is created that systematically organizes where, on the AI system, the ethical requirements based on the ethics guidelines should be confirmed (Figure 1). Specifically, the model applies the requirements definition techniques, normally used in software engineering, to match the conceptual descriptions of ethics guidelines with specific ethical requirements. It identifies problematic interactions by creating links with the results of analysis of previous incidents. This AI ethics model only needs to be created once per ethics guideline and can be used in various AI impact assessments.

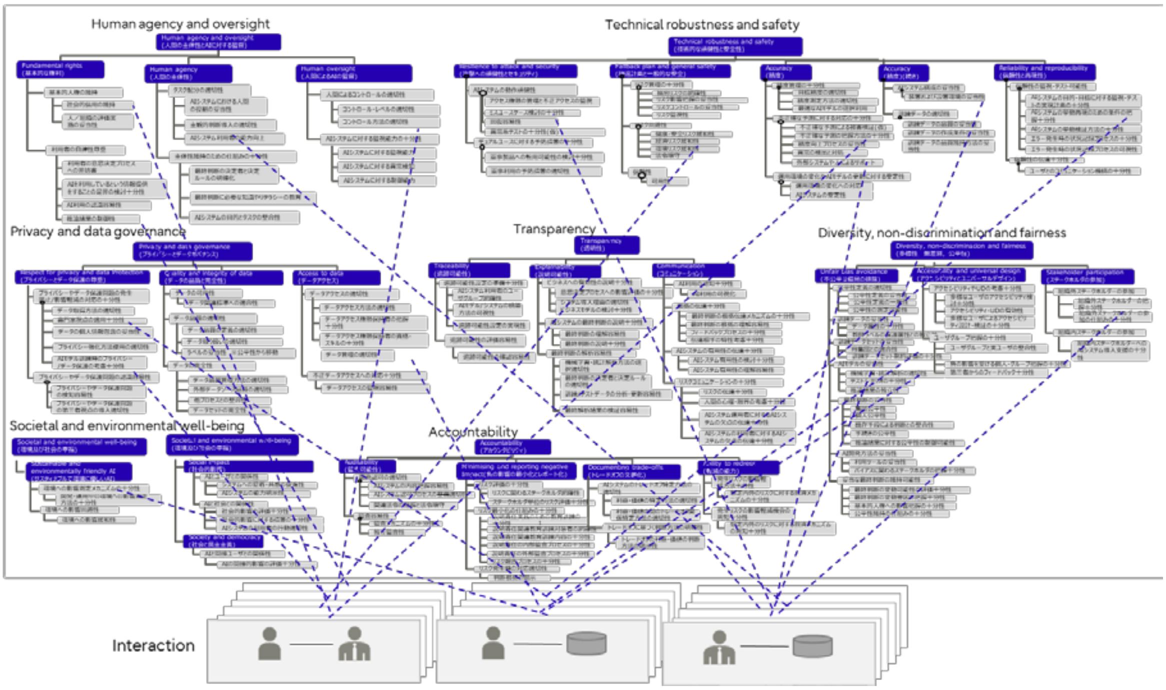

Human agency and oversight
Technical robustness and safety
Privacy and data governance
Transparency
Diversity, non-discrimination and fairness
Societal and environmental well-being
Accountability
Interaction

Figure 1. AI ethics model

Next, by mapping the interactions on the AI system to the AI ethics model, the potential ethical risks are extracted from the ethical requirements to be considered (Figure 2). The AI system is represented by a system diagram in which the components of the system and related stakeholders are represented, based on the modeling methods used in software development. This system diagram can be created as a variation of several patterns of system diagrams extracted from AI cases. This was confirmed at a joint workshop[※11] with mediaX at Stanford University. Once the system diagram is created, risks are systematically identified by extracting specific ethical requirements from the AI ethics model in accordance with the interactions that appear in the diagram.

In order to conduct this AI Ethics Impact Assessment, we will make available free of charge an AI Ethics Impact Assessment practice guide consisting of the AI ethical model, and an implementation procedure manual that includes, as an appendix, a pattern sheet of the system diagram used when creating the system diagram.

