「人と情報のエコシステム」研究開発領域

研究開発プロジェクト事後評価報告書

令和4年8月

研究開発プロジェクト名：「人と情報テクノロジーの共生のための人工知能の哲学2.0の構築」
研　究　代　表　者　：鈴木貴之　（東京大学大学院総合文化研究科 准教授））
実　施　期　間　：2018年10月～2022年3月

Ａ．総合評価

成果は得られたが限定的であると評価される。

本プロジェクトは、ディープラーニングの発達により2010年代に興った第3次人工知能ブームにおいては人工知能の社会的影響評価や倫理的検討は積極的になされているものの、第2次人工知能ブームにおいて派生したような人工知能の可能性と限界を巡る哲学的議論があまりなされていないという実情を背景に、文献調査、技術者への聞き取り調査、ワークショップ開催などを通じて従来の哲学的課題を精査することで、かつての人工知能の哲学をアップデートし今日の人工知能の隆盛に即した哲学的課題を設定することを目的に研究がすすめられた。

その結果、人工知能の可能性と限界を考察するうえで問題となる諸概念の関係の整理では、人工知能という研究分野は多様でありそのあり方を体系的に分類するための枠組みは現時点では存在していないことが明らかになり、今後適切な分類法を考案することの必要性が浮き彫りになった。また第2次人工知能ブーム期までに議論されてきた問題の核心は、意味理解の問題と関連性の問題という2つの問題に集約されることが明らかになったが、近年の深層学習においてはこの2つの問題は部分的に突破可能である可能性も示唆されるようになったことが示された。いずれにしてもこれら古典的人工知能の問題は、哲学における知能における身体の重要性という問題に収斂されると考えられるため、さらなる哲学的探求が必要であるとされた。

人工知能の可能性と限界を検討するための新たな理論的枠組みの構築上では、個々の課題に対する性能の高さよりも汎用性の高さを優先した生物知能は現時点では実現の可能性の目処はたっておらず、人類にとっては利用可能な課題は限られているがその課題に対してはほぼ理想的な性能を示す課題特化型の人工知能研究の推進が有益であることが明らかになった。

人工知能の社会実装可能性を考えるための手がかりとなる概念枠組の構築では、人間の知能の代替物としての人工知能と人間の知能の補完物としての人工知能のあり方が提示された。人間の知能の部分的代替物としての人工知能は、人間と異なる方法によって問題を解決することによって人間よりも優れたパフォーマンスを発揮することから、そのメカニズムの違いを考慮した人間の知能と人工知能それぞれの強みを生かすような実装及び利用を考えて

いく必要があるとされた。また、特化型人工知能が人間以上の性能を発揮できるのは条件や目標を明確に定式化できる課題であることから、広義の徳と呼ばれる能力を人工知能によって代替することは困難であることが明らかになった。人間の知能の補完物としての人工知能は、企業によるビックデータ分析がその一例となるが、今後我々はナッジエージェントを利用するなど意思決定の状況においてより適切な意思決定を導く可能性を得るとともに、これを利用することで我々の自律的な意思決定能力が損なわれる可能性などが示唆された。これらより、技術的分析のみならず、そもそもよい生は何かという哲学的問題の分析が今後の人工知能の実装において不可欠であることが明らかになった。

情報テクノロジーの研究開発において人文諸科学にどのような貢献の可能性があるかの検証では、人間の知能について様々な知見を持つ人文学者が人間の知能に固有の制約を明らかにすることで、高いパフォーマンスを発揮する人工知能を開発するための手がかりを与えうることが明らかになった。

これらの活動により得られた既存の課題の整理及び人工知能の今日的意義の考察は、それだけでも重要な成果であり人工知能の哲学の再構築という点で一定の進展があったと評価できる。また、これらを一般に理解可能な形でできるだけ数字を用いずに伝えることができるならば社会的インパクトも少なくないと考えられる。特に、インタビュー記録、読書ガイド、キーワード解説、文献解説、文献リストなどが掲載されている本プロジェクトのウェブサイトは、今後の発展次第では人工知能の哲学的考察をする上で参照必須のプラットフォームになりうる可能性があると考えられるため、本プロジェクトの研究開発成果の活用・展開の可能性は大きいと評価できる。しかしながら、現在のウェブサイトは一般の人が閲覧してもやや理解が難しい構成になっており改善が必要であると考えられるため、今後も継続的に本領域と関係しながら成果を様々に展開していく取り組みがなされることが期待される。

