研究領域「信頼されるAIシステムを支える基盤技術」中間評価（課題評価） 結果

1. 研究領域の概要

実社会での応用・実用化が急速に広がる人工知能（AI）技術は、新たな科学的・社会的・経済的価値を創出していく上で不可欠です。一方で、深層学習をはじめとする機械学習技術はブラックボックス問題やバイアス問題等の信頼性や安全性に関わる様々な課題を抱えており、その対策が喫緊の課題となっています。

そこで本研究領域は、人間が社会の中で幅広く安心して利用できる「信頼される高品質なAI」の実現につながる基盤技術の創出やそれらを活用したAIシステムの構築を行います。研究にあたっては、人間中心のAIシステムに関する信頼性や安全性等の定義や評価法の検討に取り組み、AIシステム全体としてその要求や要件を満たす技術の確立を目指します。

具体的には、以下の研究開発に取り組みます。

- (1) 「信頼されるAI」の実現に向けた発展的・革新的なAI新技術
- (2) AIシステムに社会が期待する信頼性・安全性を確保する技術
- (3) 人間中心のAI社会に向けたデータの信頼性確保及び人間の主体的な意思決定支援技術

上記により、社会的課題の解決や新たなサイエンス、価値の創造につなげるとともに、信頼されるAIに関連した新たな研究コミュニティの創成やAI研究における日本のプレゼンスの向上を目指します。

なお、本研究領域は文部科学省の人工知能/ビッグデータ/IoT/サイバーセキュリティ統合プロジェクト（AIPプロジェクト）の一環として運営します。

本研究領域は、文部科学省の選定した戦略目標「信頼されるAI」のもとに、2020年度に発足しました。

2. 中間評価の概要

2-1. 評価の目的、方法、評価項目及び基準

戦略的創造研究推進事業・CRESTにおける中間評価の目的、方法、評価項目及び基準に沿って実施した。

2-2. 評価対象研究代表者及び研究課題

2020年度採択研究課題

- (1) 伊藤 孝行（京都大学大学院情報学研究科 教授）
ハイパーデモクラシー：ソーシャルマルチエージェントに基づく大規模合意形成プラットフォームの実現
- (2) 乾 健太郎（東北大学大学院情報科学研究科 教授）
知識と推論に基づいて言語で説明できるAIシステム
- (3) 越前 功（国立情報学研究所情報社会相関研究系 教授）
インフォデミックを克服するソーシャル情報基盤技術
- (4) 後藤 真孝（産業技術総合研究所人間情報インタラクション研究部門 首席研究員）
信頼されるExplorable推薦基盤技術の実現
- (5) 森 健策（名古屋大学大学院情報学研究科 教授）
あいまい性を表現するReliable Interventional AI Robotics

2-3. 中間評価会の実施時期

2023年10月18日（水曜日）、10月25日（水曜日）、10月26日（木曜日）

2-4. 評価者

研究総括

相澤 彰子 国立情報学研究所コンテンツ科学研究系 教授

領域アドバイザー

岡田 浩之 玉川大学工学部 教授

奥村 学 東京工業大学科学技術創成研究院 教授

神寫 敏弘 産業技術総合研究所 主任研究員

佐藤 洋一 東京大学生産技術研究所 教授

辻 ゆかり 日本電信電話（株）研究開発担当役員 情報ネットワーク総合研究所 所長

福田 雅樹 大阪大学社会技術共創研究センター 教授

福水 健次 統計数理研究所数理・推論研究系 教授

村上 祐子 立教大学大学院人工知能科学研究科 教授

盛合 志帆 情報通信研究機構サイバーセキュリティ研究所 研究所長

横尾 真 九州大学大学院システム情報科学研究院 主幹教授

若宮 直紀 大阪大学大学院情報科学研究科 教授

鷺崎 弘宜 早稲田大学理工学術院基幹理工学部 教授

外部評価者

該当なし

研究課題別中間評価結果

1. 研究課題名： ハイパーデモクラシー：ソーシャルマルチエージェントに基づく大規模合意形成プラットフォームの実現
2. 研究代表者： 伊藤 孝行（京都大学大学院情報学研究科 教授）
3. 中間評価結果

大規模合意形成プラットフォーム「ハイパーデモクラシー」の研究開発について、本プロジェクトの開始後に登場した最新の大規模言語モデルを活用しながら、全般的に順調に研究を進めている。特に、エージェントシステムを介した合意形成の実証実験に積極的に取り組み、アフガニスタンの事例などで国際的にも大きな注目を集めた点は高く評価できる。研究コミュニティ形成に関しても、国際ワークショップ、学会における企画セッション、競技会の主催など様々な取り組みが行われている。若手研究者の育成・キャリアパス支援に関しても十分な実績となっている。

今後は、本研究課題が念頭に置く「より良い合意形成や集団意思決定」について AI 信頼性の観点から、その課題、客観的な評価の在り方、最終目標を明確化して研究に取り組むことが重要である。また、これまでの実証実験から得られたフィードバックをエージェント分野や社会科学分野における基礎的・学際的な研究に取り込んで、さらに先進的な成果を上げることが期待される。社会実装をさらに進めて行く上では、実証実験を体系的に計画して、効率的に検証が進められるとよい。

