安全かつ快適な”人とAIの共生・協働社会”の実現

アライメントに配慮した、多様な人と多様なAIの連携の推進

趣旨

- AIが急速に社会に普及する中、**進化したAIエージェントによる、他のAIや人間との協働への期待**が高まっている。
- 他方、AIが高度化・多様化することで、**人間がAIをコントロールできなくなる等の懸念**がある。

多様なエージェントが連携し、リスクを抑制しつつ価値創出を最大化するための研究開発を促進

達成目標

信頼性・公平性・安全性などを考慮しながら、
以下の実現を目指す。

1. **人とAIの共生**：知識や意図の共有技術等
2. **多様なAIの連携**：相互運用プラットフォーム等
3. **複数の人と複数のAI協働**：社会デザイン等

期待

- 複雑な社会課題の解決、社会システムの全体最適化
- 多角的な議論や合意形成
- 新たなビジネスを生むエージェント経済圏　等

懸念

- AIのシステミックリスク、AIの暴走
- 有害なAI、ブラックボックス問題
- AIが持つ影響力の強さと社会的受容性　等

将来像

公平性や多様性に配慮した包摂的な社会の実現

Well-beingや人間理解の深化と人に寄り添ったAI技術の進化

生産性向上や労働力不足の解消、新産業の創出、国際的プレゼンスの向上

令和 7 年度戦略目標

1. 目標名

安全かつ快適な“人と AI の共生・協働社会”の実現

2. 概要

人工知能（AI）関連技術が目覚ましく進化し社会に普及する中で、高機能かつ多様な特徴を有する AI エージェントが、他の AI や人間と協力してより複雑・複合的な社会課題を解決することが期待されている。他方、性質の異なる AI が乱立し、それぞれの目的のために自律的にインタラクションを行うことで、人間が予期せぬ AI の振舞いが生じるといった懸念もある。

こうした背景を踏まえ、本戦略目標では、信頼性・公平性・安全性等を考慮しながら多様な AI 及び人間が連携し、社会全体のパフォーマンスを高め、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する。

3. 趣旨

生成 AI をはじめ、AI 技術は近年加速度的に発展しており、社会や産業、科学研究等に大きな影響を与えているところ、今後ますます AI が社会に浸透し、専門家でない人にとっても身近な存在となることが予想される。また、その活用範囲の拡大に伴い、モデルの大規模化・高機能化が進むことに加え、特定の分野・用途向けあるいは組織・個人向けのカスタム AI も容易に作れるようになっており、AI は高度化・多様化することが想定される。

このように、多様な AI・人間が共存する環境となるにつれ、多角的な視点で、より複雑・高度なタスクを遂行するための、AI 間の連携や人と AI の協働に向けた取組が期待される。

一方で、高度化した AI が乱立することによって、予期せぬ振舞いが生じ、人間のコントロールが効かなくなる、粗悪な AI や有害な AI が混入して社会に悪影響を与えるといった懸念もある。

これに対しては、情報科学技術の知識のみでなく、社会学、心理学、経済学、哲学などの人文・社会科学等の知見を高度に融合させ、人間と AI が相互に補完しあいながらともに成長し、リスクを抑制しつつ価値創出の最大化を目指すことが重要である。あわせて、AI のリスク対応等には国際展開や国際標準も見据え、国際的な動向を注視しつつ検討を進めることが望ましい。

こうした背景を踏まえ、信頼性・公平性・安全性等を考慮しながら、多様な AI と人間が共存する“人と AI の共生・協働社会”を上手くデザインすることができれば、様々な社会課題の解決や社会全体のパフォーマンスの向上を図ることができるとともに、人々の Well-being や我が国の国際的プレゼンスの向上にもつながるものと期待される。

4. 達成目標

本戦略目標では、信頼性・公平性・安全性等を考慮しながら、人と AI の共生、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す。具体的には、以下を想定している。

（1）人と AI の共生

人と AI 間の共通理解（コモングラウンド）の仕組みや創発メカニズム､AI の論理的推論・説明技術、人の脅威となる行動を実行しないために AI が遵守すべき行動規範・常識等の実装方法、AI 依存による人間の主体性・思考力低下リスクの回避方策など、人間に対する理解や AI が人間に与える影響等も踏まえ、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・知見を獲得する。

（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や、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、群知能や能動的情報取得を実現するための技術等、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等を創出する。