また、本プロジェクトを実施することで日本の哲学界で人工知能をめぐる議論を活発化し関心を高めることができたことも大きな成果であったといえる。さらに、本プロジェクトのメンバーの試行錯誤の活動を通じて、非専門家でも適切な順序で学習を進めていけば人工知能研究の現状に関して一定程度実質的な理解に到達できることも確認され、その学習方法については現時点では暗黙知ではあるものの今後それが明示的になるならば人工知能研究における専門家と非専門家の共通言語の構築に向けて大きな意義があると考えられるため、今後の活動に期待したい。

一方で、申請時に予定していたポスドク研究員の雇用が上手くいかなかったことから主に哲学者の間での人工知能の理解という点にリソースが割かれてしまい、達成目標として掲げられていたコンセプトマップや類型化のチャートなどのアウトプット創出及びテクノロジーの研究開発者と人文科学研究者との交流や研究者・技術者と一般市民との問題関心の共有促

進という点は未達であり、プロジェクト活動期間内の成果創出という点においては不十分であったといえる。また内容面においても、既存の課題の整理については十分丁寧になされたいえるものの、本プロジェクトを実施することで見出された新たな課題は特に見当たらないように思われるため、哲学者ならではの固有の視点からの新たな問題提起に今後期待したい。

Ｂ．項目評価

Ⅰ．研究開発プロジェクトの研究開発内容とその成果について

1.　目標の妥当性

妥当であったと評価する。

ディープラーニングの発達により2010年代に興った第3次人工知能ブームにおいては人工知能の社会的影響評価や倫理的検討は積極的になされているものの、第2次人工知能ブームにおいて派生したような人工知能の可能性と限界を巡る哲学的議論があまりなされていない実情をふまえ、従来の哲学的課題を精査することでかつての人工知能の哲学をアップデートし今日の人工知能の隆盛に即した哲学的課題を設定する、すなわち人工知能の哲学2.0を構築するという本プロジェクトの目標は極めて妥当であり社会的必要性の高いものであったと評価する。特に、徳のあり方を特徴づける暗黙知、身体、感情という概念が人工知能の限界に共通している点に着目し、「人工知能は徳をもちうるか？」「人工知能は人間の徳の滋養にいかに役立ちうるか？」という問いを設定することで哲学と人工知能研究を繋ぐという試みは、課題の性質上または参加研究者の性質上非常に有効であったと考えられる。

一方で、もう少し事前調査がしっかりとなされた後に研究開発をスタートしたほうが有益な成果を生み出すことが可能であったとも考えられるため、本プロジェクトの終了後も研究活動を継続されることを期待したい。

２．研究開発プロジェクトの運営・活動状況

ある程度適切になされたと評価する。

人工知能に関連する文献を広くカバーし国内の技術者の聞き取り調査やワークショップを実施することで人工知能に関する理解を深めるという活動は適切であった。コロナ禍において活動が制限される中でもオンラインに切り替えることで状況に適切に対応することができたと評価する。また、哲学の学会で本プロジェクトの成果を発表することで、人工知能が哲学検討の対象に再度浮上するきっかけとなったことも高く評価できる。さらに、当初の予定になかった国際交流については追加予算を獲得し、スペインのグラナダ大学との国際ワークショップを実施した点も有意義であったと評価する。

一方で、申請時に予定していたポスドク研究員の雇用が上手くいかなかったことから、主に哲学者の間での人工知能への理解に時間をとられてしまい、プロジェクトのウェブサイトにてキーワード解説や文献解説などが記載されてはいるものの、達成目標として掲げられていたコンセプトマップや類型化のチャートなどのアウトプットがプロジェクト実施期間内には創出されず、一般社会への成果の普及という点での活動がほぼなされなかったことは運営上の課題として残ったといえる。また、人工知能についてレクチャーしてくれる適切な研究者の確保が困難であったという点に関しても、本領域のマネジメント側や担当アドバイザーに相談するなどより適切な方法があったと考えられるため、今後のプロジェクト運営については再考を期待したい。

