戦略的創造研究推進事業
(社会技術研究開発)
研究開発実施終了報告書

「人と情報のエコシステム」

研究開発領域

「　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　」

研究開発期間　平成 30 年 10 月～令和 4 年 3 月

鈴木貴之
(東京大学大学院総合文化研究科 准教授)

目次

１．プロジェクトの達成目標

１－１．プロジェクトの背景

過去 10 年ほどのあいだに、人工知能研究は大きな発展を遂げた。自律型ロボットの開発やビッグ・データの分析など、関連する情報テクノロジーも急速に発展している。同時に、人工知能によって人間の仕事が失われる可能性や、人間がみずからよりもすぐれた人工知能を作りだし、それを制御できなくなる可能性など、情報テクノロジーが引き起こすさまざまな社会的・倫理的問題についても、活発な議論が展開されている。

しかし、真の汎用人工知能や自律型ロボットはいまだ実現しておらず、実現の見通しも立っていない。このような現状をふまえれば、社会にとってより緊急性のある問いは、汎用人工知能の実現に向けた原理的な困難はどこにあるのか、どのような種類の人工知能ならば短期的に実現可能か、短期的に実現可能な人工知能にはどのような社会実装の可能性があるのか、といったことだと考えられる。人工知能をめぐる近年の議論状況においては、自動運転をめぐる法的問題をはじめとする緊急性のある具体的な問題と、デジタル失業や人工超知能の問題などの長期的で仮想的な問題は活発に議論されているが、その中間に位置づけられる上記の問題には、十分な検討が加えられていない。

第 2 次人工知能ブーム期には、哲学者をはじめとする人文科学者を巻き込んで、人工知能の可能性と限界をめぐる問題が活発に議論されていた。そしてそこでは、古典的な人工知能のさまざまな原理的問題が指摘されていた。しかし、人工知能研究そのものの停滞によって、そこで指摘された問題が十分に検討されることがないまま、議論は下火になっていった。人工知能研究がふたたび大きく前進し始めた現在、深層学習をはじめとする人工知能研究における新たな手法とその成果をふまえて、かつての人工知能の哲学（人工知能の哲学 1.0）をアップデートし、現状に即した新たな人工知能の哲学（人工知能の哲学 2.0）を構築することが不可欠である。

ここで重要な手がかりとなるように思われるのが、徳（virtue）にかんする哲学的知見である。古代ギリシア以来、哲学者は、人間の知的能力、とくに実践的な能力の核心を徳として特徴づけてきた。徳とは、ある状況において、その状況をしかるべき仕方で把握し、しかるべき感情を抱き、それにもとづいてしかるべき仕方で行為する能力である。このような実践的能力は、汎用人工知能が実現しようとしているものにほかならないように思われる。徳は、一群の明示的な規則によって規定することのできないものであり、感情を重要な要素とし、有徳な人を模倣することやしかるべき習慣を身につけることによって獲得されると考えられてきた。ここで徳のあり方を特徴づける暗黙知、身体、感情といった概念は、人工知能の限界を論じるなかで、しばしば哲学者が言及してきたものにほかならない。しかし、人工知能の可能性と限界をめぐる従来の哲学的議論においては、これらの概念の関係は十分に明らかにされてこなかった。徳概念は、ここで分析の手がかりを与えてくれると考えられる。

第 2 次人工知能ブーム以降認知諸科学に生じた研究の進展は、人工知能の社会実装を考えるうえでも重要である。たとえば、身体化された心 (embodied mind) や拡張された心 (extended mind) といった考え方にもとづく身体性認知科学研究は、人間の知的活動は脳内だけで行われるものではなく、そこでは身体や環境が重要な役割を果たすことを明らかにしてきた。このような知見は、人工知能が人間の知能の改善や拡張にどのように利用できるかを考えるうえで、重要な手がかりとなる。

これらの背景をふまえ、「人工知能は徳をもちうるか？」および「人工知能は人間の徳の涵養にいかに役立ちうるか？」という問いを手がかりとして、人工知能の可能性と限界や人工知能の社会実装を考えるための理論的枠組を構築することが、本研究開発プロジェクトの課題である。

１－２．プロジェクトの達成目標

達成目標①：徳概念を核として、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を整理する。哲学研究者以外にもその成果が容易に利用できるように、コンセプト・マップや用語集などを作成し、プロジェクトホームページで公開する。

達成目標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（人工知能の哲学 2.0）を構築する。その内容を、人工知能の研究開発者・人工知能の社会実装に携わる人々・一般市民にもアクセス可能な教科書や概説書などの形で公刊する。

