# 研究領域「信頼されるAIの基盤技術」事後評価（課題評価）結果

１．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。

２．事後評価の概要

２－１．評価の目的、方法、評価項目及び基準

「戦略的創造研究推進事業(先端的低炭素化技術開発及び先端的カーボンニュートラル技術開発(ALCA-Next)を除く。)の実施に関する規則」における「第４章　事業の評価」の規定内容に沿って実施した。

２－２．評価対象個人研究者及び研究課題

2020年度採択研究課題

（１）五十嵐　歩美（東京大学大学院情報理工学系研究科　准教授）
信頼される資源配分メカニズムの構築

（２）栗田　修平（理化学研究所　革新知能統合研究センター　研究員）
与えられた指示文章に従い言語で判断を説明するAI

（３）小林　泰介（情報・システム研究機構　国立情報学研究所　助教）
頑健性と安全性の性能限界を明らかにする深層強化学習

（４）菅原　朔（情報・システム研究機構　国立情報学研究所　助教）
説明性の高い自然言語理解ベンチマークの構築

（５）竹内　孝（京都大学大学院情報学研究科　講師）
リライアブルな意思決定のための時空間因果推論モデルの研究

（６）西田　知史（情報通信研究機構　脳情報通信融合研究センター　主任研究員）
脳情報に基づいたAIの信頼性評価技術の開発

（７）西野　正彬（日本電信電話（株）NTTコミュニケーション科学基礎研究所　特別研究員）
誤りがないことを保証する検証器つき機械学習の研究

（８）原　聡（大阪大学産業科学研究所　准教授）
機械学習モデルとユーザのコミュニケーション：モデルの説明と修正
（９）日高　昇平（北陸先端科学技術大学院大先端科学技術研究科　准教授）
機械理解の創成に向けた随伴関手の統計的推定理論の構築
（１０）藤井　慶輔（名古屋大学大学院情報学研究科　准教授）
生物集団移動の専門家が利用可能な説明・意思決定のための基盤技術
（１１）吉井　和佳（京都大学大学院情報学研究科　准教授）
人とAIの同化に基づく能力拡張型音楽理解・創作基盤

２－３．事後評価の実施時期
2023 年 12 月 4 日（月曜日）事後評価会開催

２－４．評価者
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
領域アドバイザー
石川　冬樹　　情報・システム研究機構 国立情報学研究所　准教授
宇野　毅明　　情報・システム研究機構 国立情報学研究所　教授
浦本　直彦　　花王（株）DX戦略部門　上席執行役員・データ知創戦略センター長
大野　和則　　東北大学未来科学技術共同研究センター　特任教授
岡﨑　直観　　東京工業大学情報理工学院　教授
鹿島　久嗣　　京都大学情報学研究科　教授
佐久間　淳　　東京工業大学情報理工学院　教授
櫻井　祐子　　名古屋工業大学大学院工学研究科　教授
佐倉　統　　東京大学大学院情報学環　教授／理化学研究所 革新知能統合研究センター　チームリーダー
谷口　忠大　　立命館大学情報理工学部　教授
長井　志江　　東京大学ニューロインテリジェンス国際研究機構　特任教授
藤吉　弘亘　　中部大学理工学部　教授
松井　知子　　統計数理研究所 モデリング研究系　教授
持橋　大地　　統計数理研究所 数理・推論研究系　准教授
森永　聡　　日本電気（株）データサイエンス研究所　上席主席研究員
外部評価者
該当なし

３．総括総評

人間中心のAI社会原則に基づいた先端情報社会を実現するためには、現在のAI技術の弱点や限界を克服し、社会との整合性を確保する必要があります。そのため、本研究領域では、AI技術そのものを革新的に発展させ、将来のAI社会に必要な信頼性を備えた新しいAI技術の研究開発を目指しました。具体的には、以下の研究目標を掲げ、多様な観点と方法論から、将来の AI 分野の科学技術イノベーションにつながる新しいAI技術の研究開発を推進しました。

- 現在のAI技術の限界を克服する新技術の創出
- AIシステムの信頼性・安全性を確保する技術の創出
- データの信頼性確保および意思決定・合意形成支援技術の創出

1期生は、「未来の社会を支える信頼されるAI技術とは何か」という大きな問いのもとに、数理、情報技術、社会応用などの幅広い研究領域から集結した若手研究者たちです。それぞれの研究背景と独自の着想に基づき、脳神経科学、認知科学・心理学、教育分野から、数理、メカニズムデザイン、統計科学と機械学習、制御・計測の理論と技術、画像・動画・音声・音響・自然言語の認識・処理・生成、ロボット、人間情報・都市情報など、将来のAI技術を構成する幅広い分野に関して研究を実施しました。さらに、各分野の第一級の研究者である領域アドバイザーの助言を得ながら、組織の枠を超えた研究環境で、活発な議論を行いました。互いに刺激を受けつつ、それぞれのテーマに関して、挑戦的かつ独創的な研究を推進した結果、将来の信頼される AI の創出に資する可能性が高い、創造的な技術シーズとなり得る大きな研究成果を得ることができました。