３．研究開発プロジェクトの目標の達成状況および研究開発成果（アウトプット・アウトカム）

成果は得られたが限定的である。

本プロジェクトは、ディープラーニングの発達により2010年代に興った第3次人工知能ブームにおいては人工知能の社会的影響評価や倫理的検討は積極的になされているものの、第2次人工知能ブームにおいて派生したような人工知能の可能性と限界を巡る哲学的議論があまりなされていないという実情を背景に、文献調査、技術者への聞き取り調査、ワークショップ開催などを通じて従来の哲学的課題を精査することで、かつての人工知能の哲学をアップデートし今日の人工知能の隆盛に即した哲学的課題を設定することを目的に研究がすすめられた。具体的には、①人工知能の可能性と限界を考察するうえで問題となる諸概念の関係の整理、②人工知能の可能性と限界を検討するための新たな理論的枠組みの構築、③人工知能の社会実装可能性を考えるための手がかりとなる概念枠組の構築、④情報テクノロジーの研究開発において人文諸科学にどのような貢献の可能性があるかを明らかにする、⑤情報テクノロジーの研究開発者と人文科学研究者との交流や研究者・技術者と一般市民との問題関心の共有を促進する、の5つの達成目標を掲げて実践がなされた。

①人工知能の可能性と限界を考察するうえで問題となる諸概念の関係の整理では、人工知能という研究分野は多様でありそのあり方を体系的に分類するための枠組みは現時点では存在していないことが明らかになり、今後適切な分類法を考案することの必要性が浮き彫りになった。また、第2次人工知能ブーム期までに議論されてきた問題の核心は、意味理解の問題と関連性の問題という2つの問題に集約されることが明らかになった。意味理解の問題の本質は記号接地問題として議論されてきたものでありこの問題の解決なしには汎用人工知能は実現できないとされてきたが、深層学習においては人工知能は明示的な定義なしに概念を獲得することが可能であるかもしれず、意味理解には記号接地は必ずしも必要ないという新しい

見解を示唆している可能性が見出された。関連性の問題はこれまで主にフレーム問題として論じられ人工知能の限界を示す例として問題提起されてきたが、近年の深層学習においては要素と全体の複雑な関係性を表現する関数を高い精度で近似することが可能であり、コード化不可能な性質を深層ニューラルネットワークによって捉えたりノンパラメトリックモデルで理解しうる可能性が示唆された。このように、かつての人工知能の哲学では見られなかった展開が技術的進展によってみられるようになっているものの、真の意味理解という点では現時点でも意味理解の問題と関連性の問題は残されているといえる。これら古典的人工知能の問題は、哲学における知能における身体の重要性という問題に収斂されると考えられるため、さらなる哲学的探求が必要であるとされた。

②人工知能の可能性と限界を検討するための新たな理論的枠組みの構築では、個々の課題に対する性能の高さよりも汎用性の高さを優先した生物知能は現時点では実現の可能性の目処はたっておらず、人類にとっては利用可能な課題は限られているがその課題に対してはほぼ理想的な性能を示す課題特化型の人工知能研究の推進が有益であることが明らかになった。

③人工知能の社会実装可能性を考えるための手がかりとなる概念枠組の構築では、人間の知能の代替物としての人工知能と人間の知能の補完物としての人工知能のあり方が提示された。人間の知能の部分的代替物としての人工知能は、人間と異なる方法によって問題を解決することによって人間よりも優れたパフォーマンスを発揮することから、そのメカニズムの違いを考慮した人間の知能と人工知能それぞれの強みを生かすような実装及び利用を考えていく必要があるとされた。また、特化型人工知能が人間以上の性能を発揮できるのは条件や目標を明確に定式化できる課題であることから、広義の徳と呼ばれる能力を人工知能によって代替することは困難であることが明らかになった。人間の知能の補完物としての人工知能は企業によるビックデータ分析がその一例となるが、今後我々はナッジエージェントを利用することである意思決定の状況においてより適切な意思決定を導く可能性を得るとともに、これを利用することで我々の自律的な意思決定能力が損なわれる可能性などが示唆された。これらより、技術的分析のみならず、そもそもよい生は何かという哲学的問題の分析が今後の人工知能の実装において不可欠であることが明らかになった。

④情報テクノロジーの研究開発において人文諸科学にどのような貢献の可能性があるかの検証では、人間の知能について様々な知見を持つ人文学者が人間の知能に固有の制約を明らかにすることで、高いパフォーマンスを発揮する人工知能を開発するための手がかりを与えうることが明らかになった。また③で明らかになったように、我々はどのような社会を望んでいるかという価値主導のテクノロジー開発が人工知能研究においては必要であることが示され、ここでも人文学者の貢献の可能性が示された。

⑤情報テクノロジーの研究開発者と人文科学研究者との交流や研究者・技術者と一般市民との問題関心の共有促進では、第2次人工知能ブーム期では人工知能研究者と人文科学研究者の交流の場が存在していたことやテクノロジーとウェルビーイングについて考察するためには一般市民の関与が必要であることから本プロジェクトにおいてもそのような場の構築を狙ったものの、コロナ禍において活動が制限されたことからほぼ活動ができなかったとされた。一方で、上記で得た成果を今後書籍化する計画やウェブサイトでの発信はそれなりに活発になされたといえる。

