Research area in Strategic Objective *"Trusted AI "*

The fundamental technologies for Trustworthy AI

Research supervisor: Hiroki Arimura (Professor, Division of Computer Science Graduate School of Information Science and Technology, Hokkaido University)

Overview

Systems and services utilizing artificial intelligence (AI) technologies are continuing to expand in society due to the rapid progress of mathematical sciences and information technology, and spread of the information environment such as networks and big data. This use of AI technology is expected to make it possible for all people to receive appropriate, high quality service in a Human-centric AI society, where people can demonstrate their individual abilities in harmony with society. On the other hand, recent efforts on realizing "trustworthy AI systems" that skillfully perform tasks that are important for people and society, have revealed a variety of weaknesses and limitations of the current AI technology, represented by deep learning, related to explainability, transparency, accountability, and fairness . Moreover, policies for the trustworthiness, safety, and quality assurance of so-called AI systems incorporating AI technology as a whole, and for the use of AI in harmony with society, with human beings as the starting point, are also necessary.

In this research area, our aim is to create the innovative theory and technology that will form the basis for next-generation AI technologies which break through the limitations of the current AI technology, toward the realization of Human-centric AI society. Rather than a simple extension of conventional AI technologies, we focus on the essential problems of current AI technologies and AI systems, promoting challenging research aimed at conceptualizing and formulating the problems that require solution from new perspectives, and solving those problems.

Concretely, we conduct researches on 1) basic technology on novel mathematical, computational, and analytical techniques for overcoming the weaknesses and difficulties of the current generation of AI technology, and 2) conceptualization and formulation of new AI application tasks in Human-centric AI society, including the trustworthiness, robustness, transparency, and fairness AI systems, together with new principles for their configuration and new technologies for their realization, 3) technologies for assurance of the trustworthiness, safety, and privacy of the data and information

infrastructure supporting AI systems, 4) extension of AI technology to diverse data and tasks in the real world and 5) methodologies for the design, development, and operation of AI systems.

It should be noted that this research area is administered as part of the Advanced Integrated Intelligence Platform (AIP Project) for integration of AI, big data, IoT, and cybersecurity of Japan's Ministry of Education, Culture, Sports, Science and Technology (MEXT).

Research Supervisor's Policy on Call for Application, Selection, and Management of the Research Area

1. Background

Use of systems and services that employ artificial intelligence (AI) is continuing to spread due to the rapid progress of machine learning and other core technologies. On the other hand, from the viewpoint of realizing a Human-centric AI society, various weak points and limitations of the current generation of AI technology have also become apparent. For this reason, policies for assurance of the trustworthiness, safety and quality of information systems and services incorporating AI technologies (so-called "AI systems") as a whole, and for use of AI in harmony with society, with human beings as the starting point, are also necessary. To realize "trustworthy AI" that skillfully performs tasks that are critical to both people and society, it is necessary to overcome these technological weaknesses and limitations.

For example, from the theoretical view point, there are many open problems in the principles of deep learning, which is one of the core technologies of current AI, and as a result, this technology is currently used as a kind of black box. Therefore, various weakness of deep learning have been pointed out: The explainability, transparency, and accountability of the results of predictions by deep learning are not adequately maintained; issues of fairness arise because the data used in learning contain elements of discrimination and prejudice; and AI has the weakness of being unable to respond to unknown/unexpected cases and environmental changes, and cannot understand and apply context and common sense. Moreover, new principles and technologies will be necessary in AI systems in order to guarantee the trustworthiness, safety, and quality of the systems as a whole, since these cannot be assured by existing software engineering and other conventional methodologies. Problems have also been pointed out in the data itself, including problems of trustworthiness, i.e., fake data and falsification, as well as the issue of misuse of AI in the creation and distribution of fake

data and falsification. Retrofitting will not address these issues. To solve these problems, design and implementation based on the concept of “By Design,” are necessary in order to satisfy the necessary conditions of the system from the design stage.

2. Objectives of research and development and examples of research themes

To solve the above-mentioned problems and utilize AI more widely in society, efforts to create novel concepts and methodologies that can fundamentally solve those problems are necessary. This will be achieved not by a simple extension of existing AI technology, but by clarifying the problems, weaknesses and limitations of current AI technology, and rapidly deploying the results to date in related fields such as information science and mathematical science. For this, in this research area, we promote challenging research based on free, novel concepts for next-generation AI technology and its supporting technologies which are proposed by young researchers. Through this, we aim to create the innovative theory and technology that will form the basis for “trustworthy AI” which contributes to the realization of a Human-centric AI society. Samples of research projects are presented in the following.

