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

"Fundamental Research & Development for a Symbiotic and Collaborative Society with Human and AI"

Research Supervisor: Naomi Yamashita (Professor, Graduate School of Informatics, Kyoto University)

Overview

This research area aims to advance technologies that enable the coexistence of AI and humans, as well as the collaboration of diverse AI systems, while considering aspects such as reliability, fairness, and safety. It seeks to realize cooperative interactions between multiple humans and multiple AIs.

To achieve a society where AI and humans coexist and collaborate, it is essential not only to improve the precision of systems but also to explore the design and role of AI that contribute to the well-being of individuals and society. Furthermore, beyond addressing societal challenges through AI, there is a need to anticipate and preemptively mitigate emerging social issues arising from its widespread adoption. In this research area, with these perspectives in mind, we will promote the development of AI and its foundational technologies to realize a harmonious coexistence and collaborative society involving diverse AI systems and humans.

Additionally, elucidating the impact of AI on individuals and society is a key challenge. This includes not only research related to system development but also studies incorporating perspectives from the humanities and social sciences, as well as specialized viewpoints from fields where AI is applied, such as medicine, education, and law.

This research area participates in the Ministry of Education, Culture, Sports, Science and Technology (MEXT)'s Advanced Integrated Intelligence Platform Project on Artificial Intelligence/Big Data/IoT/Cybersecurity (AIP Project).

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

1. Background

With the rapid evolution and societal proliferation of artificial intelligence (AI)-related technologies, advanced AI agents are increasingly collaborating with humans, contributing to the resolution of complex problems, the enhancement of operational efficiency, and the promotion of creativity. However, the proliferation of diverse AI systems has also raised concerns about the risks posed by

autonomous interactions between humans and AI agents, or among AI agents themselves, which may become difficult for humans to control.

To realize a society where AI and humans coexist and collaborate harmoniously, it is essential to go beyond merely improving the accuracy of systems and to explore the design and role of AI that contributes to the well-being of individuals and society. Furthermore, in addition to addressing societal challenges through AI, there is a growing need to anticipate and preemptively mitigate the social issues that may arise from its widespread adoption.

This Research Area aims to advance the development of AI and its foundational technologies to enable a harmonious coexistence and collaboration between diverse AI systems and humans. Additionally, elucidating the impact of AI on individuals and society is considered a critical challenge. We welcome not only research related to system development but also interdisciplinary studies that incorporate perspectives from the humanities and social sciences, as well as specialized viewpoints from fields where AI is applied, such as medicine, education, and law.

2. Policies on Call and Selection

(1) Basic Policy

Against the backdrop described above, this research domain aims primarily to develop elemental technologies that address complex and multifaceted social issues toward realizing a society where diverse AI and humans coexist and collaborate. This includes developing AI capable of resolving such issues or techniques to preemptively detect and prevent social challenges arising from the widespread adoption of AI.

Furthermore, elucidating the impact of AI on individuals and society is also considered an important challenge. Beyond research focused solely on system development, we actively welcome studies incorporating perspectives from humanities and social sciences, as well as specialized viewpoints from various fields where AI is applied, such as medicine, education, law, and economics.

Specifically, the following research topics will be addressed, though proposals are not necessarily limited to these. We encourage freer and more ambitious suggestions.

(2) Examples of Research Challenges

(a) Coexistence of Humans and AI

This includes the development of elemental technologies related to system development, such as those necessary for humans and AI to coexist safely and comfortably while growing together. Additionally, research on the impact of AI on humans, as well as methods to mitigate risks of diminished human cognitive abilities due to AI dependency, is included. Furthermore, in scenarios where AI responds to human inquiries—such as in telemedicine, legal consultations, or online education—this encompasses the development of technologies to ensure the reliability of responses, as well as technologies and their theoretical foundations that enable AI to empathize with human anxieties.

(b) Collaboration Among Diverse AIs

This involves developing mechanisms for mutual coordination when multiple agents exist. It also includes the development of methods to protect the privacy and sensitive data of requesters during negotiations between agents, as well as techniques to manage or control large groups of agents. For example, this could involve creating collaborative technologies that allow autonomous vehicles to adhere to traffic rules while optimizing individual actions. Additionally, preventive measures to ensure that AIs do not independently advance discussions and cause undesirable outcomes are also considered.

(c) Collaboration Between Multiple Humans and Multiple AIs

This focuses on constructing theories or models for consensus-building in environments where multiple humans and multiple AIs (agents) coexist. For instance, this could include agent simulations using large-scale language models (LLMs). It also involves developing mechanisms to maintain the overall reliability and safety of systems through mutual learning and monitoring between humans and AI, as well as research on mechanisms that foster human trust and empathy within such systems.

Research themes (a) to (c) overlap with the content of Research Area's CREST. However, in PRESTO, we expect proposals to select one or more of these themes and focus on more foundational or elemental technological aspects. This domain also places importance on fostering young researchers and anticipates research proposals with challenging themes.

3. Research Periods and Research Funds

The research duration shall be within three and a half years. The research funding (direct costs) should be applied for in an amount necessary to achieve the proposed content, with an upper limit of 40 million yen. However, please note that, because of scrutiny by the Research Supervisor, adjustments to the research funding may be made upon adoption.

4. Policies on Research Area Management

This Research Area targets individual research by natural science researchers, primarily in the field of information science and technology. Research themes may address any single challenge from (1) to (3) outlined in Section 2, or proposals from multiple perspectives are also acceptable. Furthermore, to fully consider the societal impact of AI, we encourage exchanges of opinions and cooperation between these natural science researchers and humanities and social science researchers from diverse fields such as cognitive science, behavioral social sciences, economics, and law.

Additionally, as one of the Research Areas under the "AIP Network Laboratory," which forms part of the Ministry of Education, Culture, Sports, Science and Technology's Artificial Intelligence/Big Data/IoT/Cybersecurity Integration Project (AIP Project), this Research Area will contribute to initiatives in collaboration with related research institutions, including the RIKEN Center for Advanced Intelligence Project.