これらの活動により得られた既存の課題の整理及び人工知能の今日的意義の考察はそれだけでも重要な成果であり人工知能の哲学の再構築という点で一定の進展があったと評価できる。また、これらを一般に理解可能な形でできるだけ数字を用いずに伝えることができるならば社会的インパクトも少なくないと考えられるし、本プロジェクトを実施することで日本の哲学界で人工知能をめぐる議論を活発化し関心を高めることができたことも大きな成果であったといえる。さらに本プロジェクトのメンバーの試行錯誤の活動を通じて、非専門家でも適切な順序で学習を進めていけば人工知能研究の現状に関して一定程度実質的な理解に到達できることも確認され、その学習方法については現時点では暗黙知ではあるものの今後それが明示的になるならば人工知能研究における専門家と非専門家の共通言語の構築に向けて大きな意義があると考えられるため、今後の活動に期待したい。

一方で、上述したように申請時に予定していたポスドク研究員の雇用が上手くいかなかったことから主に哲学者の間での人工知能の理解という点にリソースが割かれてしまい、達成目標として掲げられていたコンセプトマップや類型化のチャートなどのアウトプット創出及びテクノロジーの研究開発者と人文科学研究者との交流や研究者・技術者と一般市民との問題関心の共有促進という点は未達であり、プロジェクト活動期間内の成果創出という点においては不十分であったといえる。また内容面においても、既存の課題の整理については十分丁寧になされたといえるものの、本プロジェクトを実施することで見出された新たな課題は特に見当たらないように思われるため、哲学者ならではの固有の視点からの新たな問題提起に今後期待したい。

４．研究開発成果の活用･展開の可能性

期待できる可能性はあるが限定的であると評価する。

インタビュー記録、読書ガイド、キーワード解説、文献解説、文献リストなどが掲載されている本プロジェクトのウェブサイトは、今後の発展次第では人工知能の哲学的考察をする上で参照必須のプラットフォームになりうる可能性があると考えられるため、本プロジェクトの研究開発成果の活用・展開の可能性は大きいと評価できる。一方で、現在のウェブサイ

トは一般の人が閲覧してもやや理解が難しい構成になっており、改善が必要であると考えられる。第2次人工知能ブーム期に人工知能研究者と哲学者のインタラクションが活発化した背景にはそうした交流を推し進めた編集者の存在があったとされるが、そのことを考慮にいれても今後本ウェブサイトが人工知能の哲学的考察をする上で参照必須のプラットフォームとなるためには、プロのコミュニケーターの手を加えなどしてそのページを見に来た人が何らかのインサイトを得ることが可能になるような表現を検討する必要があるとも思われる。この部分について本領域としても何らか協力できることもあるとも思われるため、今後も継続的に本領域と関係しながら成果を様々に展開していく取り組みがなされることが期待される。

また上述したように、本プロジェクトのメンバーの試行錯誤の活動を通じて、非専門家でも適切な順序で学習を進めていけば人工知能研究の現状に関して一定程度実質的な理解に到達できることも確認され、その学習方法については現時点では暗黙知ではあるものの今後それが明示的になるならば人工知能研究における専門家と非専門家の共通言語の構築に向けて大きな意義があると考えられるため、この部分についてのさらなる探求と社会への発信に大きく期待する。

今回は人文科学者が人工知能研究を学ぶという点に大きなリソースが割かれそれ以外の成果の創出に至らないという点はやや残念であったものの、逆の言い方をするならば、人文科学者の基礎的な人工知能研究への理解という段階は本プロジェクト期間内に終了したため、本プロジェクト終了後も研究活動を継続することでさらなる成果が創出されるということでもあると考えられる。国際的にみても情報社会や人工知能の議論において哲学者の参画は期待されており、この分野の若手研究者の人材育成は必須であると考えられるため、ERATO池谷脳AI融合プロジェクトとの連携活動への参画含めて今後のさらなる活動に期待したい。