具体的には、次のテーマに取り組み、多様な成果を挙げています。将来の信頼されるAIの意思決定に関連して、五十嵐 歩美氏は、社会における信頼できる資源配分メカニズムの理論構築と、NPOとの協力による家事分担アプリの開発・社会提供を行いました。AIシステムの安全性と頑健性に関して、小林 泰介氏は、頑健性と安全性を保証する深層学習とロボット制御技術の研究を推進しました。西野 正彬氏は、システムに組み込まれた AI の出力動作保証に関する機械学習の基礎理論構築を行いました。信頼されるAIシステムの評価とデータ構築に関して、栗田 修平氏は、視覚情報と自然言語の対応付けによる実世界の理解・応答・探索技術の研究と、各種のベンチマークデータセットの構築を行い、菅原 朔氏は、信頼性と説明性の高い自然言語理解システムの評価手法の開発と、そのためのベンチマークデータセットの構築を行いました。実世界へのAIの適応範囲拡大の方向として、竹内 孝氏は、信頼される意思決定のための時空間因果推論モデルの研究開発と、都市情報処理への応用可能性の検証を行いました。藤井 慶輔氏は、実世界の集団・群れの移動に関する説明・意思決定のための機械学習技術の研究開発と、自然科学・人間科学への適用可能性の追求を行いました。新原理に基づく AI 技術の研究開発に関して、西田 知史氏は、脳情報と AI の融合技術と、それに基づく AI の信頼性評価技術の研究開発を行い、日高 昇平氏は、人間の認知過程に関する数理構造に基づく機械理解の一般的枠組みに関して基礎研究を行いました。社会におけるAIと人間との協働に関して、原 聡氏は、システムとユーザ間の対話に基づく機械学習モデルの説明・修正可能性の研究開発を行いました。吉井 和佳氏は、音楽情報処理における人と AI の同化に基づく能力拡張型の音楽理解・創作技術の実現を目指して、暗黙知と形式知を統合した音楽AIシステム構築の新しい枠組みの提案と要素技術の研究開発を行いました。

以上のように、本研究領域の研究者は、「信頼されるAIの実現」を目指して、多様な観点と方法論に基づき、新しい理論の展開・構築と、新しい技術の開発や、プロトタイプ構築、実証実験を行いました。これらの研究により、将来の科学技術イノベーションに大きな影響を与える多くの研究成果を挙げると共に、信頼されるAIの基盤技術分野を開拓しました。

# 研究課題別事後評価結果

１．研究課題名：　信頼される資源配分メカニズムの構築

２．個人研究者名

五十嵐　歩美（東京大学大学院情報理工学系研究科　准教授）

３．事後評価結果

現代社会において、タスク・資源（「財」と呼ぶ）の公平な配分と、そのための AI と数理に基づく社会制度とサービスの設計と実装は重要な課題である。とくに、仕事、家事、医療物資に代表される分割可能性を仮定できない「不可分財」の公平な配分に関して、効率性と公平性を両立させる理論保証のある配分方式を明らかにすることは、AI による信頼される社会サービス実装のための喫緊の課題である。本研究は、新しい公平配分理論の構築と、社会実装による実証を目指した、独創的かつ挑戦的な研究であり、以下の 4 つのテーマについて成果を挙げている。

理論的な成果として、第 1 に、マトロイド制約と呼ばれる自然な数理構造の導入により、従来の理論では達成が難しいとされてきた不可分財の近似公平性・効率性・戦略的操作不可能性を同時に満たすことができる配分メカニズムを研究開発した。第 2 の成果として、財の組み合わせに関する時間的・物理的な制約構造と連結制約を考慮した近似的な公平性を満たす配分の存在を理論的に示した。第 3 に、フードバンクやライドシェアリングサービスなど、資源が逐次的に到着する状況における効率的かつ公平な配分のオンラインアルゴリズムを設計した。上記の公平配分に関する一連の理論的成果は、当該分野において高く評価され、トップレベルの国際会議や学術誌に採択されている。一方、社会実装の成果として、JST のサイエンスインパクトラボ参加を契機として NPO の協力を得て、公平な資源配分メカニズム理論を実社会に適用し、カップルやルームメイト間の家事分担を可視化し、公平な分担を提案する家事分担アプリを開発した。このアプリは、2022 年 5 月に NHK スペシャルで紹介され、そのウェブサイトに 1 万を超えるアクセスがあるなど（2022 年 7 月時点）、社会的にも大きな反響を呼んだ。この成果は、AI に基づく新たな社会基盤技術として、今後の社会に大きなイノベーションをもたらすものと期待される。

これらのさきがけ研究を通じて、社会における資源配分の公平性を実現するための理論的研究と社会実装を着実に進め、当該分野のトップランナーの地歩を築いている。今後は、計算的社会選択理論研究の第一人者として、また、将来の AI・情報科学・社会科学の融合分野をリードする研究者として、さらなる世界的発展を期待する。