（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等を見据えた複数の人と複数の AI による協働環境・社会デザインをはじめ、認知バイアスを持ちうる人間と AI の民主的討議・多角的意思決定や、相互学習・相互監視による信頼性・安全性・品質向上技術など、複数の人と AI が協働することでより良い社会に繋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等を図る。

5. 見据えるべき将来の社会像

「4. 達成目標」の実現を通じ、様々な特徴を有した AI と人間が相互に補完し、ともに成長することで、認知バイアスを排した民主的討議や多角的な意思決定、有害な AI の排除、入手情報の偏りによる社会の分断の回避等を実現し、公平性や多様性に配慮した包摂的な社会の実現を目指す。

また､AI という異なる知能とのインタラクションを通じて､Well-being や人間理解の深化を図るとともに、それを基に人に寄り添った AI 技術の進化を目指す。

さらに、人と AI の効率的なタスク分担や、人に寄り添った価値観に基づく AI の自律的な行動等により､生産性の向上や労働力不足の解消に寄与するとともに､異なる AI 間の相互運用プラットフォームを世界に先駆けて構築するなど、新産業の創出や国際的プレゼンスの向上に繋がることを期待する。

6. 参考

6-1. 国内外の研究動向

国内外ともに、生成 AI や AI エージェント技術に関する研究が盛んに行われているほか、産業界においても自律化・多様化する AI エージェントへの注目が高まっている。

（国内動向）

平成 26 年度戦略目標「人間と機械の創造的協働を実現する知的情報処理技術の開発」では予測符号化理論等の認知発達の理論的枠組など、平成 29 年度戦略目標「ネットワークにつながれた環境全体とのインタラクションの高度化」では文脈と解釈を同時推定する対話 AI 等の知見が得られたほか、令和 4 年度戦略目標「文理融合による社会変革に向けた人・社会解析基盤の創出」では社会特性を導入したマルチエージェントシミュレーションなど、令和 5 年度戦略目標「人間理解とインタラクションの共進化」では人間・社会の理解に基づく新たなインタラクティブシステム等の研究が進められている。

また、望む人が誰でも身体的能力、認知能力及び知覚能力を強化・拡張できる技術の創出を目指した内閣府 ムーンショット型研究開発制度のムーンショット目標 1 においても、脳科学、心理学等の領域の研究者が課題推進者として参画して研究が進められている。さらに、人と一緒に成長する AI ロボットの実現を目指す同制度のムーンショット目標 3 や、こころの安らぎや活力を増大する技術等の実現を目指す同制度のムーンショット目標 9 においては、人間への影響に焦点をあてたインタラクションの研究が進められている。

加えて、大規模な社会データから人間行動や社会現象を理解しようとする計算社会科学等における活動が活発化しており、人文・社会科学と情報学の融合したコミュニティが醸成されつつある。

（国外動向）

AI による人間の能力強化や人間拡張をキーワードとした研究が盛んに行われているほか、人と AI エージェントが問題解決・意思決定のためにチームを組む「Human-AI Teaming」や、マルチエージェント間コミュニケーション等の研究が進められている。

また、自律エージェント及びマルチエージェントに関する主要な国際会議である AAMAS[1]においては、投稿論文数が令和 4 年 615 件、令和 5 年 1,015 件、令和 6 年 1,113 件と増加の傾向を示しているほか、AAAI[2]で関係するワークショップが予定されており、「自動交渉エージェントのソフトウェアに関する競技会 [3]」の活動も活発化している。

その他、欧州では AI に関する包括的な規制である AI Act を発行し、人間の健康・安全や民主主義・法の支配に重大な害を及ぼす恐れのある AI については、ハイリスクな AI システム

1 International Conference on Autonomous Agents and Multi-Agent Systems
2 The Association for the Advancement of Artificial Intelligence
3 Automated Negotiating Agents Competition

として規制を導入しているほか、様々な枠組みにおいて、AI ガバナンスに関する議論が活発に行われている。

6-2. 検討の経緯

「戦略目標の策定の指針」（令和元年 7 月科学技術・学術審議会基礎研究振興部会決定）に基づき、以下のとおり検討を行った。

1. 我が国あるいは世界の基礎研究を始めとした研究動向について、科学計量学的手法を用いた論文分析や科学技術振興機構（JST）研究開発戦略センター（CRDS）の有する知見、科学技術・学術政策研究所（NISTEP）の各種調査結果、JST の有する過去の研究領域の評価結果や事業運営から得られた知見等を収集・蓄積し、研究動向を俯瞰した。

