Research area in Strategic Objective *"Realization of a safe and comfortable society where "humans and AI coexist and collaborate""*

Creation of Interdisciplinary System Foundation for a Symbiotic and Collaborative Society with Humans and AI

Research supervisor: Kiyoshi Izumi (Professor, School of Engineering, The University of Tokyo)

Overview

In this research area, we aim to advance technologies that enable the symbiosis of AI and humans, as well as the collaboration of diverse AIs, and to realize the cooperation between multiple humans and multiple AIs based on these technologies, taking into consideration factors such as reliability, fairness, and safety.

As AI is expected to become more advanced and diverse in the future, while there are high expectations for multifaceted efforts towards collaboration between humans and diverse AIs, there are also concerns that the disorderly growth of advanced AI could lead to unexpected behaviors in the real world and loss of human control. Furthermore, there is a risk that crude or malicious AI could have negative impact on society. This research area addresses these expectations and challenges by integrating insights from not only information science and technology but also from humanities and social sciences such as sociology, psychology, and economics, promoting an interdisciplinary approach.

Specifically, we will advance research and development on the following challenges: (1) acquiring the technologies and insights necessary for humans and AI to coexist safely and comfortably and to grow together, (2) creating information science technologies required for diverse AIs to collaborate, leveraging their unique characteristics in both cyber and physical domains, and (3) establishing methods to integrate and evaluate theoretical frameworks and related individual technologies in virtual and real world fields for better collaboration between multiple humans and AIs. The advancement of this research area and its projects will also take into account international standards and developments in AI risk management and regulations.

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

1. Background

In recent years, artificial intelligence (AI) technologies have been advancing rapidly, significantly influencing various fields such as socioeconomics, industry, and scientific research. AI is expected to become increasingly pervasive in society and familiar not only to specialists but also to the general public. With the expansion of AI applications, these advancements will lead to a more sophisticated and diverse AI ecosystem including large-scale AI, highly functional AI, domain-specific AI, and autonomous AI. In an environment where these various types of AI and humans coexist, it will be necessary to develop a new type of AI that is capable of performing more complex and advanced tasks. However, there are also concerns about the risk of uncontrolled and disorderly growth of sophisticated AI, which could lead to unexpected behavior in the real world and loss of human control. Furthermore, it has been pointed out that shoddy or malicious AI could have a negative impact on society. To address these challenges, an interdisciplinary approach is essential, integrating information science with humanities and social sciences such as sociology, psychology, and economics. Humans and AI must complement each other and co-evolve to maximize value creation while mitigating risks. Additionally, AI risk management and regulation should be developed in line with international trends and standards. Given this background, this research area aims to enhance overall societal performance by fostering collaboration between diverse AI systems and humans while ensuring reliability, fairness, and safety.

2. Research and Development Goals and Examples of Research Topics

Based on the background, this research area aims to enhance the overall performance of society by promoting research and development toward the realization of a "Symbiotic and Collaborative Society with Humans and AI," ensuring reliability, fairness, and safety. Specific research efforts will include, but are not limited to, the following:

(1) Human-AI Symbiosis

Acquiring technologies and knowledge necessary for safe and comfortable coexistence and co-evolution between humans and AI, considering human understanding and the effects on humans, including:

- Establishment of a common understanding (common ground) infrastructure between humans and AI, which is necessary for mutual understanding between humans and AI in real world situations.

- Development of technologies to implement codes of conduct and “common sense” to govern AI behavior so that it does not pose a threat to humans.
- Methods to prevent the decline of human autonomy and thinking due to AI dependency.

(2) Collaboration among Diverse AIs

Creating information science technologies necessary for diverse AIs to collaborate by leveraging their unique features in both cyber and physical spheres including:

- Developing a platform that enables interoperability among AIs with different characteristics.
- Methods to manage and control large AI groups while protecting privacy-sensitive data.
- Technologies for achieving swarm intelligence and active information acquisition.

(3) Collaboration Among Multiple Humans and AIs

- Designing collaborative environments and social systems involving multiple humans and AIs to solve complex social issues and optimize overall social systems in real-world scenarios.
- Developing theoretical frameworks for discussions and multifaceted decision-making among humans and AIs to contribute to a better society.
- Integrating and evaluating related individual element technologies in virtual or real fields.

In addition to the above, research arising from humanities and social sciences perspectives, such as research on the introduction of AI into the real world to influence people and change their behavior through contact with people, international comparisons of the social acceptability of AI and discussions on the direction of future AI development, are also included in the scope of this category.

Related Technology Keywords: AI Agent, Generative AI, Large Language Model (LLM), Data Protection, Cyber-Physical Systems, Human-AI Collaborative Systems

3. Expected Research Approach

Combining several examples of the research topics mentioned above, it is expected that developed AIs will function as intended within a "Symbiotic and Collaborative Society with Humans and AI." Given the rapid progress in the AI field, it is possible to review and modify targets based on actual technological advancements at the time of the interim evaluation. Regarding interactions with society, collaboration with humanities and social science researchers specializing in fieldwork is considered valuable. Therefore, it is recommended to include humanities and social science researchers in the team from the application stage. Depending on the research theme, appropriate humanities and social science researchers may not be clear at the outset; hence, adding members at the Research Area Meeting for the selected projects or collaboration with PREST researchers under the same Strategic

Objectives is welcomed.

4. Research Funding and Period

The research period is up to five and a half years. Research funds (direct costs) should be requested for the amount necessary to achieve the proposal, with an upper limit of 300 million yen. Please note the research fund may be adjusted during the selection based on the evaluation by the Research Supervisor.

5. Points to Note for Application

This research area targets team type research conducted by natural science researchers centered on information science and technology. Research themes can focus on any of the challenges (1) to (3) indicated in section 2, either individually or from multiple perspectives. To thoroughly consider the societal impact of AI, it is recommended that research teams include natural science researchers and those from various fields such as cognitive science, behavioral social sciences, economics, and law.

This research area is part of the "AIP Network Lab," forming an integrated project for Artificial Intelligence/Big Data/IoT/Cybersecurity (AIP Project) by the Ministry of Education, Culture, Sports, Science, and Technology. It collaborates with related research institutions, including the RIKEN Center for Advanced Intelligence Project.