〇戦略目標「信頼される AI」の下の研究領域

信頼される AI の基盤技術

研究総括：有村 博紀　（北海道大学 大学院情報科学研究院　教授）

研究領域の概要

ネットワークやビッグデータ等の情報環境の広がりと、数理科学と情報技術の急速な発展によって、人工知能(AI)技術を用いたシステムやサービスが社会に広がりつつあります。このような AI 技術の利活用により、あらゆる人々が適切で高品質なサービスを受け、社会と調和しつつ個人の能力を発揮して暮らしていける人間中心の AI 社会の実現が期待されています。その一方で、人と共に社会の重要なタスクをこなす「信頼される AI システム」の実現において、深層学習に代表される現在の AI 技術には、説明性や納得性、安定性、公平性等に関するさまざまな弱点や限界があることが判明してきました。また、AI 技術を組み込んだいわゆる AI システム全体やデータの信頼性・安全性・品質保証に関して、さらに、人間を基点として社会と調和した AI の利活用に関する方策も必要です。

本研究領域では、人間中心の AI 社会の実現に向け、現在の AI 技術の限界を突破する次世代 AI 技術の基盤となる革新的な理論・技術の創出を目指します。従来の AI 技術の単なる延長ではなく、現在の AI 技術や AI システムが持つ本質的な問題点に取組み、解くべき問題を新たな視点で概念化・定式化し、その解決を目指す挑戦的な研究を推進します。

具体的には、1）現在の AI 技術の弱点や困難を克服するための新しい数理・計算・解析手法に関する基礎技術や、2）AI システムの信頼性・頑健性・透明性・公平性等、社会における新たな AI 応用タスクの概念化・定式化と新しい構成原理・実現技術、3）これらを支えるデータや情報基盤の信頼性・安全性・プライバシーの保証技術、4）多様なデータやタスクに対する AI 技術の拡張、5）AI システムの設計・開発・運用の方法論、等の研究に取り組みます。

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。

募集・選考・領域運営にあたっての研究総括の方針

１．背景

機械学習等の中核技術の急速な発展によって、人工知能(AI)技術を用いたシステムやサービスの利活用が広がりつつあります。その一方で、人間中心の AI 社会の実現には、現在

のAI技術にはさまざまな弱点や限界があることが判明してきました。そのため、AI技術を組み込んだ情報システムやサービス（いわゆる「AIシステム」）全体の信頼性・安全性・品質を保証することや、人間を基点として社会と調和したAIの利活用に関する方策も必要です。人間と共に社会の重要なタスクをこなす「信頼される AI」を実現するためには、これらの技術的な弱点や限界を克服して行く必要があります。

例えば、現在のAIの中核技術の一つである深層学習（ディープラーニング）については、理論的には原理が未解明な部分も多く、一種のブラックボックスとして用いられています。そのため、予測結果の説明性や、納得性、透明性の担保が不十分であることや、データに含まれている差別や偏見を学習してしまう公平性の問題、未知・想定外ケースや環境変化に対するぜい弱性、文脈や常識の理解と適応ができていない等の弱点を持つことが指摘されています。また、AI システムについては、既存のソフトウェア工学等の方法論ではシステム全体の信頼性や安全性、品質を保証することができないため、それらを担保するための新たな原理や技術が必要とされています。データ自体についても、贋データや改ざんといった信頼性の問題や、これらの贋データの作成・流通や改ざんにAIが悪用されるといった問題も指摘されています。さらに、これらの問題に対して、後付けでなく、設計の段階からシステムの必要要件を満足させるための「バイ・デザイン」に基づく設計や構築も必要です。

２．研究開発の目標と研究課題の例

上記のような課題を解決し、AI技術がより広く社会で活用されるためには、現在のAI技術の単なる延長ではなく、現在のAI技術の問題点・弱点・限界を明らかにし、情報科学・数理科学等の関連諸分野のこれまでの成果を飛躍的に発展させて、新たな概念や方法論を創出し、問題の根本的な解決をはかることが必要です。そのため本研究領域は、若手研究者による次世代AI技術とその支援技術に関する自由で新しい発想に基づいた挑戦的な研究課題を推進し、これらを通して、人間中心のAI社会の実現に寄与する「信頼されるAI」の基盤となる革新的な理論・技術の創出を目指します。以下に、研究課題のサンプルを示します。

（１）現在のAI技術の弱点の克服

ア　深層学習に代表される現在の AI 技術の原理や、可能性・限界の解明と、それらに立脚した新しいAIの構成原理や技術の提案

イ　AI 技術の社会への浸透にともなって生ずる新しいタスクや多様なデータを対象とし、新しい数理モデルや、アルゴリズム、システムの提案、それらに伴う数理・情報技術・社会応用の困難性を解決する研究

