〇戦略目標「信頼される AI」の下の研究領域

信頼される AI システムを支える基盤技術

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）

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。

募集・選考・領域運営にあたっての研究総括の方針

１．背景

近年、機械学習等の AI 技術が著しく発展して様々なシステムやサービスに活用されていますが、一方でその信頼性や安全性等については懸念も示されています。AI 技術そのもの、特にその中心技術である深層学習（ディープラーニング）については、出力結果の説明性や納得性が不十分である、データに含まれているバイアスを学習してしまう、未知・想定外ケースや環境変化に対してぜい弱である、文脈や常識の理解ができていない等の弱点が挙げられています。また AI 技術が応用されたシステムやサービスについては、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、新たな方法論が必要である等の指摘があります。そしてデータ自体についても、フェイクの流通や改ざんの恐れがあり、加えて、これらフェイクの作成・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。

２．研究開発の目標と研究課題の例

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ための技術の確立を目指します。

具体的な研究課題の例を以下に示します。ただし、募集課題はこれに限りません。

（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

ア　深層学習のような帰納的な処理と知識・言語による推論・プランニング等の演繹的な処理を最適に融合させたAI技術の研究

イ　大量教師データが与えられなくても、実世界環境との相互作用を通して、知識獲得・成長するAI技術の研究

ウ　人間の脳情報処理や認知発達過程に関する知見に基づく新しいAI原理の研究

（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

ア　判断・推論の根拠を説明できるAIシステムを実現するための技術の研究

イ　データ拡張やデータバイアス除去やデータ匿名化などデータを加工する技術の研究

ウ　未知・想定外ケースや環境変化にも頑健な AI システムを実現するための技術の研究

エ　AIシステム全体の信頼性・安全性の確保、品質保証を可能とする技術の研究

（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

ア　データ改ざんやねつ造（フェイク）等を検知し対処する技術の研究

イ　人間が主体性・納得感を持って、適切かつ迅速に判断を下したり合意を形成したりすることを支援する技術の研究

３．想定する研究の進め方

本研究領域は人間中心の AI 社会に資する信頼される高品質な AI 技術の実現に向けて、新たなサイエンスや価値を創造して、社会的課題を解決することを目指します。領域に参加する研究チームはハイインパクトな研究成果に関する目標を自ら設定して、バックキャス

ト的にその目標を達成するための AI 基盤技術やそれを活用した AI システム構築に関する研究開発を推進します。

また、本研究領域では最終的なゴールの一つとして信頼される 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。そのため、領域内のみならず、同じ戦略目標の下に実施する、さきがけ「信頼される AI の基盤技術」を始めとする、領域外との連携によるコミュニティ作りを積極的に行うことを推奨します。

４．研究期間と研究費

研究期間は約 5.5 年間（2021 年 10 月から 2027 年 3 月末まで）とします。研究期間全体における研究費は 3 億円(間接経費を除く)を上限とします。必要に応じて研究加速等の支援を行います。

フランス ANR との共同提案においても、ANR 側の予算規模にかかわらず CREST の基準で応募してください。

５．応募にあたっての留意点

本研究領域では、チーム型研究の「CREST」として運営します。本研究領域の研究課題を２で例示しましたが、チーム構成としては、一つの課題の解決を目指す構成でも、複数の課題の解決を目指す構成でも構いません。また、実績のある研究者のみならず、若手研究者のチャレンジングな研究提案も推奨します。

本研究領域では、AI システムの信頼性に関する個別の要素技術の発展のみならず、人間中心の信頼される AI システムをどう構築するかが研究の重要なポイントとなります。そのため、情報学分野の研究者だけでなく、倫理、法律および哲学を含む人文・社会科学系の研究者や AI システムのエンドユーザとなる産業界の関係者など、AI に関わる様々な分野・セクターを巻き込んだ幅広いメンバー構成を必要に応じて検討してください。

また、本研究領域の応募にあたっては、提案研究が信頼される AI の実現に向けて、学術や社会にどのようなインパクトを与えようとするものであるかを明らかにするとともに、5.5 年間での達成目標および、3 年後のマイルストーンについてできるだけ具体的に記載してください。

２回目となる本年度の公募においては、領域全体として「概要」にある３つの研究開発項目を展開できるよう、研究テーマのバランスや理論と実装のポートフォリオも考慮しながら選考を進めたいと考えています。本領域の趣旨、目標を十分に勘案していただき、領域全体が活性化されるような研究が数多く提案されることを期待しています。

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、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。

AIP ネットワークラボでは、大学院生を含む若手研究者の育成・教育を目的とした取組みの１つとして、「AIP チャレンジプログラム」を実施しています。このプログラムは、CREST 研究チームに所属する若手研究者に CREST 課題に資する独自のテーマでの個人研究を支援するものです。成果報告会では、他領域の若手研究者や研究総括、領域アドバイザーと交流し、様々な刺激を受ける良い機会となっています。是非、研究チームに若手研究者を呼び込み、この「AIP チャレンジプログラム」への参加を促してください。詳細は以下をご覧ください。

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