■Example 1: Research on overcoming the weaknesses of current AI technology

- Understanding of the principles, the possibilities and limitations of current AI technology, such as deep learning, and proposals for new AI configuration principles and technologies based thereon.
- Proposals for novel mathematical models, algorithms, and systems for the next-generation AI that works with new tasks and diverse types of data arising in Human-centric AI-society, and research to solve the difficulties in their social application.
- Research for realizing AI which integrate inductive information processing such as deep learning, and deductive information processing including symbolic reasoning, knowledge processing, and simulation techniques.
- Research on AI related to high-order machine learning tasks, not limited simply to accurate prediction of input/output relationships, but also including discovery of causal relationships, counterfactual explanations, and statistically significant predictions.
- Research on novel AI principles based on understanding of information processing in the human brain and the and human cognition process.
- For the above, we expect research that can solve problems not only in existing application targets, but also the characteristic problems of the new applications targets that are supposed to occur as a result of the permeation of AI technology in society.

■Example 2: Research to assure the trustworthiness and safety of AI-systems/services

・Research on new mathematical approaches, methodologies and technologies in connection with the design, development, and operation of information systems in which AI is the core technology (hereinafter referred to as “AI systems”).
・Research for the realization of AI systems/services which are robust against unknown/unexpected cases, environmental changes, data bias, malicious attacks (hacking) and others.
・Research on quality assessment/control and safe design of AI systems/services.
・New methodologies that enable assurance from the design stage (“By Design”) of the characteristics desired in AI systems/services in collaboration with people/society, namely, the trustworthiness, transparency, safety, privacy and fairness of AI systems/services.

■Example 3: Research on assurance of the trustworthiness of data for AI systems and decision-making and consensus-building using AI technology
・Proposals for novel mathematical models for realizing decision-making and consensus-building with desirable characteristics such as transparency and fairness in services and systems in which large numbers of participants and stakeholders in society will participate, and research on related formulas, algorithms, and systems.
・Research on technologies for detecting and taking action against fake data, data falsification and other fraudulent actions (including those attributable to AI) when using the results of data collection/management and systems using AI systems.
・For the above, research on novel concepts, mathematical approaches, information technologies, and social implementation, not limited to conventional research topics, but also including those that will bring about innovation as basic technologies of the Human-centric AI society of the future.

Please note that the possible research projects in this research area are not limited to the examples mentioned above. We are seeking challenging research concepts based on new thinking, which contribute to the realization of a Human-centric AI society, and trustworthy AI for that purpose, from diverse standpoints including mathematics, information technology, social application, etc.

3. Assumed methods to advance research

The aim of this research area is “trustworthy AI technology that contributes to a Human-centric AI society.” Because there are various conceivable approaches to realizing this AI technology, we hope to receive research proposals not only from fields in which AI is the core technology, but also from a diverse range related to the mathematics, information technologies, and social applications of AI technologies and systems. In promoting research, we intend to train human resources for research, who can conceive novel ideas and carry out advanced research contributing to the future, and to

construct a human network of researchers which will lead to future collaborations, by providing a platform for exchanges and mutual inspiration among fellow young researchers in different fields of specialization. This will include area-wide research area Meetings held, in principle, two times each year. In addition, in order to achieve the research objectives of this research area, perspectives other than science and technology in the narrow sense are necessary for promoting research. Therefore, we recommend exchanges and collaboration with related academic fields such as the humanities and social sciences related to this area, and with diverse stakeholders, including companies and governmental agencies; we also plan to provide opportunities for this in the research area. We hope that the participating researchers themselves will decide flexibly on participation in these opportunities, based on the researcher's own specialty and research mode.

4. Research periods and research funds

The budget for one research project at the beginning is 300 million yen at the maximum (direct expenses). The research period begins in fiscal year 2020 and ends in fiscal year 2025 (five and a half years or lesser).

5. Precautions for application

Please note that this research area is one research area in the "AIP Network Laboratory" comprising the Advanced Integrated Intelligence Platform Project (AIP Project) for integration of artificial intelligence, big data, IoT, and cybersecurity sponsored by Japan's Ministry of Education, Culture, Sports, Science and Technology (MEXT). As such, we are also contributing to the integrated operation of the AIP Project through efforts in research and development projects, while also collaborating with related research institutions and others, beginning with the RIKEN Center for Advanced Intelligence Project (AIP).

For details concerning the JST AIP project, please see:

https://www.jst.go.jp/kisoken/aip/en/index.html