達成目標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を構築する。具体的には、人間と人工知能のとりうる関係を類型化し、各類型の実現可能性や長所・短所などを明らかにしたチャートなどを作成し、プロジェクトホームページで公開する。

達成目標④：情報テクノロジー研究開発者へのインタビューや研究会の開催などを通じて、情報テクノロジーの研究開発において、哲学をはじめとする人文諸科学に（倫理的問題の検討以外に）どのような貢献の可能性があるかを明らかにする。

達成目標⑤：学会ワークショップやシンポジウムの開催などを通じて、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を促進する。さらに、一般向けのトークイベントや公開討論会の開催や新聞・雑誌における記事の執筆などを通じて、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。

２．研究開発の実施内容

２－１．実施項目およびその全体像

グループ①：人工知能の哲学 2.0 の構築

実施項目 1：人工知能研究の歴史と現状のレビュー

以下の各項目を実施する準備作業として、合同研究会などの開催を通じて人工知能研究の歴史と現状に関する理解を深める。合同研究会では、人工知能研究に関するテキストの講読を行うとともに、人工知能研究者による講演会を随時開催し、本研究開発プロジェクトの前提となる知識の共有を図る。

実施項目 2：人工知能の哲学の体系的再検討

関連文献のレビュー、第 2 次人工知能ブーム期にこの問題を論じていた哲学者などを対象としたインタビュー調査、主要な論者を講演者に迎えての研究会の開催などを通じて、人工知能の可能性と限界をめぐる過去の哲学的議論において、これまでどのような問題が論じられてきたのかを明らかにする。

実施項目 3：徳概念を手がかりとした各論点の関係の分析

人工知能の限界を論じる際に持ち出される論点には、明示的規則の限界、暗黙知、文脈理解の能力、直観、感情、身体の重要性、事実的判断と価値的判断の違いなどがある。本実施項目においては、これらの話題に関する近年の哲学的・経験科学的知見を参照しつつ、一連の論点の関係を明らかにする。

実施項目 4：人工知能の可能性と限界の解明

実施項目 1 から 3 の成果をふまえ、汎用人工知能や自律的ロボットを実現するうえでの本質的な問題がどこにあるか、人間の知能と人工知能にはどのような共通点と相違点があるかを明らかにする。具体的には、身体をもたない人工知能によって言語理解を実現しようとする現在の人工知能研究にはどれだけの可能性と限界があるか、新たなアプローチが必要であるとすればそれはどのようなものか、といった問いに具体的な答えを与えることを目指す。

実施項目 5：成果物の作成

実施項目 1 から 4 の成果を、第 2、第 3 グループの活動成果とあわせて、人工知能研究者・人工知能の社会実装に携わる人々・人間と情報テクノロジーの関係に関心のある一般市民など、さまざまな人にアクセス可能な形式の書籍にまとめ、出版する。

グループ②：徳と人工知能

実施項目 6：徳にかんする哲学的議論のレビュー

人工知能による徳の実現可能性を検討するための準備作業として、西洋哲学における徳研究の体系的なレビューを行う。徳倫理学と徳認識論という西洋哲学における二つの主要な徳研究領域のレビューを通じて、徳はどのような要素から成り立ち、どのように涵養されるものと考えられてきたのかを明らかにする。

実施項目 7：経験的知見をふまえた徳理解のアップデート

伝統的に徳倫理学が提唱してきた「徳」理解は、パーソナリティ心理学および社会心理学をはじめとする近年の経験科学の知見をふまえた批判的な検討をうけている。本実施項目では、経験的知見に基づいた批判とそれにたいする応答・修正の検討を通じて徳理解のアップデートを行い、人間にかんする科学的知見とも整合的な徳理解を提示する。

実施項目 8：人工知能による徳の実現可能性

実施項目 6 および 7 の成果をふまえ、人工知能による徳の実現の可能性と限界を考察する。具体的には、徳を構成する各要素のどのような側面は人工知能によって実現可能であるのか、またどのような側面は実現が困難であるかを明らかにする。また、情報テクノロジーによって人間の徳を涵養する方法としては具体的にどのようなものが考えられるかなどについて、具体的な分析と提案を行う。

実施項目 9：成果物の作成

実施項目 6 から 9 の成果を、第 1 グループ、第 3 グループの成果とあわせて書籍にまとめ、実施項目 5 の一部として出版する。

グループ③：拡張された心と人工知能

実施項目 10：拡張された心にかんする先行研究のレビュー

拡張された心や身体性についてのこれまでの研究・議論の要点を明らかにする。具体的には、①身体性認知科学（embodied cognitive science）が心の科学において市民権を得るまでに行われた論争および②4E（embodied, embedded, enactive, extended）の内実、現状、批判、論争についてのサーヴェイを行う。