2. 上記情報収集の結果、科学技術・学術審議会 情報委員会 情報科学技術分野における戦略的重要研究開発領域に関する検討会の審議のまとめ及び CRDS ワークショップ「人・AI 共生社会のための基盤技術」等を参考にして分析を進めた結果、信頼性・公平性・安全性等を考慮しながら、多様な AI 及び人間が連携し、社会全体のパフォーマンスを高めることが重要であるとの認識を得て、「安全かつ快適な「人と AI の共生・協働社会」の実現に向けた新研究領域」を注目すべき研究動向として特定した。

3. 令和 6 年 12 月に、文部科学省と JST は共催で、注目すべき研究動向「安全かつ快適な「人と AI の共生・協働社会」の実現に向けた新研究領域」に関係する産学の有識者が一堂に会するワークショップを開催し、「安全かつ快適な“人と AI の共生・協働社会”」に必要となる技術やその重要性、戦略的に取り組むべき内容等について議論いただき、ワークショップにおける議論や関連する有識者からのヒアリング等を踏まえ、本戦略目標を作成した。

6-3. 閣議決定文書等における関係記載

『新しい資本主義のグランドデザイン及び実行計画 2024 年改訂版』（令和 6 年 6 月 21 日）

Ⅴ. 投資の推進　3. AI

生成 AI は社会経済システムに大きな変革をもたらす一方で、偽・誤情報の流布や犯罪の巧妙化など様々なリスクも指摘され、安全・安心の確保が求められる。

米国企業等が先行する中、我が国もそれに追随すべく計算資源の整備や大規模モデルの開発が進んでおり、また、小規模・高性能なモデルや複数モデルの組合せの開発等、新たな研究も進んでいる。

AI の開発や利活用等のイノベーションが社会課題の解決や我が国の競争力に直結する可能性がある。生成 AI を含む AI の様々なリスクを抑え、安全・安心な環境を確保しつつ、イノベーションを加速する。

『第 6 期科学技術・イノベーション基本計画』（令和 3 年 3 月 26 日閣議決定）
第 2 章 Society 5.0 の実現に向けた科学技術・イノベーション政策
1. 国民の安全と安心を確保する持続可能で強靭な社会への変革
（6）様々な社会課題を解決するための研究開発・社会実装の推進と総合知の活用
（c）具体的な取組
①総合知を活用した未来社会像とエビデンスに基づく国家戦略の策定・推進
〇ＡＩ、バイオテクノロジー、量子技術、マテリアルや、宇宙、海洋、環境エネルギー、健康・医療、食料・農林水産業等の府省横断的に推進すべき分野について、国家戦略に基づき着実に研究開発等を推進する。（略）

7. その他

本領域は AI・人間・社会それぞれの在り方を今一度見直すものであり、様々な分野が関係していることから、AI 分野だけでなく、インタラクションやロボティクス、認知科学、人文・社会科学、教育等幅広い分野の研究者の参画、及び関係領域の発展を期待する。各分野の相互発展を目指した共同研究等が行われることで、融合領域のコミュニティの醸成につながるとともに、例えば、AI という異なる知能とのインタラクションが図られることで、人間の特性が明らかとなり、それがさらに AI に求められる条件につながるといった相乗効果が生じることで、本戦略目標の効果が最大化するものと考えられる。

他方で、本領域は進展が速い分野であるとともに、世界的にも今後の検討領域であるといえることから、特に個人型研究においては、独創的かつ柔軟なアイディアが期待される。

いずれの場合においても、国際的な動きが重要となる分野であるため、国際展開や国際標準も見据え、国際的な観点を積極的に取り入れることが望ましい。

また、令和 5 年度戦略目標「人間理解とインタラクションの共進化」や令和 7 年度戦略目標「実環境に柔軟に対応できる知能システムに関する研究開発」等の戦略目標に基づく JST 戦略的創造研究推進事業、関係するムーンショット型研究開発制度や国立情報学研究所大規模言語モデル研究開発センター(LLM 研究開発センター)、情報通信科学・イノベーション基盤創出（CRONOS）等の取組と連携・情報共有することにより、新たな研究進展や成果創出の加速を促すことが望まれる。

最後に、本戦略目標は多様な AI が普及した将来の“人と AI の共生・協働社会”の在り方を問うものであることから、既存技術の単純利用やその軽微な改良等に留まらず、自身の研究成果が社会にどのような影響をもたらすのか、どういった社会像の実現に資するのかといったビジョンを持った研究提案を期待する。