ウ　深層学習のような帰納的な情報処理と、記号推論や知識処理、シミュレーション等の演繹的な情報処理を融合したAIを実現するための研究

エ　入出力関係の高精度な予測にとどまらず、因果関係の発見や、反事実説明、統計的に有意な予測等の高次の機械学習タスクに関わるAIの研究

オ　人間の脳情報処理や認知過程等に関する知見に基づく新しい AI 原理の研究

カ　上記ア～オについて、現在の適用先だけではなく、AI 技術が社会に浸透することで新たに生じると想定される適用先特有の課題についても解決できる研究を期待します。

（2）AI 技術が導入されたシステム・サービスの信頼性・安全性を確保するための研究

ア　AI 技術を中核とする情報システム（以降、AI システムと呼ぶ）の設計・開発・運用に関わる新しい数理と方法論、技術の研究

イ　未知・想定外ケースや環境変化、データのバイアス、悪意ある攻撃等に対して頑健な AI システム・サービスを実現するための研究

ウ　AI システム・サービスの品質評価、安全設計に関する研究

エ　AI アルゴリズムやシステムの信頼性や透明性、安全性、プライバシー、公平性といった、人間や社会との協働において望まれる性質を設計段階から保証（「バイ・デザイン」）可能にする新しい方法論

（3）AI システムのためのデータの信頼性保証や、AI 技術を用いた意思決定等に関する研究

ア　社会における多数の参加者・利害関係者が参加するサービスや制度において、納得性や公平性などの望ましい性質をもった意思決定や合意形成を実現するための新しい数理モデルの提案や、アルゴリズム・方式・制度の研究

イ　AI システムが利用するデータの収集・管理やシステムの結果の活用において、フェイク、データ改ざん等の不正（AI によるものを含む）を検知し、対処する技術の研究

ウ　上記について、従来の研究トピックにとどまらず、既存の分野を横断し、将来の人間中心の AI 社会の基盤技術として革新をもたらすような、新しい概念・数理・情報技術・社会実装の研究

上記の研究課題例に限られることなく、数理・情報技術・社会応用等のさまざまな立場から、人間中心の AI 社会と、そのための信頼される AI の実現に資する、新しい発想に基づいた挑戦的な研究構想を求めます。

３．想定する研究の進め方

本研究領域が目指す人間中心の AI 社会に寄与する信頼される AI 技術の実現には、さまざまなアプローチが考えられます。そのため、AI 技術の中核技術の分野だけでなく、AI 技術と AI システムの数理・情報技術・社会応用に関連する多様な分野からの研究提案を期待します。研究推進にあたっては、年間 2 回を原則とする領域全体会議等を通して、異なる専

門分野の若手研究者同士が交流し、相互に触発する場を設けることで、新しいアイデアの創発や、未来に貢献する先端研究を推進する研究人材の育成、将来の連携につながる研究者の人的ネットワーク構築をはかります。さらに、本研究領域の研究目標を達成するには、研究を進めていく上で狭義の科学・技術以外の観点も必要となります。そのため、領域に関連する人文社会科学等の関連学問分野や、企業、行政等の多様なステークホルダー（関係者）との交流や連携を推奨し、その機会を領域で設ける予定です。ただし、これらへの参加については、参加研究者自身の特性や研究形態に応じて、柔軟にご判断いただければと思います。

４．研究期間と研究費

研究期間は 3 年半以内、予算規模は、総額 4,000 万円（間接経費を除く）を上限とします。

５．応募にあたっての留意点

数理・情報技術・社会応用等の多様な立場から、人間中心の AI 社会の実現に資する新しい発想に基づいた独創的かつ挑戦的な研究構想を求めます。本研究領域では、達成が容易でなくても真にインパクトの大きい研究を実施していただきたいと考えています。したがって、失敗を恐れず、さきがけでなければできないような、独創的アイデアに基づいた挑戦的かつ革新的な研究提案をしてください。申請では、従来の研究や技術の単なる延長ではなく、現在の AI 技術または AI システムの問題点や課題がどこにあるか、それをどのように解決したいのか、そのためにどのような新しい概念や方法を作り出したいのかについて、申請者ご自身の言葉で語っていただくことを期待します。また、ご自身の目標の実現に際して、どのような領域参加者とどのような連携をしたいかについても自由に記載いただけると良いと思います。

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）を構成する「AIP ネットワークラボ」の 1 研究領域として、CREST「信頼される AI システムを支える基盤技術」や理化学研究所革新知能統合研究センターをはじめとした関係研究機関等と連携しつつ研究課題に取り組むなど、AIP プロジェクトの一体的な運営にも貢献していきます。詳細は以下をご覧ください。

https://www.jst.go.jp/kisoken/aip/index.html