# 研究課題別事後評価結果

１．研究課題名：　与えられた指示文章に従い言語で判断を説明する AI

２．個人研究者名

栗田　修平（理化学研究所　革新知能統合研究センター　研究員）

３．事後評価結果

本研究は、AI が社会において人間と協働する際の重要な鍵となる、実世界における言語理解に焦点を当て、仮想世界や実世界と、言語情報を結びつけるベンチマークタスクとデータセットの構築の方法論と、実際のデータセットの構築を目的とした独創的かつ社会的な波及効果の高い研究である。本研究では、以下の二つの観点から研究を推進した。

(1) 実世界の情報と自然言語との対応付け技術については、テキストから参照された物体を画像や動画から探索する AI の基礎技術として「参照表現理解タスク」に着目し、主観視点動画からテキストで参照された物体を探すタスクや、全周カメラで撮影されたシーンをテキストにて説明するタスクなど、実世界とテキストをつなぐ各種のタスクを開発した。(2)自然言語による人間の指示を理解し、実世界・仮想世界から情報を取得し、行動や質問応答を行う技術については、視覚と言語を融合した実世界探索において、画像キャプションモデルを利用して言語指示と動作・視覚の結合や、3 次元の写実的なシーンからの質問応答、詳細な物体の探索などについて、タスクやモデルの研究開発と提案を行い、データセット構築を行なった。

これらの成果は、多数の優れた論文として、視覚・言語処理から深層学習にいたる AI 分野のトップ国際会議等に採択され、国際的にも注目されている。さらに、タスクとモデルの提案、データセットの公開などを通じて、今後の先端的な言語・視覚 AI の研究に不可欠な研究情報基盤の構築に貢献している点は国際的にも意義がある。これらは、従来の言語理解手法では扱うことができない物理世界の情報をテキストと対応付けて取扱うことを可能にする新しい「信頼される AI の基盤技術」の開発を促進・支援するものであり、学術・産業・社会的な波及効果が高い。今後の自然言語処理、画像処理、ロボティクスなどの幅広い分野の AI 技術基盤となると期待される。

本さきがけ研究を通じて、他分野のさきがけ研究者との連携や、実社会応用の検討などの活動も積極的に行い、AI 分野を横断した「実世界での言語理解」という新しい融合分野を着実に切り拓いている。また研究成果のオープンソース・オープンデータとしての国際的な公開による社会貢献も、今後の AI 研究の推進方法として高く評価できる。実世界で活躍できる AI システムの開発と評価に関する第一人者となることを期待する。実世界における言語理解の立場から、本研究で培った成果を、現在飛躍的に発展している生成モデル等の新しい AI システムで生じている種々の課題にも展開することを期待したい。

# 研究課題別事後評価結果

１．研究課題名：　頑健性と安全性の性能限界を明らかにする深層強化学習

２．個人研究者名

小林　泰介（情報・システム研究機構　国立情報学研究所　助教）

３．事後評価結果

本研究は、自律ロボットなどの環境を探索する自律型 AI に焦点を当て、実世界における信頼性を改善するための深層強化学習を研究開発している。とくに、深層強化学習の信頼性に大きく関与する頑健性（外乱への補償能力）と安全性（制約条件を満たす能力）の 2 つに着目し、モデルベース強化学習に基づいて、以下の 3 つの観点から、これらを積極的に保証する新しい機構を開発した。

(1) 確率的予測モデルの表現力向上に関しては、予測の最悪ケースを考慮可能な最適化問題を導出し、そこに潜むバイアス・バリアンスのトレードオフを高効率に調整可能なメタ最適化手法を開発した。さらに、確率分布が持つ表現能力を効率的に高めつつも、分布の平均的挙動を解析的に算出可能な制約付き確率分布モデルを開発した。(2) 実時間内での安全性保障に関して、高次元空間での低効率な予測を避けるため、制御可能な限界まで低次元空間に圧縮するためのスパース表現学習を開発し、ロボットマニピュレータによる実証実験で有用性を実証した。さらに、制約条件を破綻させない準最適解を高速に獲得するため、行動系列候補から失敗を積極的に回避する更新則を持つモデル予測制御を開発・実装して、GPU を用いずに CPU のみの計算環境においても、円滑な人とロボットのインタラクションを実現可能なことを実証した。(3) 頑健性の制約付き最大化に関しては、不等式制約付きの最適化問題を適切に扱うため、不等式制約の正則化項への変換手法を開発し、これをロボットアームに実装し、学習時に経験しなかった未知物体の把持や人とのインタラクション時の適応的な振る舞いを実現することができた。予測モデルの頑健性を最大化するために、敵対的学習において予測精度の低下率に不等式制約を設ける新たな枠組みを開発した。