Ⅱ．研究開発プロジェクトの領域への貢献

研究開発プロジェクトの運営と活動、および得られた研究開発成果は領域の目標達成にある程度貢献できたと評価する。

「人工知能が人の雇用を奪う」などと人工知能の誇大広告がなされていた2010年代において、人工知能研究の今日的進展対して哲学的観点からの考察を丁寧に実施し基礎概念を再検討した本プロジェクトの活動は、社会がより冷静で客観的な議論を展開する基盤を整えるために必要な重要な活動であったと高く評価する。本領域内では他にも人工知能の雇用代替性についてエビデンスベースで検証した山本プロジェクトや、冪即を評価軸として人工知能の現在の達成度を検証した田中プロジェクトなど人工知能の現在を冷静に評価したプロジェクトがあるため、それらのプロジェクトの成果とあわせて何らかのアウトプットをすることで社会にインパクトを与える可能性もありうると考えられる。今後も継続的に本領域と関係しながら、成果を様々に展開していく取り組みがなされることが期待される。

また、本プロジェクトの研究代表者は本領域として実施しているERATO池谷脳AI融合プロジ

ェクトとの連携活動において、ELSI検討チームの中核的メンバーであり、その点においても重要な貢献を果たしていると考えられるため、本プロジェクトの領域への貢献は非常に大きかったと高く評価する。

以上

令和4年8月
社会技術研究開発センター

「人と情報のエコシステム」研究開発領域における
2021年度　研究開発プロジェクト事後評価結果について（概要）

社会技術研究開発事業「人と情報のエコシステム」研究開発領域の研究開発プロジェクトに対し、以下のとおり事後評価を実施した。

１．評価対象

下表のプロジェクトを評価の対象とした。【６件】

プロジェクト名称	研究代表者	所属・役職 （事後評価実施時点）
データポータビリティ時代におけるパーソナル情報のワイズ・ユース実現支援プラットフォームに関する研究	柴崎　亮介	東京大学　空間情報科学研究センター　教授
パーソナルデータエコシステムの社会受容性に関する研究	橋田　浩一	東京大学　大学院情報理工学研究科　教授
人と情報テクノロジーの共生のための人工知能の哲学2.0の構築	鈴木　貴之	東京大学　大学院総合文化研究科　准教授
想像力のアップデート：人工知能のデザインフィクション	大澤　博隆	筑波大学　システム情報系　助教
過信と不信のプロセス分析に基づく見守りAIと介護現場との共進化支援	北村　光司	産業技術総合研究所　人工知能研究センター　主任研究員
人と新しい技術の協働タスクモデル：労働市場へのインパクト評価	山本　勲	慶應義塾大学　商学部　教授

２．評価の進め方

以下の手順で評価を行った

・令和4年2月　　評価用資料の作成
　　「終了報告書」提出
・令和4年2月　　事前査読
・令和4年2月23・24日　　ヒアリング評価
・令和4年3月　　評価報告書（案）の検討
・令和4年8月　　評価報告書の確定
　　評価報告書の内容に関する事実誤認および非公開事項の有無等確認を研究代表者等に対して実施。再検討、修正等を適宜行った後、評価報告書を確定。

３．評価項目

以下の評価項目により、評価結果を「評価報告書」として取りまとめた。

A．総合評価

B．項目評価

（１）研究開発プロジェクトの研究開発内容とその成果について

①目標の妥当性

②研究開発プロジェクトの運営・活用状況

③研究開発プロジェクトの目標の達成状況および研究開発成果

④研究開発成果の活用・展開の可能性

（２）研究開発プロジェクトの領域への貢献

４．評価者（所属・役職は事後評価実施時点）

＜領域総括＞

國領　二郎	慶應義塾大学 総合政策学部　教授

＜領域総括補佐＞

城山　英明	東京大学　大学院法学政治学研究科　教授

＜領域アドバイザー＞

加藤　和彦	筑波大学　副学長・理事（総務人事・情報環境担当）
久米　功一	東洋大学　経済学部　教授
河野　康子	一般財団法人日本消費者協会　理事
砂田　薫	国際大学ｸﾞﾛｰﾊﾞﾙ・ｺﾐｭﾆｹｰｼｮﾝ・ｾﾝﾀｰ　主幹研究員
信原　幸弘	東京大学　名誉教授
松原　仁	東京大学　大学院情報理工学研究科　教授
丸山　剛司	元 中央大学　理工学部　特任教授
村上　文洋	株式会社三菱総合研究所　ICT・ﾒﾃﾞｨｱ戦略ｸﾞﾙｰﾌﾟ　主席研究員
村上　祐子	立教大学　大学院人工知能科学研究科・文学部　教授

＜評価専門アドバイザー＞

村田　潔	明治大学商学部　専任教授
奥和田　久美	北陸先端科学技術大学院大学　客員教授

以上