実施項目 11：情報テクノロジーによる人間知性の拡張可能性の検討

実施項目 10 の成果をふまえ、また、関連諸分野（認知・学習・社会的協働行為・知的道具のデザインの分野）で提示されたさまざまな相互作用の類型を参考にして、人間と人工知

能の関係の類型化を行い、情報テクノロジーによる知能の拡張可能性という観点から、各類型の長所・短所を明らかにする。

実施項目 12：人間知性と人工知能の協働可能性の検討

実施項目 10、11 を踏まえて、さらに集団的知性や認知的分業にかんする心理学研究や認知科学研究を参照し、人工知能が人間とは異なる種類の自律的知的存在となりうる場合には、人間と人工知能にはどのような協働の可能性があるかを検討する。さらに、医療現場を事例として、協働可能性について具体的な分析を行う。

実施項目 13：成果物の作成（2021 年度）

実施項目 10 から 12 の成果を、第 1 グループ、第 2 グループの成果とともに書籍にまとめ、実施項目 5 の一部として出版する。

実施項目	2018 年度	2019 年度	2020 年度	2021 年度
1. 人工知能研究の歴史と現状のレビュー	→	→	→	→
2. 人工知能の哲学の体系的再検討		→	→	
3. 徳概念を手がかりとした各論点の関係の分析			→	→
4. 人工知能の可能性と限界の解明			→	→
5. 成果物の作成				→
6. 徳に関する哲学的議論のレビュー	→	→		
7. 経験的知見をふまえた徳理解のアップデート		→	→	
8. 人工知能による徳の実現可能性			→	→

9. 成果物の作成				→
10. 拡張された心に関する先行研究のレビュー	→	→		
11. 情報テクノロジーによる人間知性の拡張可能性の検討		→	→	
12. 人間知性と人工知能の協働可能性の検討			→	→
13. 成果物の作成				→

２－２．実施内容

グループ①：人工知能の哲学 2.0 の構築

実施項目 1：人工知能研究の歴史と現状のレビュー

(1) 目的：人工知能研究の歴史と現状に関する理解を深める。

(2) 内容・方法・活動：

人工知能研究の歴史と現状に関する理解を深め、プロジェクトメンバー間で知識の共有を図ることを目的として、人工知能に関する教科書をテキストとした読書会や人工知能研究者を講演者とする研究会などを開催した。具体的には以下のような活動を実施した。

・2018 年 10 月から 2019 年 1 月に 3 回の全体研究会を開催し、プロジェクトメンバーの発表を通じて各メンバーの知識および問題関心の共有を図った。

・2018 年 12 月の全体研究会で、鈴木貴之が Boden, *Artificial Intelligence: A Very Short Introduction* (Oxford University Press, 2018) の内容を報告した。

・2019 年 2 月に尾形哲也氏（早稲田大学）を講演者とした全体研究会を開催し、深層学習研究の現状に関する理解を深めた。

・2019 年 8 月に小野哲雄氏（北海道大学）および飯塚博幸氏（北海道大学）を講演者とした全体研究会を開催し、人間の意思決定の改善に人工知能を活用する手法に関する理解を深めた。