これらの成果は、設計時から頑健性と安定性の要求を新たな最適化問題として組み込み、そのための理論に裏付けられた最適化手法を設計することで、深層強化学習器の信頼性を向上させた。単なる新奇な理論的アイディアの提案を行うだけにとどまらず、粘り強く、かつ、確実に、実ロボットにおける実装と実証実験を進め、人とロボットの円滑な相互作用の実現における提案手法の有用性を示した。まとめると、実世界で人と協調して働くロボットに要求される頑健性と安全性を、タスクの設計段階から数理的にモデル化し、性能保証をもつ最適化手法を設計することで、実世界の信頼されるロボットの開発・運用に際して重要な課題に一つの解を示した研究である。本さきがけ研究は、上記の課題に関して、個別のタスクに対する要素技術の開発を対象としたものである。今後は、将来出現すると予想される人と協調して働くさまざまな汎用のロボット運用基盤へも対応できるように、本研究で得られた方法論の一般化と汎用化にも取り組むことを期待したい。

# 研究課題別事後評価結果

１．研究課題名：　説明性の高い自然言語理解ベンチマークの構築

２．個人研究者名

菅原　朔（情報・システム研究機構　国立情報学研究所　助教）

３．事後評価結果

人間のように文章を理解する AI システムの開発における評価指標と品質保証に着目し、それらを実現するための説明性の高いベンチマークの構築と、その持続的開発を可能にする構築の方法論（ロードマップまたはチェックリスト）の研究開発を目標とする独創的かつ波及効果の高い取組である。本研究では、以下の 3 つのテーマに関して研究を行なった。

(1) 持続的開発を可能にするロードマップの作成に関して、心理学における文章理解の議論に基づき、ベンチマーク評価手法が満たすべき要件を整理した。はじめに読解の定義と評価方法について理論的な基礎を整備し、次に心理測定学における妥当性議論という枠組みを参照することで、測定結果の解釈の妥当性を手続き的に向上させるための段階的手順を提案し、自然言語理解ベンチマークにおけるチェックリストの形で具現化して、当該分野のトップ国際会議の論文として提案した。(2) データセット収集の方法論に関しては、ニューヨーク大学との国際共同研究により、クラウドソーシングを用いた高品質なデータ収集の方法論を開発し、4 つの収集プロセスに関して各 1000 件を超える問いを収集・編纂し、作成したデータセットを国際的に公開した。この結果は、2 編の論文として、当該分野の最難関国際会議に採択され、発表された。(3) ベンチマークデータセットの提案に関しては、上記で実施した二つの研究項目の知見を生かして、状況に応じて結末が変わるような文脈における常識推論や、賛否のある複雑な議論における論理的推論、埋め込み的に複雑な状況にある語用論的推論などの高度な文章理解力を評価するためのタスクとデータセットを研究開発した。これらは、現在出現しつつある極めて高い性能をもつ大規模言語モデルに対しても評価可能な、高度な言語能力と言語現象を必要とする言語理解タスクを提案しており、当該分野の最難関国際会議で採択・発表されるなど、注目を集めた。

さきがけ研究における一連の研究は、多数の優れた研究論文として当該分野の複数の最難関国際会議に採択・発表され、文章を理解する AI システムの信頼性に関して国際的にも高く評価されている。データセットの公開も積極的に行なっており、将来の信頼される AI の実現のために不可欠である。その一方で、客観的評価が困難とされるシステムとデータの信頼性の評価と品質保証に関して、さきがけでの分野融合的議論にも積極的に参加しながら、言語理解の本質に迫る深い理解の下に着実に研究を遂行し、研究を進展させた。さきがけでの成果は、現在、急速に発展し、社会・経済に影響を与えつつある大規模言語モデルに関しても、今後の開発の基盤となる評価データ・手法を提供するものであり、今後のさらなる飛躍を期待する。

# 研究課題別事後評価結果

１．研究課題名：　リライアブルな意思決定のための時空間因果推論モデルの研究

２．個人研究者名

竹内　孝（京都大学大学院情報学研究科　講師）

３．事後評価結果

本研究は、都市空間における人やモノの移動など、膨大な時空間データから高信頼な予測を行い、高度 AI 社会における信頼される意思決定に貢献する、時空間因果推論モデルの実現を目指した挑戦的かつ独創的な研究である。

従来の因果推論は大規模な空間データに対応しておらず、データの空間的な偏りが推論の歪みを引き起こすという問題点をもっていた。これに対して、本研究では、機械学習技術を時空間情報と統合し、新たなデータ駆動型予測モデルを開発することで、時空間データにおける機械学習と数理統計技術の新たな可能性を拓き、実世界における時空間情報を反映した信頼される意思決定の実現に貢献している。

