Research area in Strategic Objective *"Trusted AI "*

Core technologies for trusted quality AI systems

Research supervisor: Akiko Aizawa (Professor, Digital Content and Media Sciences Research Division, National Institute of Informatics)

Overview

Artificial Intelligence (AI) technologies are being applied to a rapidly expanding range of applications in the world at large, and have become indispensable for the creation of new value in scientific, social and economic spheres. However as a consequence of their "black box" nature, facing built-in biases and other such limitations, "deep learning" and other machine learning technologies present a variety of reliability/safety related issues that must be addressed before widespread application.

Therefore our Research Area involves creating fundamental technologies, and constructing AI systems that incorporate these fundamental technologies, leading to the realization of trusted quality AI that humans can use widely and safely in society. In our research we also address such issues as the definition and assessment of the reliability/safety of Human-centric AI systems, the determination of requirements for such systems, and the establishment of technologies to meet those requirements.

More specifically, we direct our efforts toward the following research and development areas:

(1) Revolutionary/evolutionary AI technologies toward the realization of trusted AI.

(2) Technologies to ensure the reliability and safety of AI systems expected from a Human-centric society.

(3) Technologies to ensure data reliability and support human decision making within a Human-centric AI society.

Through these efforts we aim to open avenues to the resolution of various social issues, promote the creation of new science and value, foster a community for research into trusted AI and related fields, and heighten the presence of Japanese research within such fields.

This research area is managed as part of the AI, big data, IoT, and the cyber security integration project developed by the Ministry of Education, Culture, Sports, Science and Technology (AIP Project).

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

1. Background

AI technologies have been employed within a wide variety of systems and services within recent years in step with marked advances in such AI-related technologies as machine learning. However, there remain concerns over the reliability and safety of such systems/services. AI technology itself, along with the "deep learning" conceptual core upon which it is centered, has notable weaknesses including: insufficient explainability and persuasiveness of output, learning biased information included in data, vulnerability over unknown/unexpected cases and shifting environments; an inability to comprehend context; and a lack of "common sense". Also, the overall quality, reliability and safety of the systems and services to which AI technologies are commonly employed cannot be adequately assured with pre-existing software engineering and other methodologies, a deficiency that many observers believe will require the establishment of a whole new methodology to fully address. And there are also issues with regards to data integrity, including the creation, dissemination and altering of "fake data," and even to the application of AI for these malicious purposes.

2. Objectives of R&D and examples of research themes

The Research Area involves creating fundamental technologies, and constructing AI systems that incorporate these fundamental technologies, leading to the realization of trusted quality AI that humans can use widely and safely in society. In our research we also address such issues as the definition and assessment of the reliability/safety of Human-centric AI systems, the determination of requirements for such systems, and the establishment of technologies to meet those requirements. Presented below are some concrete examples of research themes that fit this category. Note, however, that we also welcome proposals for other themes beyond what we show here.

(1) Revolutionary/evolutionary AI technologies toward the realization of trusted AI.

- Research into AI technologies that optimally combine inductive information processing (as in deep learning) and deductive information processing (as in knowledge/language-based inferencing, planning and the like).
- Research into AI technologies that support knowledge acquisition/evolution through mutual

interaction with real-world environments (and do not require a vast amount of training data).

- Research on novel AI principles inspired by human brain information processing and human cognition process.

(2) Technologies to ensure the reliability and safety of AI systems expected from a Human-centric society.

- Research into technologies for constructing AI systems capable of explaining the basis for their decisions/inferences.
- Research into data processing technologies such as data expansion, data bias removal and data anonymization.
- Research into technologies for constructing AI systems with the robustness to accommodate unknown/unexpected cases and changing environments.
- Research into technologies for guaranteeing overall AI system safety/reliability and for enabling quality assurance.

(3) Technologies to ensure data reliability and support human decision making within a Human-centric AI society.

- Research into technologies for detecting/treating data falsification, data faking, etc.
- Research into technologies for providing data to assist humans in making prompt and appropriate decisions independently in a convincing manner.

3. Desirable research methods

The aim of this Research Area is to promote scientific innovation and value creation toward the realization of trusted quality AI technology that contributes to a Human-centric AI society and helps resolve related social issues. Research teams active in this area are to set high-impact targets themselves and promote R&D into fundamental AI technologies and AI systems that incorporate these technologies using a backcasting approach to achieve these goals.

One ultimate goal of this Research Area is to foster a community for research into trusted AI and related fields, and heighten the presence of Japanese research within such fields. Participants are thus encouraged to actively build communities not only within this field but also outside it, with PRESTO "The fundamental technologies for Trustworthy AI" that under the same strategic objectives.

4. Research period and budget

The research period is 5.5 years (from October 2021 through March 2027), and the upper limit of research budget is a total of ¥300 million per project (excluding indirect costs). If necessary, we may provide additional support to accelerate research.

Even for the JST-ANR joint proposals, the maximal budget will be allotted to Japanese side team. Please refer instruction for joint proposal for details.

5. Recommendations when applying

Research in this area is conducted under CREST team-style research. While we do present some examples themes under “2. Objectives of R&D and examples of research themes,” a team may be configured to address a single theme or multiple themes. Also, we encourage the participation of not only researchers with an established track record, but also younger researchers with some ambitious research proposition.

An important point with regards to this Research Area is that efforts are to go beyond the advancement of some discrete technological aspect of AI system reliability to also address the question of how to best configure Human-centric AI systems themselves. Thus, we ask applicants to consider how their team can be arranged to include a broad range of members in other fields/sectors/specialties having some bearing on AI if necessary; that is, we also recommend that a team include not just researchers within IT/information sectors, but also researchers active in sectors throughout the social sciences and humanities, among them ethics, law and philosophy.

When preparing your application, please elaborate what sort of impact your proposed research would have on social and academic endeavors toward the realization of trusted AI, what goal you hope to attain 5.5 years ahead, and what milestones you anticipate to reach 3 years ahead. Please be as specific as possible.

In this second Call for Applications, we would like to proceed with the selection in consideration of the balance of research themes and the portfolio of theory and implementation so that the three research and development areas shown on "Overview" can be developed as a whole of this Research Area. Please fully consider the objectives and goals of this Research Area. And we hope that many researches will be proposed that will activate this Research Area.

Please note that this Research Area is one under 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). We are contributing to the integrated operation of the AIP Project through R&D activities, while also collaborating with related research institutions including the RIKEN Center for Advanced Intelligence Project (AIP).

The AIP Network Lab conducts an AIP Challenge Program as one aspect of its efforts to further the

education and professional development of promising young researchers, including graduate students. Under this program, we provide support for young researchers affiliated with a CREST research team as they pursue individual research projects/themes.

Meetings are held at which researchers gather to announce their results. They are lively, stimulating and provide an excellent opportunity to meet others active in various fields, including young researchers, research supervisors, and research area advisors. Please encourage young researchers within your research team to participate in the AIP Challenge Program. For more information, please refer to the following link:

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