2024 年度年次報告書
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

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

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 ユニットリーダー）
小山 聡（名古屋市立大学 データサイエンス学部 教授）
森 純一郎（東京大学 情報基盤センター 教授）

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て、従来のコンピュータシステムの信頼性指標を出発点としたヒューマンコンピュテーション（HC）における信頼性の新たな指標の定義や、それを実現するための技術開発を行ってきた。また、人-AI 協働系が社会に受容されるために必要な倫理的課題の特定とその解決手法の検討、さらには HC の応用シナリオとして、多人数によるデータ解析や創造的課題解決の現場における具体的な技術課題の解決にも取り組んできた。

4 年目となる今年度は、これまでに進めてきた HC の要素技術の開発に加え、近年急速に発展・普及している大規模言語モデル（LLM）を見据えた研究も進めた。LLM は、従来 HC において人間が担っていた多くの役割を代替できるようになりつつあり、これを前提とした新たな HC の枠組みが必要となっている。こうした状況を踏まえ、本年度は LLM の存在を前提とした HC の再設計とその発展に向けた研究を推進した。

まず、これまで本研究が取り組んで来た「HC における信頼性」という新たなコンセプトに基づく研究の土台として、HC における信頼性の定義と分類を行い、関連する既存研究の体系的整理を実施した成果をサーベイ論文としてまとめ、出版に至った。

HC 信頼基盤技術グループでは、HC の信頼性向上のための主要な技術である回答集約アルゴリズムを中心に研究を行った。 Surprisingly Popular（SP）投票は、回答者自身の意見の「珍しさ」を予測するメタ認知を活用する集約手法である。メタ認知が社会的つながりの中で形成されることに着目し、回答者がつながりのある他者の意見を参照することで、SP 投票の性能がどのように影響を受けるかを調査した。さまざまな社会ネットワーク構造を用いたシミュレーションにより、つながりの参照とネットワーク構造が SP 投票の性能やメタ認知に与える影響を明らかにした。また、近年の大規模言語モデルなど AI 技術の進展により、人間と AI が共同でクラウドソーシングを行う状況が注目されているが、回答集約の最適なアルゴリズムは明確ではない。そこで、特に AI の能力が極端に偏っていることや、人間より大量のタスクをこなせることに着目し、シミュレーションによって回答集約の重要な要素を分析した。

HC の社会受容グループでは、人−AI 協働系における公平性について、データ作成におけるバイアスについて、特に社会的バイアスの観点から研究を進め、日本語における社会的バイアスの質問応答データセットの作成を行った。さらに、多数の人間による意見を統合する際にマイノリティの意見が反映されず、それがワーカーの能力評価にも反映される問題に取り組み、グループ能力評価及び意見統合を行う方法を開発した。

人間参加型機械学習グループでは、人間と大規模言語モデル（LLM）が協調して意思決定や学習を行うための方法論の開発に取り組んだ。まず、LLM が人間の役割を代替することを念頭に置いた場合、LLM が人間と同様の性質を示す可能性を把握し、それを制御可能にすることが重要である。そのような性質の一つに「認知バイアス」がある。本研究では、LLM における認知バイアスの網羅的な調査を実施するとともに、人間の認知バイアスを軽減する既存の手法が、LLM にも有効に働くかどうかを検証した。また、LLM を評価者として用いる「LLM-as-a-judge」の枠組みに注目し、LLM が多様な意見を多様な視点から総合的に評価できる方法を開発した。具体的には、数

理的意思決定手法の一つである階層的分析法（AHP）を LLM に適用することで、複数の基準に基づく評価を効果的に行う手法を実現した。

最後に、HC による知的創造活動支援グループでは、「科学技術・教育領域における HC による知的創造活動支援に関する研究」を引き続き進めた。また、LLM の急速な発展と普及を受けて新たな設定した研究項目である「HC による LLM の安全性に関する研究」を進めた。まず、知的創造活動支援への応用として科学技術領域では、大規模な学術文献データの解析をもとに学際的な学術分野における文献のインパクトを予測するモデルを新たに開発した。さらに、教育領域では、これまでに開発してきた学習者の能力推定技術の学習支援システムへの実応用に関する取り組みを進めた。 次に、HC による LLM の安全性に関する研究について、LLM の学習に用いられる一般的なコーパスに適用可能な、汎用的な逆蒸留手法を開発した。当該研究の成果は国際会議 ICLR2025 に採択されるとともに、国内においては言語処理学会年次大会において優秀賞を含む複数の賞の受賞に至った。

【代表的な原著論文情報】

1) Hisashi Kashima, Satoshi Oyama, Hiromi Arai, Junichiro Mori. Trustworthy Human Computation: A Survey. Artificial Intelligence Review, 57(322):1-45, 2024.
2) Yu Yamashita, Yuko Sakurai, Satoshi Oyama, Masaki Onishi, Atsuyuki Morishima. Analysis of Surprisingly Popular Voting for Opinion Aggregation on Social Networks. IEEE Access, 13:23371-23383, 2025.
3) Takumi Tamura, Hiroyoshi Ito, Satoshi Oyama, Atsuyuki Morishima. Simulation-based Exploration for Aggregation Algorithms in Human+AI Crowd. In HCOMP2024 Works-in-Progress and Demos Track, 2024. (Best Work-in-Progress Paper)
4) Yasuaki Sumita, Koh Takeuchi, Hisashi Kashima. Cognitive Biases in Large Language Models: A Survey and Mitigation Experiments. In Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing (SAC), 2025.
5) Xiaotian Lu, Jiyi Li, Koh Takeuchi, Hisashi Kashima. AHP-Powered LLM Reasoning for Multi-Criteria Evaluation of Open-Ended Responses. In Findings of the Association for Computational Linguistics (EMNLP Findings), 2024.
6) Jiyi Li. Label Aggregation of Composite Crowd Tasks by Worker Ability Constraint Satisfaction. In Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI), 2025.
7) Emiko Tsutsumi, Tetsurou Nishio, Maomi Ueno. Deep-IRT with Temporal Convolutional Network for Comprehensive Reflection of Student Ability History Data. In Proceedings of the 25th International Conference on Artificial Intelligence in Education (AIED), 2024.
8) Masaru Isonuma, Ivan Titov. Unlearning Reveals the Influential Training Data of Language Models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 2024.

9) Huimin L, Masaru Isonuma, Junichiro Mori, Ichiro Sakata. UniDetox: Universal Detoxification of Large Language Models via Dataset Distillation. In Proceedings of the 13th International Conference on Learning Representations (ICLR), 2025.