成果として、まず、空間偏在を補正したデータ分布推定モデルに関して、新たな方式を研究開発し、公共交通ビッグデータを用いた新駅の需要予測において実証した。反事実仮想を用いた空間的介入効果推定に関しては、新しい予測モデルとして時空間データを空間グラフとして表現し、グラフニューラルネットワークと因果推論の介入効果推定法を組み合わせることで、人流誘導シミュレーションや購買活動における空間介入効果の推定を可能にした。移動系列データに対する因果推論に関しては、軌跡カーネルを用いた新しい機械学習手法を開発し、今までにない人やモノの移動軌跡データを対象とした空間推論と意思決定のためのモデルを研究開発した。これらを実装し、スポーツ科学、動物生態学、自動運転車シミュレーションの空間移動データの解析が可能なことを実証した。これらの研究成果は、空間における移動体のモデリング・予測のための基盤技術として重要な技術となると期待される。以上の成果は、多くの優れた論文として、人工知能、機械学習、空間情報等の分野の最難関国際会議に採択・発表され、空間情報学分野における国際賞 1 件と国内賞 3 件を受賞し、15 件の国内報道に掲載されるなど、国内外で高く評価されている。

さきがけ領域内外との異分野連携を行いつつ、さきがけ研究によって、従来、未開拓であった時空間データにおける機械学習・数理統計技術について、基本方式の創出、技術開発、新しいタスクの設計、社会における実証実験までを完了しており、本さきがけ研究スタート時の想定を大きく上回るすばらしい成果を示した。本研究の成果は、実世界時空間における高信頼な予測を可能とし、社会におけるさまざまな意思決定の信頼性を向上させ、安全で安心な社会の実現のための新たな AI 技術の可能性を示した。今後のさらなる飛躍を期待する。

# 研究課題別事後評価結果

１．研究課題名：　脳情報に基づいた AI の信頼性評価技術の開発

２．個人研究者名

西田　知史（情報通信研究機構　脳情報通信融合研究センター　主任研究員）

３．事後評価結果

本研究は、AI の提示情報に対する人間の主観的信頼性のメカニズムを解明し、AI に対する主観的信頼性を脳活動から解読・評価するための技術基盤を開発する、極めて挑戦的かつ独創的な取組である。このために、人間の主観的信頼性を生み出す脳内メカニズムに着目し、最先端の脳科学の知識と脳活動計測・解読技術に基づき、以下の 2 つのアプローチで研究を進めた。

1 つ目のアプローチとして、人間による AI への信頼メカニズムの解明に関し、認知神経科学の方法論と知見を用いて、AI への信頼性を生み出す脳内基盤の探究を行った。AI 生成画像に対する人間がもつ魅力度判断における脳内機序を、最新の機能的脳活動計測機器である fMRI を用いて、約 60 名の被験者による大規模計測実験を行うことで、AI 生成情報に対する人間の負の価値判断バイアスと脳応答の関係を明らかにした。さらに同規模の実験で、AI 不安感の個人差を生み出す脳内基盤を分析した。2 つ目のアプローチとして、人間らしい判断を行う AI の実現に関して、脳情報のモデルを AI へ融合する技術である脳融合深層ニューラルネットワーク（脳融合 DNN）の研究・開発を行った。映像の意味認知タスクの実験では、脳融合 DNN が個人差を反映した高精度な判断能力をもつことを示唆する結果を得た。

これらの成果は、脳神経科学、計算生物学、人工知能分野のトップクラスの国際会議・学術雑誌の論文として発表され、1 件の国際学会賞と 2 件の国内学会賞を受賞するなど、高い評価を受けた。これらの成果は、最先端の脳神経科学と AI 技術の融合により、AI の提示情報に対する人間の主観的な信頼性の仕組みを解明し、人間の個性を反映した AI システムの実現の可能性を示した点で画期的な研究であり、その学術的意義は極めて大きい。また、さきがけ研究領域においては、脳・人間科学と情報科学の境界を越える異分野間の議論と連携を積極的に進め、将来の信頼される AI に関する議論をリードし、深めてくれた。本研究は、未来の AI 社会において人と共存・協働する信頼される AI 技術の実現に向けた重要な一歩となると思われる。脳情報と AI 技術の融合の基礎研究を確実に進めるとともに、将来的な実用化に向けた継続的な取組も必要となるであろう。信頼される AI の地平を切り開く、脳 AI 融合の第一人者として、今後のさらなる飛躍を期待する。

# 研究課題別事後評価結果

１．研究課題名：　誤りがないことを保証する検証器つき機械学習の研究

２．個人研究者名

西野　正彬（日本電信電話（株）NTT コミュニケーション科学基礎研究所　特別研究員）

３．事後評価結果

従来の機械学習アルゴリズムは、本質的に統計的な仕組みを内包するため、予測の出力仕様の性能保証が困難であった。本研究は、自然言語による文章生成システムや、医療や自動操縦システムなど、実際の多くの AI システム応用において、検証可能な誤りをデータベースとして準備可能である点に着目した。そして、検証器つき機械学習モデルという新しいアイディアに基づき、動作保証付き AI システムの実現方式を開発するという斬新なアイディアを提示し、統計学習理論に基づく解析により、この提案の有効性を示している。成果として、外付けの検証器付き機械学習システムにおいて、任意の仕様のもとで検証器を用いても、元のモデルのもつ汎化誤差の上限が悪化しないことや、元のモデルが一定の予測性能をもつ場合には、推論時にのみ検証器を付加しても性能が悪くならないことなどの理論的性能保証を与えている。本研究の提案の適用対象としては、最近、注目されている自然言語による文章生成システムにおける事実との整合性保証や、医療や自動操縦での危険行動回避による安全性保証などが該当する。このような一見すると複雑な AI タスクにおいて、基本的かつ理論的性能保証をもつ方式を提案し、その有用性を示したことは、高く評価できる。さらに、仕様がわかっている場合に積極的に検証器を利用することで、予測の汎化誤差の改善に繋げられるという興味深い結果も得られている。また自然言語処理システムにおける機械学習アルゴリズムの信頼性検証のための手法として、数理計画法ソルバーを用いて機械学習モデルの訓練データの微小変動に対する頑健性を検証する手法を研究開発している。

