

○戦略目標「安全かつ快適な“人と AI の共生・協働社会”の実現」の下の研究領域

人と AI の共生・協働社会を実現する学際的システム基盤の創出

研究総括：和泉 潔 （東京大学 大学院工学系研究科 教授）

研究領域の概要

本研究領域では、信頼性・公平性・安全性等を考慮しながら、AI と人の共生や多様な AI の連携を可能とする技術の発展及びそれらを基にした複数の人と複数の AI による協働の実現を目指します。

今後、AI がさらに高度化・多様化していくと見込まれる中で、人と多様な AI の協働に向けた多角的な取り組みが期待されている一方、高度化した AI が無秩序に乱立することで、実世界における予期しない振る舞いや人間の制御が及ばなくなるリスクや粗悪な AI や悪意のある AI が社会に悪影響を及ぼす可能性も指摘されています。本研究領域では、これらの期待や課題に対応するために、情報科学技術の知見だけでなく、社会学、心理学、経済学といった人文・社会科学の知見を高度に融合し、学際的なアプローチによる取り組みを進めます。

具体的には、（１）人と AI が安全かつ快適に共生し、ともに成長するために必要となる技術・知見の獲得、（２）多様な AI 同士それぞれの特徴を活かしてサイバー・フィジカル両面で協働するために必要となる情報科学技術の創出、（３）複数の人と AI が協働することでよりよい社会につなげるための理論体系や関係する個別要素技術を統合して仮想空間や実フィールドにおいて評価する手法の構築、等の課題に関わる研究開発を推進します。領域および研究課題の推進にあたっては、AI のリスク管理や規制のあり方に関する国際的な標準や展開も見据える等、国際的な動向を踏まえて取り組みます。

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）に参画します。

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、人工知能（AI）関連技術は加速度的に発展し、社会経済、産業、科学研究など多岐にわたる分野に大きな影響を与えています。今後、AI はますます社会に浸透し、専門家に限らず一般の人々にとっても身近な存在となることが予想されます。また、AI の活用範囲の拡大に伴い、大規模かつ高機能な AI モデルが登場するだけでなく、特定分野・用途に特化した「領域特化型 AI」や自律的に学習・判断を行う「自律型 AI」も容易に開発できるよ

うになり、今後さらに高度化・多様化していくと考えられます。

こうした多様な AI と人間が共存する環境下では、より複雑で高度なタスクの遂行を可能とするため、AI 同士の連携や人と AI の協働に向けた多角的な視点の取り組みが期待されています。しかし、高度化した AI が無秩序に乱立することで、実世界において予期しない振る舞いが発生したり、人間の制御が及ばなくなったりするリスクも懸念されます。さらに、粗悪な AI や悪意のある AI が社会に悪影響を及ぼす可能性も指摘されています。

これらの課題に対応するためには、情報科学技術の知見だけでなく、社会学、心理学、経済学といった人文・社会科学の知見を高度に融合し、学際的なアプローチを推進することが不可欠です。人間と AI が相互に補完しあいながら共進化し、リスクを抑制しつつ価値創出の最大化を図ることが求められます。あわせて、AI のリスク管理や規制のあり方については、国際的な標準や展開も見据え、国際的な動向を踏まえた検討を進めることが望ましいと言えます。

このような背景のもと、本研究領域では、信頼性・公平性・安全性等を考慮しながら、多様な AI と人間が連携し、社会全体のパフォーマンスを向上させることを目指し、人々が安全かつ快適に暮らせる「人と AI の共生・協働社会」の実現に向けた研究開発を促進します。

2. 研究開発の目標と研究課題の例

上記の背景を踏まえ、本研究領域では、信頼性・公平性・安全性等を考慮しながら、多様な AI と人間が連携し、社会全体のパフォーマンス向上を図ることを目指し、人々が安全かつ快適に暮らせる「人と AI の共生・協働社会」の実現に向けた研究開発を推進します。具体的には以下のような研究に取り組めますが、必ずしもこれらに限定するものではなく、より自由で挑戦的な提案を期待します。

(1) 人と AI の共生

現実世界の場面において、人と AI の相互理解のために必要となる、人と AI 間の共通理解（コモングラウンド）基盤の確立、および AI の論理的推論・説明技術の開発。人の脅威とならないように AI の行動を律するための行動規範や“常識”の実装技術の開発。AI 依存による人間の主体性・思考低下リスクの回避方法など、人間に対する理解や人間に与える影響を踏まえ、人と AI が安全かつ快適に共生し、共進化するために必要な技術・知見の獲得。

(2) 多様な AI 間の連携

性質の異なる AI 同士が相互運用できる協調プラットフォームの開発、プライバシー・センシティブデータを保護しながら大規模 AI 群を管理・制御する手法の開発。群知能や能動的情報取得を実現するための技術など、多様な AI がそれぞれの特徴を活かしてサイバー・フィジカル両面で協働するために必要となる情報科学技術の創出。

(3) 複数の人と複数の AI による協働

- ・ 実社会を想定し、より高度な社会課題の解決や社会システムの全体最適化を見据えた、複数の人と複数の AI による協働環境・社会デザイン
- ・ 認知バイアスを持ち得る人間と AI の討議・多角的意思決定など、複数の人と AI が協働することで、よりよい社会につなげるための理論体系
- ・ 関係する個別要素技術を統合し、仮想空間あるいは実フィールドにおいて評価する手法の構築など

これらの他にも、AI が実社会に導入され、人との接触を経て人に影響を与え、行動変容を図るための研究、AI の社会受容性の国際比較と今後の AI の開発の方向性に関する議論など、人文・社会科学的な視点から生じる研究も範疇に含まれる。

(関連する技術キーワード)

AI エージェント、生成 AI、大規模言語モデル (LLM)、データ保護、サイバー／フィジカルシステム、人と AI の協働システム

3. 想定する研究の進め方

上述の研究課題例からいくつかを組み合わせて、「人と AI が共存し協働する」社会で開発した AI がシステムとして想定通り機能することを期待します。なお、AI の分野は極めて速く進歩が進むため、中間評価の時点で適宜目標を見直し、実際の技術進展に即した内容に変更することも可能とします。

また、社会との関わりについては、フィールドワークを主体とする人文・社会科学系の研究者との協力が有用と考えられることから、応募時点から人文・社会科学系の研究者をチームに加えることを推奨します。研究テーマによっては、どういう分野の人文・社会科学系研究者を加えるか、明確でないこともあり得るため、初年度採択後に実施される領域会議等でのメンバー追加、同目標下のさきがけ研究者との協業も歓迎します。

4. 研究費と研究期間

研究期間は5年半以内とします。研究費（直接経費）は提案内容の達成に必要な額を申請することとし、上限は3億円とします。ただし、総括による精査の結果、採択にあたっては研究費の調整を行う場合がありますので、予めご了承ください。

5. 応募にあたっての留意点

本研究領域では、情報科学技術を中心とした自然科学系の研究者によるチーム型研究を対象とします。研究テーマは、2. で示した（1）～（3）の課題のうち、どれか単独、あるいは複数の観点からの研究提案でも構いません。また、AI の社会への影響について十分に考慮するため、これら自然科学系研究者と、認知科学、行動社会科学、経済学、法学など多様な分野の人文・社会系研究者を加えたチームの研究も推奨します。

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）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た取り組みなどにも貢献していきます。