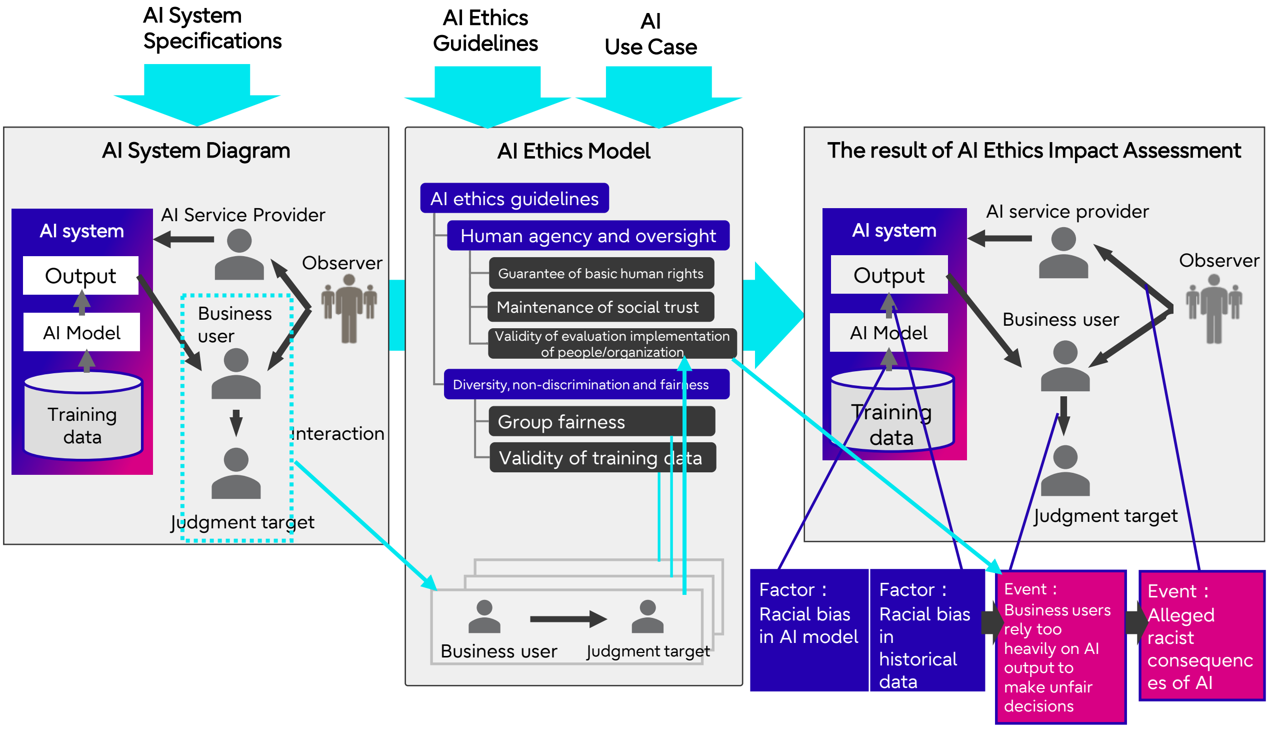

AI System Specifications
AI Ethics Guidelines
AI Use Case
AI System Diagram
AI system
Output
AI Model
Training data
AI Service Provider
Observer
Business user
Judgment target
Interaction
AI Ethics Model
AI ethics guidelines
Human agency and oversight
Guarantee of basic human rights
Maintenance of social trust
Validity of evaluation implementation of people/organization
Diversity, non-discrimination and fairness
Group fairness
Validity of training data
Business user
Judgment target
The result of AI Ethics Impact Assessment
AI system
Output
AI Model
Training data
AI service provider
Observer
Business user
Judgment target
Factor : Racial bias in AI model
Factor : Racial bias in historical data
Event : Business users rely too heavily on AI output to make unfair decisions
Event : Alleged racist consequences of AI

Figure 2. Overview of AI Ethics Impact Assessment

[※11] https://mediax.stanford.edu/event/ai-ethics-workshop/

Effect of AI Ethics Impact Assessment

What can be learned by conducting an AI Ethics Impact Assessment is explained in the analysis results (Figure 3) applied to a fictitious AI system aiming to predict recidivism risk. In this system, the AI learns from historical defendant data, and predicts the recidivism risk of a new defendant based on information about the new defendant. Judges use the AI's prediction results to help determine whether a defendant is to be released on parole or given a sentence. From the extracted risks shown in Figure 3, we can assume the possibility that the judgment of the judge will be biased or even picked up by the media due to biases or unfair prediction results caused by the attributes in the historical data. In some parts of the United States, a similar AI system for predicting recidivism risk was actually adopted, and there was an incident in which the media published an article about “AI for predicting recidivism risk is discriminatory [Note 1]”, which caused controversy. Although the incident in the United States may differ slightly from the fictitious case because the only available information is in the articles, the analysis results of this fictitious case show that the AI Ethics Impact Assessment was able to extract the risks that caused the incident [Note 2]. It also makes it easier to take concrete actions due to visualizing the interactions where the different risks are taking place. In this way, the AI Ethics Impact Assessment is conducted systematically in accordance with the procedure manual, enabling full assessment of the impact within the scope of ethics guidelines.