本さきがけ研究において中心となる検証器つき機械学習アルゴリズムの性能解析に関する成果は、機械学習分野における最難関会議に採択・発表されている。また関連する機械学習の信頼性向上に関する研究について、国内招待講演と学術誌の解説記事を発表するなど、その研究活動は高く評価されている。今後、AI 技術の社会への受容が進むとともに、動作保証付きの AI システムへの要求はより大きくなると予想される。着実な基礎研究を継続するとともに、さまざまな社会応用分野での実証を進めることで、本さきがけ研究で得られた成果が、将来の信頼される AI システム構築の基盤技術として大きく発展することを期待する。

# 研究課題別事後評価結果

１．研究課題名：　機械学習モデルとユーザのコミュニケーション：モデルの説明と修正

２．個人研究者名

原　聡（大阪大学産業科学研究所　准教授）

３．事後評価結果

AI の基盤技術である機械学習の信頼性向上を目指し、その主要な問題点の一つであるモデルのブラックボックス性の解決に向けて、「機械学習モデルとユーザ間のコミュニケーション」を可能にする革新的な研究の枠組みを提唱し、機械学習モデルの説明・修正技術の開発と、その実現可能性について研究した極めて独創的な研究である。これは、従来の機械学習技術の在り方を根本から変える革新的な提案であり、モデルの説明と修正という双方向コミュニケーションを、機械と人間の間で実現することで、ユーザにとって望ましい AI システムの性質（安定性、頑健性、説明可能性、透明性、公平性など）を備えた機械学習モデルの実現を目指している。これは、機械学習の信頼性の向上に対して、統一的な視点からの解決策を与える極めて重要な研究といえる。現在、アドホックに研究されている機械学習運用における各種の問題点の解決に新展開をもたらすものであり、将来の学術的・産業的なインパクトは極めて大きい。

具体的な成果として、次の 3 つのテーマについて研究した。(1)モデルの説明に関して、説明のための関連性指標を研究開発し、モデル損失関数のパラメータ勾配のコサイン類似度が優れた指標であることを示した。一方、モデルの修正に関しては、(2)再学習による修正のための学習安定化問題に取り組み、アルゴリズム安定性理論を援用した決定木モデル向けの新しい手法を開発した。(3)さらに、完全分散化の環境で効率よく計算可能なハイパーパラメータ最適化法を初めて開発した。これらの研究成果は、機械学習分野における世界最高峰国際会議に採択されており、複数の国内学会賞を受賞し、報道掲載されるなど、国内外で高く評価されている。上記の研究成果は、「機械学習モデルとユーザ間でのコミュニケーション」という新しい概念を提唱し、独創的なアイディアと理論展開により「説明性」と「修正」という双方向コミュニケーションを実現する技術である。今後の AI の信頼性の研究開発の上で、研究進行のプロセス自体も重要な成果である。上記の成果は、従来の技術課題や分野の機械学習問題に狭く囚われることなく、本質的な解決策の議論を行なった上で、深層学習、画像類似性、差分プライバシー、安定アルゴリズム、分散計算などの複数の分野・技術を融合したアプローチの研究開発により生み出されている。このことは、本領域が目指す分野横断的研究を具現し、若きさきがけ研究者の良き手本となっている。また、本さきがけ領域の研究者とともに、当該分野の国内研究会において「信頼される AI」をテーマに企画セッションを企画・開催し、産業界との連携も進めるなど、提案する枠組みと研究成果の発信も積極的に行なっている。

このように、信頼される AI の中核となる機械学習モデルの信頼性向上のための研究に関して、本さきがけ研究を通して目覚ましい飛躍を成し遂げており、高く評価できる。今後は、機械学習の信頼性に関する第一人者として、またこれからの機械学習・AI 技術をリードする世界的研究者として、さらなる活躍を期待する。