・2019 年 4 月から岡谷貴之『深層学習』をテキストとした読書会を開催し、深層学習に関する理解を深めた。

・2019 年 4 月に西垣通氏（東京大学）へのインタビューを実施した。
・2019 年 8 月に堀浩一氏（東京大学）へのインタビューを実施した。
・2020 年 10 月から Russell and Norvig, *Artificial Intelligence: A Modern Approach (Fourth Edition)*（Pearson, 2021）をテキストとした読書会をオンラインで開催し、人工知能研究の現状に関する理解を深めた。
・2021 年 3 月に、三宅陽一郎氏（スクウェアエニックス）および中島秀之氏（公立はこだて未来大学）へのインタビューを実施した。
・2021 年 4 月から、パールら『入門統計的因果推論』の読書会をオンラインで開催し、統計的因果推論に関する理解を深めた。
・2021 年 6 月に古宮嘉那子氏（東京農工大学）を講演者とする全体研究会を実施し、自然言語処理研究の現状に関する理解を深めた。
・2021 年 12 月に尾形哲也氏へのインタビューを実施した。
・2022 年 1 月に小野哲雄氏へのインタビューを実施した。
・2022 年 2 月に小町守氏（東京都立大学）へのインタビューを実施する（予定）。
(3) 結果：
・一連のテキストの読書会などを通じて、現在の人工知能研究は手法、目的いずれにおいてもきわめて多様であり、人工知能を深層学習と同一視する一般社会における理解は、人工知能研究の現状を必ずしも正確に反映していないことが明らかになった。
・現在の人工知能研究における手法の多くは統計的推測に基づいており、現在では人工知能研究と統計学が不可分な関係にあることが明らかになった。
・人工知能研究の課題や限界を検討するためには、各手法がどのような関係にあるのか（競合関係にあるのか、併用可能なのかなど）を明らかにすることが不可欠だが、現時点では、多様な手法や事例を整理するためのよい分類枠組みが存在しないということも明らかになった。
・インタビューを通じて、多くの人工知能研究者は、汎用人工知能の実現に向けては現在でも多くの課題が残されていると考えていることが明らかになった。
・活動成果のうち、インタビュー（本人確認済み分）、*Artificial Intelligence* 各章の要約の一部を資料としてプロジェクトウェブサイトに掲載した。
(4) 特記事項：
・プロジェクトメンバー向け研究会で講師となる人工知能研究者を見つけることができず、人工知能研究の現状理解に関してはメンバーによる読書会に頼らざるを得なかったため、この実施項目に予想以上に多くの時間を要することになってしまった。
・他方で、試行錯誤を通じて、非専門家が人工知能研究を学ぶための道筋が明らかになったため、プロジェクトウェブサイトで情報共有を進めていきたい。

実施項目 2：人工知能の哲学の体系的再検討

(1) 目的：人工知能の可能性と限界をめぐる過去の哲学的議論においてどのような問題が論じられてきたのかを明らかにする。

(2) 内容・方法・活動：

人工知能の可能性と限界をめぐる主要な哲学的論点とそれらの関係を明らかにするために、第 2 次人工知能ブーム期までの人工知能の哲学に関する文献の分析や、当時の主要な論者へのインタビューなどを行った。具体的には以下のような活動を実施した。

・2019 年 2 月から 3 月に、柴田正良氏（金沢大学)、黒崎政男氏（東京女子大学)、土屋俊氏（大学改革支援・学位授与機構)、松原仁氏（公立はこだて未来大学）へのインタビューを実施した。

・2019 年 3 月に、東京大学駒場キャンパスでシンポジウム「人工知能の哲学 2.0 の構築に向けて」を開催し、鈴木貴之が「人工知能の哲学のアップデートのために」という発表を行い、近年における人工知能研究の進展をふまえた哲学的考察の必要性を論じた。

・2019 年 8 月に開催された全体研究会において、鈴木貴之が「第 2 次人工知能ブーム期における人工知能の哲学：議論と教訓」という発表を行い、過去の人工知能研究における主要な論点は意味理解の問題と関連性理解の問題であり、両者は身体の重要性などいくつかの論点を共有することを明らかにした。

・東京大学の大学院生の協力を得て、人工知能の哲学に関連する文献リストを作成した。

(3) 結果：

・第 2 次人工知能ブーム期までの哲学的な議論は、人工知能の原理的な限界を指摘するものが中心であること、その主要な論点は意味理解と関連性理解の問題であること、いずれにおいても身体性が鍵となると考えられていることが明らかになった。

・他方で、意味理解と身体の関係や関連性理解と身体の関係に関しては、従来の哲学的議論においては十分な分析がなされていないことが明らかになった。

・以上の分析をふまえると、意味理解や関連性理解における身体の役割を明らかにすることと、深層学習をはじめとする人工知能研究における新たな手法がこれらの問題の克服をもたらすかを明らかにすることが、議論をさらに進展させるための鍵となることが明らかとなった。

・活動成果のうち、文献ガイド 1 点と文献リストの一部を資料としてプロジェクトウェブサイトに掲載した。

(4) 特記事項：

・実施項目 1 に時間を要したこともあり、文献解説、キーワード解説、コンセプトマップと行った成果物の作成を十分に行うことができなかった。これらについては、引き続きプロジェクトウェブサイトに随時情報を追加していく計画である。

実施項目 3：徳概念を手がかりとした各論点の関係の分析

(1) 目的：徳概念を手がかりとして、人工知能の限界を論じる際に言及される身体性や暗黙知といった諸概念の関連を明らかにする。

(2) 内容・方法・活動：

人間の知能において人工知能による再現が困難であるのはどのような側面であり、どのような理由で再現が困難であるのかを明らかにするために、倫理学における徳の分析や心の哲学における合理性の分析などに関する哲学文献の検討を行った。具体的には以下のような活動を実施した。

