CREST「信頼される AI システムを支える基盤技術」
研究領域中間評価報告書

1. 研究領域としての成果について

(1)研究領域としての研究マネジメントの状況

本研究領域は、戦略目標「信頼される AI」の下、「AI を信頼する主体としての人間」を AI 技術の研究開発の中核とする、新しい AI 研究の在り方を考究し、社会的課題の解決、新たなサイエンス、価値の創造につなげることをねらいとした。これは、内閣府の統合イノベーション戦略推進会議が定める人間中心の AI 社会原則の理念に向けた、AI 技術開発の具体的な道筋を示すためのものと位置づけられる。

本研究領域では、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、評価法の検討と、それに立脚した基盤技術の確立、および社会実装を目指した。本研究領域での「信頼」をキーワードに、実世界環境とのインタラクション、脳情報処理、認知発達過程を組み込んだ技術開発と社会受容性の両面を重視した取り組みは非常に意義深い。

研究課題は、2020～2022 年度で 12 件を採択した。選考においては、研究総括が新たに出口から見て設定した 4 つの観点(①社会課題解決、②人間 AI 協調、③原理解明、④基盤構築)でバランスよく採択している。また、若手研究者もグループリーダとして多く参加しており、AI 分野の人材育成にも貢献している。

領域アドバイザーは、実世界システムの幅広い技術分野と、情報通信法や情報哲学など社会科学の人材をバランスよく人選している。AI・機械学習の品質保証や、AI 研究開発倫理、因果推論の専門家からのアドバイスを得る機会を設定するとさらによいと思われる。また、信頼という観点から、関連する社会学者や、社会を技術で変えようとしている政策実務者、企業研究者などのアドバイスも有効と考える。

本研究領域の運営では、予想を超えた生成 AI の発展に対応し、いくつかの研究課題へは、積極的にこの技術を取り入れ強化するよう方向づけを行ったり、軌道修正とともにより難度の高い課題に取り組むよう指導したりするなど迅速な計画見直しを図った。

本研究領域は、「AI を信頼する主体としての人間」を AI 技術の研究開発の中核とする新しい難解な試みである。そのため、信頼する側の人間と、信頼される側の AI を区別せずに AI の性能追求にならないように、また、人間は主観的判断で裏づけのない場合でも信頼することから、これと AI の品質の信頼との間にはずれがあることを意識づけた。さらに、各研究課題の「信頼される AI」の定義と実現について、領域横断で討議を行い認識の共有化を図った。定義自体が幅広く難しい分野の中で、考え方の整理や共通認識の醸成、研究の方向性の指導を強く進めたと認められる。しかしながら、AI の品質への信頼を向上すれば、信頼できる AI になると捉えられるところも、未だ若干あるため、継続したリードを望む。また、研究期間後半でも当初計画に捉われることなく、技術や潮流の急激な変化に応じて柔

軟に対応し、継続的に変更していくことを望む。

領域会議での研究進捗把握は適切であり、加えて参加者アンケートなどの工夫で次回の領域会議の立案や運営の見直しをきめ細かく行っている。

各チーム持ち回りで各研究課題の基礎知識を参加者に紹介する異分野融合・連携のための領域セミナー(計 12 回開催)も良い試みである。予算配分も各研究課題のインセンティブになるよう工夫し実施している。

アウトリーチについては、文部科学省 AIP プロジェクトの一環として JST が実施している AIP ネットワークラボに参加している他の研究領域と連携し、独自の 2nd International Workshop on IAA(Intelligence Augmentation and Amplification)2022、JST-ERCIM (European Research Consortium on Informatics and Mathematics)ワークショップ(2021～2024 年の毎年)、仏コート・ダジュール大学における特別企画ワークショップ(2024 年)などの企画や参画で AI・IoT・サイバーセキュリティ分野の国際的ネットワーク構築に大きく貢献している。

若手研究者育成では、AIP ネットワークラボのチャレンジプログラムへの応募推奨を行った。2021～2024 年で 47 件と多く採択されている。また、キャリアパス支援のため、昇進、受賞、外部研究予算の獲得のための協力・指導に力を入れており、多くの若手研究員が昇進や上位職への異動を実現している。

コロナ禍での研究領域発足という背景もあるが、CREST の趣旨からも重要な領域内連携の推進が必ずしも十分でないように思われる。領域内の技術を共有することは領域の研究を推進する源となるため、そのポテンシャルを活かすことを望む。サイトビジットもコロナ禍の延長でリモートとなっているが、参加者の出席率と、現地開催のメリットを計り適宜開催の仕方を見直していくことも期待する。また、さらなるプレゼンスの強化のための著名な国際会議でのワークショップの企画・参加、米国ビッグテック、AI Alliance との連携や、本分野のキーとなる社会受容性強化のための JST の RISTEX(社会技術研究開発センター)の関連プログラムなどとの連携も視野に入れることを期待する。

(2)研究領域としての戦略目標の達成に向けた状況

国際的に見ても高い水準の研究成果が生まれており、著名な論文誌、国際会議で多く採択されていることから十分な内容と評価できる。

論文(査読あり、学術雑誌・会議録)は国際 530 報、国内 46 報、招待講演は国際 103 件、国内 196 件ある。特許出願は国際 2 件、国内 4 件、受賞は 128 件、メディア掲載やプレスリリースは 456 件、ワークショップは 151 件あり、社会実装に向けて重要なソフトウェア、ライブラリ、データベースは、ほぼ全てのチームで GitHub、Web サイト等で公開している。

特筆すべき研究成果の例としては、以下のようなものがある。

伊藤チームのハイパーデモクラシーのための合意形成プラットフォームは「信頼される AI」のまさに的を射た研究トピックである。本研究は Google DeepMind など AI 研究で著名

な研究チームの最新論文にも引用されている。本技術は、アフガニスタンで市民に利用され大きな話題になった。また、研究成果活用企業であるスタートアップ AGREEBIT 株式会社において、自治体、教育機関、企業などで利用が拡大している。

乾チームの大規模言語モデル(LLM)の内部表現と説明可能性に関する理論研究は、国際的なプレゼンスが高い研究成果を創出している。

越前チームの AI で生成されたフェイクメディアによる攻撃を検出・防御し、サイバー空間での人間の免疫力を高めるソーシャル情報基盤技術は、国際的に顕著な学術的成果に加え、メディア掲載 204 件、招待・依頼講演 58 件、書籍著者・解説記事 35 件など注目を集め、本分野の潮流を創出している。また、自動真贋判定プログラム SYNTHETIQ VISION を開発し、数社の国内企業にライセンスし国内初のサービスにつなげている。

後藤チームの信頼される音楽推薦基盤は、情報学、神経生理学、社会心理学を融合した学際的な研究である。これにより音楽推薦の内部状態を可視化した世界初の音楽発掘サービス「Kiite」などを開発・一般公開し実証している。受賞 12 件、メディア掲載 228 件と社会の注目度が高い。

上記の他にも独自の優れた研究成果をあげており評価に値する。実世界システム、人間の認知システムを扱った研究成果についても今後の進展を期待する。

「信頼される AI」は、まさに今の社会の喫緊の課題であり、かつ政府政策へも影響が大きい。社会の耳目を惹く研究領域名のため、重要性、影響に注意しながら、研究者のみならず政策実務者、一般社会へ丁寧にメッセージを発信することを望む。その際、「信頼される AI」をどう作るか、得られた技術の効能とともに、その信頼の定義、使われる前提条件や、保証できる範囲、利用における注意点などを明確にし、発信することを望む。

以上を総括し、本研究領域は優れていると評価する。

以上