# 研究課題別事後評価結果

１．研究課題名：　機械理解の創成に向けた随伴関手の統計的推定理論の構築

２．個人研究者名

日高　昇平（北陸先端科学技術大学院大先端科学技術研究科　准教授）

３．事後評価結果

人の理解の認知過程を捉える新しい枠組みとして視覚・言語の理解を統一的に扱う新しい学問分野「機械理解」の分野の創成を目指した独創的かつ野心的な研究である。従来の深層学習と異なる、群に代表される代数的構造に基づく理論的枠組みを提案し、視覚・言語の理解の認知メカニズムを提案するとともに、実証実験を行い、提案の有効性を実証している。従来の深層学習は、ブラックボックスであり、予測の理解や説明が困難であった。このことが、深層学習を、信頼性や透明性を必要とする社会応用から遠ざけている。この観点に基づき、本研究では、視覚・言語の理解に潜む群の代数構造に着目し、連続空間上に埋め込まれた視覚や言語データを、その深層構造である群構造へ橋渡しして、有限かつ離散な組み合わせ論的構造で表現することで、理解の透明性を向上させた。具体的な研究成果として以下の２つが挙げられる。

(1)視覚的理解に関しては、複数の立体像を視覚に生起させる、ネッカーキューブなどの線画データを用いて、視覚理解を、群構造の最大化による曖昧性を解消する情報処理として捉えた数理モデルを提案した。このモデルは、多様な線画が人間によって立体と知覚される度合いを定量的に予測し、同時に、対称性を増減させることで、立体と知覚される度合いを制御した画像生成に成功した。本研究成果は、視覚の心理学分野で長年議論されてきた２つの仮説に包括的な説明を与えることを可能とする画期的な成果である。(2)言語理解に関しては、高次元ベクトル空間への単語埋め込みに基づく自然言語処理技術に着目し、なぜこうした言語モデルが単語共起関係に基づく統計量を用いた機械学習手法によって構成可能かを説明する理論を構築した。この理論は、現代の大規模言語モデルの成功の理由を説明できる可能性をもち、興味深い成果である。

本研究の成果は、多くの学術論文誌や、国際会議、国際・国内研究集会などで発表された。視覚理解に関する成果は、認知科学分野の論文誌で論文賞を受賞し、言語理解に関する成果は、言語処理分野の国内会議発表で委員特別賞を受賞するなど、独創性が高い評価を得ている。提案の研究テーマに関する 2 回のオーガナイズドセッションを人工知能分野の国内学会で企画・開催するとともに、新聞・テレビ等で関連研究の報道発表を行うなど、研究成果に関する情報発信も積極的に行った。

本研究で推進した表層構造と深層構造の代数的融合に基づく機械理解の研究は、現在の AI 研究で主流となっている、機械学習と数理統計に基づく理解の研究を補完する重要な試みであり、将来の信頼される AI の実現に寄与すると思われる。独自の発想に基づき、新しい枠組みの提案と、粘り強い研究推進により着想を実証した点は高く評価できる。今後、さきがけ研究の成果を継続的に発展させて、その有効性を実証することで、次世代の人工知能と認知科学を根本から革新し、「機械理解」という新しい融合的学問分野を確立することを期待する。

# 研究課題別事後評価結果

１．研究課題名：　生物集団移動の専門家が利用可能な説明・意思決定のための基盤技術

２．個人研究者名

藤井　慶輔（名古屋大学大学院情報学研究科　准教授）

３．事後評価結果

人間やその他生物の多様な集団移動運動を研究対象とし、自然科学、社会科学、交通、都市基盤分野など、実環境での集団移動を扱う専門家にとって、説明・操作・意思決定可能な人工知能基盤技術を開発するという独創的かつ発展性の高い研究である。計測技術の発展により、様々な生物や人工物の集団移動・行動データが収集可能となり、社会や科学分野への貢献が期待される一方で、情報技術に詳しくない実環境の集団運動の専門家（例えば生物学の研究者やスポーツチームの監督等）にとって、これらの集団行動データの分析や活用は難しい課題となっている。そこで、単なる予測ではなく、学習で得られた機械学習モデルから、集団行動の分析と意思決定を支援する信頼される AI 技術を研究開発した。本研究により、以下の 3 つの主要な成果をあげた。

(1) 観察データに基づく集団運動規則の抽出に関して、自然科学分野の多種多様な生物を研究対象とする研究者と協力し、機械学習技術を用いて、動物行動学の理論モデルとデータを統合し、集団の移動軌跡から個体間の相互作用規則を推定可能な画期的な情報抽出技術を研究開発した。この技術は、群体移動のシミュレーションデータや、コウモリ、カツオドリ、マウス、ハエ等の様々な野生動物の行動データに基づき評価・改良され、多くの研究論文が出版されるなど、スポーツ科学や生物学におけるさまざまな研究に活用されており、その有効性は高く評価できる。(2) 反事実予測に基づく介入効果推定などの集団運動評価技術に関して、マルチエージェントの複雑なシナリオにおける反事実的な経時的介入結果を推定する手法を発案した。従来の機械学習方法では考慮されていなかった共変量の反事実的予測とマルチエージェント関係の構造を組み入れることで、集団行動の予測だけでなく、推定結果に基づいた能動的な意思決定を実現する画期的な手法である。その他にも、反事実予測をデータから実現可能な特性をもつ、様々な集団スポーツに適用可能な評価手法を研究開発した。(3) 強化学習に基づく集団運動シミュレーションと意思決定評価に関しては、スポーツデータに基づく価値関数の推定手法と、生物のように協力的な移動を行う自律型強化学習モデルのシミュレーション技術を研究開発した。従来の課題であった強化学習における実世界と仮想環境との間のドメインギャップを解決する新手法も提案した。

