December 2022
JDLA

Study Group 'AI governance and its Evaluation'
Report on the Session #3 (Phase III)

1. Introduction

Japan Deep Learning Association establishes study groups as a forum for deepening knowledge and discussing domestic and international policy trends related to artificial intelligence (hereafter AI) and Deep Learning (hereafter DL). This study group, 'AI Governance and its Evaluation,' defines 'governance' as a system of management and evaluation by various actors and launched a study group in July 2020 to investigate what forms of governance are possible to help build trustworthy AI systems, and the Phase III began this year, 2022.

In the third session of Phase III (November 11, 2022), the theme of "Ethical Framework" for realizing AI governance was discussed, and Sachiko Onodera (Research Center for AI Ethics, Fujitsu Limited) presented the AI Ethics Impact Assessment (AIEIA) developed by Fujitsu Limited as a method to assess the ethical risks of AI.

2. Introduction of AI Ethics Impact Assessment (AIEIA) by Sachiko Onodera (Research Center for AI Ethics, Fujitsu Limited)

Overview of AIEIA

Fujitsu has been actively working on AI ethics from a relatively early stage in Japan, such as announcing its AI Commitment and starting AI ethics education for all employees in 2020. In April 2021, the Research Center for AI Ethics was established to promote not only research and development on AI ethics, but also collaboration with related external initiatives and dissemination of information.

AIEIA was developed in response to the awareness of how to incorporate the ethical principles and the ethics guidelines into the actual development and provision of AI services, as they have been presented by various countries and organizations.

In developing the AIEIA, past incidents were analyzed to identify situations where ethical issues have arisen. What emerged was that it is important to identify who uses AI systems, where, and how. Another point is that ethical issues can arise in the areas where stakeholders are directly or indirectly involved with AI systems as elements, and where those elements are involved with AI systems (interactions). AIEIA, which was developed based on the above points, is a method to derive possible risks by following

steps: 1) organize the components of AI systems and interactions with each stakeholder as an AI system diagram, 2) identify the ethical requirements corresponding to the interactions in the system diagram in 1) above, then 3) identify the opposite state of those ethical requirements as "risks".

AIEIA application case studies

The AIEIA is available to the public [1], including assessment results using representative past incidents extracted from the Partnership on AI's AI Incident Database. The main cases applied included the case of a recruiting AI that screens candidates for interviews based on their resumes, and a loan screening AI that makes loan decisions. By conducting risk assessment using AIEIA, we can organize what could happen in each interaction (risk likelihood factors), and also see that the risks extracted will differ depending on how the same AI system is implemented.

For example, in the case of using AI to support the judgment of loan officers in loan screening at a bank, verification was conducted on two different AI system configurations. In Case 1, the AI's decision results are notified to the loan officer, who then makes the final decision based on the results; in Case 2, the AI's decision results are automatically notified directly to the loan applicant. In both cases, the risk that the AI's screening results may have a bias toward a particular gender or race due to factors such as bias in the training data is extracted. In addition, in Case 2, the risk that the applicant will be directly notified of the results, thereby depriving him or her of the opportunity to appeal is also extracted. Thus, the ethical issues that can be anticipated will vary depending on the method of implementation.

Future Development

AIEIA is a method that makes it possible to systematize the guideline, which are vast amounts of written information, and to identify where and what risks occur in a system by focusing on interactions. By making its contents public, it aims to put into practice AI ethics initiatives with domestic and international partners. AI ethics is an issue that should be addressed by society, and Fujitsu intends to promote discussion and consideration of this issue together with various stakeholders. The official version is expected to be released by the end of FY2022.

3. Main comments from the participants

The main discussion topics are as follows.

[1] https://www.fujitsu.com/jp/about/research/technology/aiethics/ (Japanese)
https://www.fujitsu.com/global/about/research/technology/aiethics/ (English)

- **Assess and address the risks identified**
  - ✓ AI ethical requirements tied to interactions are automatically extracted, but the concretization and evaluation of events that violate the requirements (risks) is a human task. Depending on the individual business, industry, and situation, a person with knowledge of the business interprets and picks up risk factors.
  - ✓ The decision on how to deal with the risks identified in the evaluation, including the acceptable range and how to reach a consensus with the user side, will be made by human. As a preliminary step, AIEIA can be used as a method to understand what kind of risks exist.
  - ✓ Even when the same AI system is evaluated, there may be cases where the requirements extracted are not risks, depending on the purpose of its use and the culture and ethics of the country. Appropriate judgment should be made as to where the real risk points are. In addition, it is important that the stakeholders involved are fully informed and convinced of the purpose of use and possible issues.

    It goes without saying that it is important to identify risks in advance before introducing an AI system, but since interactions often increase during the operational phase, it is necessary to extract, analyze, and evaluate interactions at each timing of changes in the scope of application. Similarly in the development phase, as development progresses and AI models become more detailed, some risks may become apparent, so that risk assessment at each stage of the AI lifecycle will help prevent omissions.

- **AIEIA Verification**
  - ✓ AIEIA is intended to be examined with not only the cases in the AI Incident Database, but also domestic cases such as AI risks using camera images, which have been the focus of much attention recently.
  - ✓ It is necessary to accumulate case studies of verification targeting the operational phase in the future.

- **Scope and update of AI ethics model**
  - ✓ The AIEIA's AI ethics model is based on the ethics guidelines and is structured to identify risks within the guidelines (within the minimum line requirements in other words). Updates to the model will be made as the guidelines are revised.
  - ✓ In some cases, broader requirements than the guidelines (e.g., social acceptability) are considered, and in other cases, the ethics themselves change, so discussions are necessary in each case. Whatever the case may be, it is necessary to identify the risks within the scope of the minimum requirements to be followed, and then foster a common understanding of the other areas through discussions with

stakeholders.

- ✓ Currently, one model per guideline is supported (AIEIA is based on the Ethics Guidelines for Trustworthy AI). It is possible to create another model from another guideline, but since guidelines are sometimes tied to each other, the composition of a model with multiple additional guidelines is a future issue.

- **Human resources expected in the future**
  - ✓ AIEIA can identify risks in accordance with the guidelines, but the final decision on how to deal with those risks is made by a human, and that person must have an understanding of not only the technical aspects but also the business aspects. Instruction and training in these areas are also necessary.
  - ✓ Since the final decision is made by human, it is important to develop human resources who can not only use but also update the ethics guidelines. Widespread use of tools such as AIEIA and the sharing of knowledge to raise the level of human resources will make a significant contribution internationally. It can also be linked to JDLA's human resource development activities.

The 3rd Session (Phase III) of the Study Group
Date/Time: November 11th (Tuesday) 13:00-14:00 (On Zoom)
Contents:
- "AI Ethics Impact Assessment (AIEIA)" by Sachiko Onodera (Research Center for AI Ethics, Fujitsu Research, Fujitsu Limited)
- Questions & Discussion