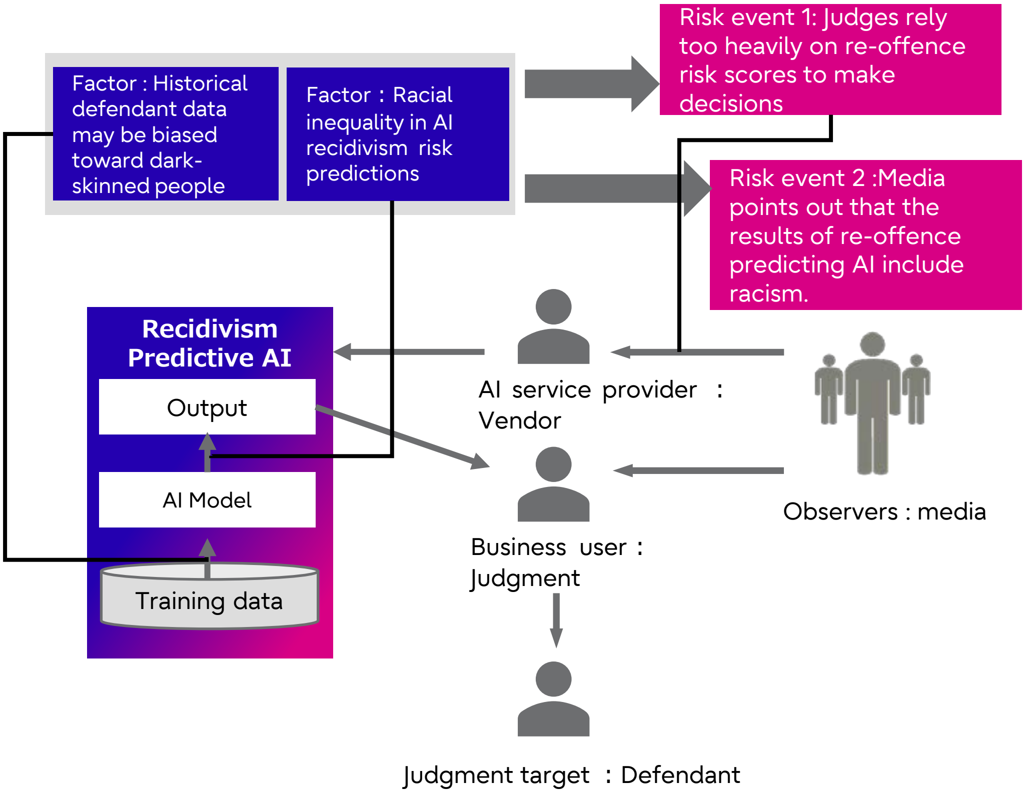

Factor : Historical defendant data may be biased toward dark-skinned people
Factor : Racial inequality in AI recidivism risk predictions
Risk event 1: Judges rely too heavily on re-offence risk scores to make decisions
Risk event 2 :Media points out that the results of re-offence predicting AI include racism.
Recidivism Predictive AI
Output
AI Model
Training data
AI service provider : Vendor
Observers : media
Business user : Judgment
Judgment target : Defendant

Figure 3. Application example of AI Ethics Impact Assessment for recidivism risk prediction AI

3. AI Ethics Impact Assessment

The AI Ethics Impact Assessment was applied to previously unseen incidents to verify its effectiveness. Of more than 150 AI incident cases registered in the AI incident database provided by PAI, approximately 15 cases were selected as test incidents, categorized by industry and type of AI application. As a result of testing the AI Ethics Impact Assessment, we confirmed that the risks causing the different incidents can be mapped to interactions between AI and stakeholders, and that all of them can be extracted. It should be noted that the AI ethics model used in this verification was created from Trustworthy AI. We will make available, free of charge, an AI Ethics Impact Assessment practice guide consisting of this AI ethics model and an analysis procedure manual, and an AI Ethics Impact Assessment application casebook consisting of application examples to representative cases.

4 . Conclusion

The AI Ethics Impact Assessment systematically and comprehensively aims to assess the ethical impacts of AI, and provides information on where ethical issues caused by AI can occur. By applying it before the development or delivery of AI systems, the ethical issues posed by these systems can be proactively addressed. We will develop the AI Ethics Impact Assessment, and aim to spread trustworthy AI to society by adopting not only the technology of AI, but also varied knowledge and viewpoints from fields such as law and philosophy, along with countries and organizations which have a common philosophy regarding trustworthy AI.

Notes

[1] Machine Bias - There's software used across the country to predict future criminals. And it's biased against blacks, by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica, May 23, 2016 https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] The examples and analysis figures presented here include some speculative descriptions. Generally, the AI Ethics Impact Assessment conducted by Fujitsu will clarify the details of a case and proceed with analysis. However, please note that the facts may not necessarily be as described in this article, since some portions of the cases introduced here have not been disclosed in detail.

FUJITSU LIMITED
Research Unit　Research Center for AI Ethics
E-mail： fj-labs-aie-info@dl.jp.fujitsu.com

Trademarks

All brand names and product names are trademarks or registered trademarks of their respective holders.

Disclaimer

FUJITSU DOES NOT GUARANTEE THAT THERE ARE NO RISKS OTHER THAN THOSE IDENTIFIED BY THIS METHOD, AND IT IS NECESSARY FOR EACH ORGANIZATION OR INDIVIDUAL USING THIS METHOD TO DETERMINE AND IMPLEMENT MEASURES IN SPECIFIC CASES. FUJITSU SHALL NOT BE LIABLE FOR ANY DAMAGES WHATSOEVER RESULTING FROM THE USE OF THIS METHOD OR THIS MANUAL.

Version 1.0.1, February 2022