研究課題別中間評価結果

1. 研究課題名： 知識と推論に基づいて言語で説明できる AI システム

2. 研究代表者： 乾 健太郎（東北大学大学院情報科学研究科 教授）

3. 中間評価結果

人が説明するのと同様の仕方で、判断の過程や理由を言語で説明できる新しい AI システムの設計・実現に向けて、「知識と推論に基づく説明可能 AI」および「説明のコミュニケーション」の2つの研究項目について着実に研究を進め、学術面で優れた成果を上げている。これまでの研究活動によって、理論言語学と深層学習を融合するアルゴリズムなど新規性の高い顕著な成果が得られている一方で、2つの研究項目の連携が必ずしも十分とは言えない側面も見受けられる。

大規模言語モデルが急激に普及・拡大しつつある中、人間にとって納得がいく説明とは何かという視点に根差した説明可能 AI の構築は、ますます重要性を帯びてきている。今後は、AI 安全性などのトピックにも十分に配慮しつつ、研究グループ間や領域内での連携を強化することが強く求められる。AI を取り巻く環境の劇的な変化を踏まえつつ、各グループの知見を活かしてシナジー効果を創出するための研究マネジメントの強化をはかっていただきたい。社会還元的な展開活動は現時点ではまだないが、企業との共同研究で成果が上がり始めていることから、今後の展開に期待したい。

研究課題別中間評価結果

1. 研究課題名： インフォデミックを克服するソーシャル情報基盤技術
2. 研究代表者： 越前 功（国立情報学研究所情報社会相関研究系 教授）
3. 中間評価結果

フェイクメディアの生成技術、検出・防御技術、無毒化技術、インフォデミックを緩和して多様な意思決定を支援する情報技術等について、優れた研究成果を論文として発表するとともに、ベンチマークセットを公開するなど研究コミュニティに大きく貢献している。また、多くのメディア掲載や国内外での招待講演や書籍出版等を通して、研究成果を効果的に発信している。さらに、フェイク顔画像を自動検出する基本特許を押さえ、この基本特許に基づく真贋判定プログラム SYNTHETIQ VISION を国内企業に有償ライセンスし、国内初の実サービスにつなげるなど、実用化及び社会還元的な展開活動も顕著に進められている。若手研究者の研究費獲得、受賞、キャリアアップなどにも力を入れている。

フェイクメディアという AI の信頼性・安全性に関する重要技術に関して、本 CREST を契機に他省庁や産業界などから注目を集め、他の国プロ立ち上げにつなげる等、同研究分野をけん引して大きなムーブメントを創出しており、さらなる発展が期待される。今後は社会実装を進める中でグループ間の連携を強化しつつ、世界の動向に素早く対応した研究開発を継続して行っていただきたい。

研究課題別中間評価結果

1. 研究課題名： 信頼される Explorable 推薦基盤技術の実現
2. 研究代表者： 後藤 真孝 (産業技術総合研究所人間情報インタラクション研究部門 首席研究員)
3. 中間評価結果

Explorable な推薦基盤技術という新たな枠組みを提唱し、推薦誘導、推薦過信、推薦束縛の3つの問題を整理した上で、音楽メディアの鑑賞と創作の両面から研究に取り組み、優れた成果を上げている。情報学、神経生理学、社会心理学の研究者によるグループ間の連携も十分取られている。また、実証面でも推薦の挙動を探索・説明可能な音楽推薦システムを開発し、企業と連携して実運用を開始、多数のコラボ企画を実施している。「信頼される AI」を探求する上で、研究→開発→社会実装をアジャイルで回す本プロジェクトの手法は、新たな取り組み方として高く評価できる。

本領域が目標とする AI の信頼性に関し、従来から広く研究されてきた「説明可能性」に対して相補的な役割を果たす「探索可能性」の構想と具体化を目指しており、科学イノベーションに寄与する着実な成果が見込める。また、実用化および社会還元的な展開活動においては、利用者からのフィードバックに基づく新たな推薦サービスの開拓も期待される。グループ間の連携により、音楽嗜好推定のための生理・脳機能計測に関する基礎研究や推薦受容傾向の概念分析および尺度に関する研究をさらに展開していただきたい。

研究課題別中間評価結果

1. 研究課題名： あいまい性を表現する Reliable Interventional AI Robotics

2. 研究代表者： 森 健策（名古屋大学大学院情報学研究科 教授）

3. 中間評価結果

あいまい性を表現可能な AI 治療ロボットの技術確立に向けて、あいまい性を可視化して作業指示や作業支援を行うための医療画像処理の基礎的な研究および医療現場への展開を想定したロボットシステムの開発を進めている。医療分野への AI システムの適用という社会的ニーズの大きい分野に挑戦して、着実に優れた学術的成果を上げている点は評価できる。また応用面でも、内視鏡手術支援や微細操作支援の AI ロボットで実証を進めており、製品化も視野に入れるなどインパクトのある技術的成果を上げている。若手研究者の育成・キャリアパス支援も順調である。

今後は、最終目標である「Reliable Interventional AI 基盤技術」を確立し、医師によって信頼される医療 AI ロボットを実現するために、さらに研究を加速することが望まれる。これによって、既存の医用画像解析手法や手術ロボットシステムにはない、本研究課題ならではの独創性をさらに追求していただきたい。このために、あいまい性を AI への信頼性につなげる方向性を意識してグループ間連携を強化するとともに、評価手法や評価尺度を検討して AI 信頼性に関する実証を具体化することが求められる。