信頼される AI システムを支える基盤技術
2021 年度採択研究代表者

2021 年度
年次報告書

鹿島 久嗣

京都大学 大学院情報学研究科
教授

人と AI の協働ヒューマンコンピュテーション基盤

§1. 研究成果の概要

本研究では、信頼できる人-AI 協働系の設計論の確立を目指し、その信頼性の定義として、従来のコンピュータシステムの信頼性指標を起点としたヒューマンコンピュテーション(H/C)のための指標の定義とこれを達成するための技術開発、また、人-AI 協働系が社会で受容されるための倫理的課題の特定と解決技術の開発、さらに、H/C の活用シナリオとしての、多人数データ解析や創造的課題解決における具体的な技術的課題の解決と実践を行う。

初年度では、次年度以降の本格的な研究の開始へ向けた調査と要素技術の開発を中心に、4 グループで研究を進めた。

まず、H/C 信頼基盤技術グループでは、信頼性指標（RASIS）の定義ならびにその向上のために新規に技術開発が必要な項目を検討した。H/C タスクの種類に応じて重視すべき RASIS 項目の検討を行うとともに、機械学習・ソフトウェア工学・情報セキュリティ等の分野の技術で活用できるものを調査した。また、H/C の RASIS の実証基盤として用いるコンペティション型多人数データ解析システムの開発方式の検討を行った。

次に、H/C の社会受容グループでは、人ーAI 協働系における公平性について、主に言語データ作成における認知バイアスや社会的バイアスの影響と評価方法について検討するとともに、非専門家に対するAIの判断の説明のあり方について目的帰属型の説明を検討することで、人-AI 協働に求められる公平性及びその説明性についての検討項目を明らかにした。

そして、人間参加型機械学習グループでは、少量データから難しい学習課題を達成する要素技術開発を中心に実施した。専門家による大量の教師データ作成のボトルネックを解決するために、データの類似度評価など、少ない専門的知識を要するデータを用いて、深層学習で利用可能な表現を抽出する手法を開発した。また、タスク参加を通じて人間がスキルを成長させていくために、スキルの獲得状況を推定・提示する方法を開発した。

最後に、H/C による知的創造活動支援グループでは、学術文献データや、データ解析コンペティションのログデータ、オンラインコミュニケーションデータなど、研究の基盤となる大規模実データ収集と整備を行なった。また、意思決定における認知バイアスの気づきを獲得するための技術開発に向けて、認知バイアスが意思決定に与えうる負の影響を軽減するために、認知バイアスの気づきを人に提供する意見集約・情報提示技術を開発した。

§2. 研究実施体制

(1) 人間参加型機械学習グループ

① 研究代表者:鹿島 久嗣 (京都大学大学院情報学研究科 教授)

② 研究項目

・限られたデータから難しい学習課題を達成する人間参加型深層学習開発

・多数の人間が解析に参加するクラウドソーシング型データ解析を実現する技術開発

(2) ヒューマンコンピュテーション信頼基盤技術グループ

① 主たる共同研究者:小山 聡 (北海道大学大学院情報科学研究院研 准教授)

② 研究項目

・ヒューマンコンピュテーションの RASIS(Reliability, Availability, Serviceability, Integrity, Security)の定義

・ヒューマンコンピュテーションの RASIS の定量的評価指標の確立

・ヒューマンコンピュテーションの RASIS を実現する基盤技術の開発

(3) ヒューマンコンピュテーションの社会受容グループ

① 主たる共同研究者:荒井 ひろみ (理化学研究所革新知能統合研究センター・チームリーダー)

② 研究項目

・人-AI 協働系において求められる公平性及びその説明の明確化

・公平性およびその説明を実現する方法論開発

・人-AI 協働系におけるプライバシー保護技術の開発

(4) ヒューマンコンピュテーションによる知的創造活動支援グループ

① 主たる共同研究者:森 純一郎 (東京大学大学院情報理工学系研究科 准教授)

② 研究項目

・人の知的創造活動を支援するための実証研究

・人の意思決定における認知バイアスの気づきを獲得するための技術

【代表的な原著論文情報】

1) "Label Aggregation for Crowdsourced Triplet Similarity Comparisons", International Conference on Neural Information Processing (ICONIP), 2021.

2) "Interpretable Knowledge Tracing: Simple and Efficient Student Modeling with Causal Relations", AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI), 2022.

3) "Neural collaborative filtering with multicriteria evaluation data", Applied Soft Computing, vol. 119, 2022.

4) "Unsupervised Abstractive Opinion Summarization by Generating Sentences with Tree-

Structured Topic Guidance", Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021.
5) “Homophily に基づくサイレントマジョリティの意見推定”, 言語処理学会第 28 回年次大会, 2022.