本研究の成果は、機械学習分野とシミュレーション分野等の当該分野の最難関学術誌と国際会議を含む多くの論文として採択・発表されるなど、その独創性と有用性が国際的に高い評価を受けている。集団移動研究に AI 技術を導入するという独創的な発想に基づき、生物学、スポーツ科学、交通科学など、幅広い分野への応用を切り拓いている。提案技術のコードの実装と、シミュレーションと実データを用いた技術検証による実効性の保証だけでなく、将来の社会実装に向けて、実社会における多様な応用の検討と、データを収集・保有する団体や企業との連携の検討がなされていることも高く評価できる。今後、さきがけ研究で得られた研究成果を継続的に理論展開し、社会実装方式を明らかにすることで、実世界の人間や生物の集団行動の理解を深め、多様な社会問題の解決に貢献するだけでなく、集団行動への AI 応用の第一人者として、信頼される AI 基盤技術の新たな可能性を拓くことを期待する。

# 研究課題別事後評価結果

１．研究課題名：　人と AI の同化に基づく能力拡張型音楽理解・創作基盤

２．個人研究者名

吉井　和佳（京都大学大学院情報学研究科　准教授）

３．事後評価結果

人間の音楽理解の計算モデルとして、音楽データから自律的に知識を獲得可能な統計的推論法を確立し、音楽理解と音楽創作支援を統合した信頼される音楽 AI の実現を目指すという独創的かつ挑戦的な研究である。音楽を専門的に学んでいない人でも、音楽を楽しみ、鑑賞し、暗黙・明示の両面から音楽理解を深め、音楽知識を蓄積することで、音楽創作の基盤を築くことができるとの着想の下、本研究は、大量の音楽データから、人の音楽理解と同様に、解釈可能な音楽知識を獲得し、暗黙知と形式知の両面を統合した音楽 AI の実現を目指すものである。とくに、以下の 2 つの点で特徴的である。1 つ目は、音楽理解と創作支援の統合である。従来の音楽は、音響や信号としての音楽情報の分析や生成に特化していたが、形式知としての楽譜の推論は苦手であった。音楽データから楽譜を推論する過程と、楽譜から音楽を生成する過程を対称的かつ融合的にとらえて、音楽理解と創作支援を統合する。これにより、二面性をもつ音楽データをシームレスに行える。2 つ目は、暗黙知と形式知の統合である。暗黙知の獲得を得意とする深層学習と、形式知の獲得を得意とする確率的生成モデルを組み合わせることで、音楽を理解し、創作を支援するための機械学習技術を実現している。

音楽データのシームレスな理解と操作を可能にする音楽 AI の実現に向けた具体的な研究成果として、以下の 3 つが挙げられる。(1)信号から記号への変換技術に関しては、音楽音響信号を楽譜に変換する「真の自動採譜」の実現を目標とし、歌声解析、メロディ解析、リズム解析、ハーモニー解析、声部構造解析、リズム音源分離等のさまざまな音楽理解タスクに関して、各種の生成モデルに基づく統計的推論と深層学習に基づくモデルの End-to-End 学習の融合による形式知と暗黙知の獲得の統合した画期的な音楽理解手法を開発した。(2)音楽知識の記号表現技術に関しては、音楽の形式知の背後にある暗黙知の体系化に取り組み、音楽の本質的構造である周期性や階層性を捉えた統合的な音楽言語モデルを研究開発した。音楽モデルが捉えたモデルの各要素の確率的な偏りを利用した音楽創作支援の有効性を実証した。(3)協調的な記号操作技術に関しては、学習で得られた形式知としての音楽知識からユーザの音楽創作能力を拡張することを目指して、ユーザの技量に合わせてピアノ譜の難易度を無段階に調節する手法や、バンド譜からピアノ譜面への編曲など、ピアノ譜と任意編成の吹奏楽譜の間の自動編曲技術を開発した。

これらの研究成果は、当該分野のトップレベル国際会議と学術誌を含む 28 報の論文として発表されており、プレスリリースや、書籍の分担執筆等、成果公表も積極的に行った。とくに、開発された採譜技術は、サービス利用可能な高い水準で実装されており、国際的な競争力をもっている。音楽言語モデルは、今後、大規模音楽言語モデルに繋がる可能性が高く、今後、国際的な音楽基盤モデルとしての地位の確立が期待できる。これらを通じて、さきがけ研究の成果は、音楽情報処理にとどまらず、暗黙値と形式知を扱う全ての AI 分野における信頼される AI の基盤技術となり得るものである。継続的に発展させることで、音楽 AI の発展に大きく貢献し、人々の音楽の楽しみ方を大きく変える可能性を秘めた、革新的な音楽 AI 技術となることを期待する。