

2023 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ーテーション基盤

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 大学院情報理工学系研究科 准教授)

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、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ーテーション(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目指す。

3 年度目となる今年度は、H/C における信頼性を高めるための要素技術の開発を中心に、各領域への展開へ向けて、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、H/C の信頼性の観点の 1 つであるセキュリティに関連して、不正なタスクの実行などの攻撃に対してどのようなペナルティを与えるかという問題を抽象化した、報酬とペナルティの両方を考慮した新たなセキュリティゲームを提案した。その上でペナルティの設計によって防御者および攻撃者の最適戦略がどのように変化するかを解析した。また、前年度までに開発をしたコンペティション型多人数データ解析システムを、大学におけるデータ解析の演習において実際に運用した。その過程で得られたフィードバックに基づき、環境設定の簡易化やグループ管理機能の改善、様々な評価関数の導入といった機能追加を行い、オープンソースソフトウェアとして公開した。

H/C 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た。人間が社会的なバイアスをもった判断を行っている場合に、公平なAIからのフィードバックによって、公平な判断を促す方法を提案した。また、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。また、各国のパスワード特性における機械学習を利用したパスワード推測器の評価を行い、人間が生成するデータの国ごとの傾向を考慮すると安全性に対する評価が変わることを示した。

人間参加型機械学習グループでは、AI と人間が適切な形で協力して課題を達成するための技術開発を実施した。たとえば、属性に偏りのある人間の集団の意見をまとめて結論を出す際に、属性について公平な結果を導く手法を開発した。また、自然言語処理タスクにおける AI の出力の解釈手法をクラウドソーシングによって評価・比較する方法を開発した。さらに、大規模言語モデル(LLM)による人間の代替あるいは人間との協業のための技術開発を行った。たとえば、LLM がもつ認知バイアスの調査を行い、その軽減手法についての検討を行った。また、LLM によってタスクを実施させるにあたり、LLM に対してタスクの機微な情報をできるだけ漏らさないようにこれを実施させる手法についての検討を行った。また、LLM へのプロンプトエンジニアリングをコンペティションの形で開催する際のセキュリティ問題を指摘し、これに対処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、人間参加型機械学習の基盤技術の研究および科学技術・教育領域における H/C による知的創造活動支援に関する研究を進めた。まず、人間参加型機械学習の基盤技術として、大規模言語モデルのアライメントを行うために、人間によ

って設計されたプロンプトを活用した最適なプロンプトの交互最適化技術を開発した。また、大規模言語モデルの学習データの影響を評価するため逆学習による言語モデルの解析手法を開発した。次に、知的創造活動支援への応用として科学技術領域では、学術分野のサーベイを自動的に生成するタスクとデータセットの構築を行った。また、引用には現れない潜在的なハイインパクトな研究の性質を計量書誌学の観点から定量化することに成功した。さらに、教育領域では学習者の能力の時系列変化を考慮した項目反応理論の新たな基盤技術の開発と学習支援システムへの応用を行った。

【代表的な原著論文情報】

- 1) Riku Yoshioka, Yuko Sakurai, Satoshi Oyama, Masato Shinoda. Proposing a New Security Game with Reward and Penalty. In Proceedings of the 2023 IEEE International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2023 (Best Student Paper).
- 2) Ryosuke Ueda, Koh Takeuchi, Hisashi Kashima. Fair Opinion Aggregation for Voter Attribute Bias. In Proceedings of 6th AAAI/ACM Conference on AI, Ethics, and Society (AIES), 2023.
- 3) Xiaotian Lu, Jiyi Li, Zhen Wan, Xiaofeng Lin, Koh Takeuchi, Hisashi Kashima. Evaluating Saliency Explanations in NLP by Crowdsourcing. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING), 2024.
- 4) Jiyi Li. A Comparative Study on Annotation Quality of Crowdsourcing and LLM via Label Aggregation. In Proceedings of the 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2024.
- 5) Mingzhe Yang, Hiromi Arai, Naomi Yamashita, Yukino Baba. Fair Machine Guidance to Enhance Fair Decision Making in Biased People. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI), 2024.
- 6) Nozomu Miyamoto, Masaru Isonuma, Sho Takase, Junichiro Mori, Ichiro Sakata. Dynamic Structured Neural Topic Model with Self-Attention Mechanism. In Findings of the Association for Computational Linguistics: ACL 2023 (Findings of ACL), 2023.
- 7) Masaru Isonuma, Junichiro Mori, Ichiro Sakata. Differentiable Instruction Optimization for Cross-Task Generalization. In Findings of the Association for Computational Linguistics: ACL 2023 (Findings of ACL), 2023.
- 8) Tetsu Kasanishi, Masaru Isonuma, Junichiro Mori, Ichiro Sakata. SciReviewGen: A Large-scale Dataset for Automatic Literature Review Generation. In Findings of the Association for Computational Linguistics: ACL 2023 (Findings of ACL), 2023.
- 9) 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.