・2019 年度秋学期に、鈴木貴之が大学院で合理性のコード化不可能性をテーマとしたゼミを開講し、あわせて関連文献の文献調査を行った。

・2022 年 3 月に徳と人工知能をテーマとしたワークショップをオンラインで開催し、鈴木貴之が「徳のコード化不可能性と人工知能（仮)」という発表を行う（予定)。

(3) 結果：

・分析哲学では、合理性、徳、美といった現象は明示的な規則の集合によっては捉えることができないというコード化不可能性テーゼが主張されており、これは人工知能研究におけるフレーム問題と密接な関係をもつように思われることが明らかになった。

・徳に関しては、われわれは明示的な規則ではなく具体的な事例を通じて徳を身につけることが指摘され、合理性や美に関しては、要素が全体に与える影響が文脈によって根本的に変化するということがコード化不可能性をもたらすと指摘されている。このような分析がフレーム問題を考察する上でも重要なヒントになることが明らかになった。

・深層ニューラルネットワークのような複雑なパラメトリックモデルやノンパラメトリックモデルは、人間がなぜこのような能力をもちうるのかを理解する手がかりとなるが、これらだけでは十分な説明は与えられないことも明らかになった。

(4) 特記事項：

・実施項目 3 については十分な活動ができなかったため、プロジェクト終了後も学会ワークショップなどの形で活動を継続する計画である。

実施項目 4：人工知能の可能性と限界の解明

(1) 目的：汎用人工知能や自律型ロボットを実現するうえでの本質的な問題がどこにあるか、人間の知能と人工知能にはどのような共通点と相違点があるかを明らかにする。

(2) 内容・方法・活動：

実施項目 1 から 3 の成果および第 2 グループ、第 3 グループの研究成果をふまえ、現在の人工知能の可能性や課題を考察する。具体的には以下のような活動を実施した。

・2019 年 11 月に開催された日本科学哲学会第 52 回大会において、鈴木貴之がオーガナイザとなりワークショップ「機械学習・深層学習の哲学的意義」を開催し、鈴木貴之が「深層学習の哲学的意義：認知科学の哲学と人工知能の哲学の場合」という提題を行った。提題で

は、深層学習によって汎用人工知能を実現するためには転移学習やメタ学習などが不可欠であること、フレーム問題などにはまだ解答が得られていないことを論じた。

・2019 年 11 月にスペイン・グラナダ大学で開催された国際ワークショップ Japanese-European Meeting on Artificial Intelligence and Moral Enhancement において、鈴木貴之が “Toward an Update of Philosophy of Artificial Intelligence” という発表を行い、深層学習などの新しい手法の可能性と限界に関する理論的考察の必要性を論じた。

・2020 年 11 月に開催された国際学会 Philosophy of Human-Technology Relations Conference 2020 (PHTR2020) において、鈴木貴之がオーガナイザとなり、パネルセッション “Artificial Intelligence as a Tool” を開催した。パネルセッションでは、鈴木貴之が “Two Conceptions of Artificial Intelligence” という発表を行い、人工知能研究の短期的な目標としては、人間の代替物を作ることよりも、人間にとって有用な道具を作ることを目指すことが望ましいと論じた。

・2021 年 6 月に開催された科学基礎論学会の国際シンポジウム Fairness, Integrity and Transparency of Formal Systems: Challenges for a Society Increasingly Dominated by Technology において、鈴木貴之が指定討論者として “Transparency in AI: Identifying the Real Issue” という発表を行い、人工知能の不透明性やバイアスに関して、人間に由来する問題と人工知能自体に由来する問題を区別することが重要であると論じた。

・2022 年 1 月に開催された東京大学 AI センターのシンポジウム「AI 時代の哲学を考える」で、鈴木貴之が「人工知能の哲学 2.0 に向けて」という発表を行い、現在の人工知能は生物知能とは大きく異なるあり方をしており、汎用人工知能の実現への道筋は明らかでないと論じた。

・2022 年 3 月に開催予定の総括シンポジウムで、鈴木貴之が実施項目 1 から 4 までのおもな成果を報告する (予定)。

(3) 結果：

・現在の人工知能研究において大きな成果が得られているのは課題特化型の人工知能であり、生物がもつ汎用知能とは大きく異なることが明らかになった。

・生物進化のような過程を経ずに汎用人工知能を実現するための明確な方法は存在しないことが明らかになった。

・人工知能研究には、人間と同様の汎用知能を人工的に実現することと、人間にとって有用な知的道具を作ることという 2 つの異なる目標が考えられ、後者は前者を達成することなしに達成可能であるため、人工知能研究の短期的な目標としては後者が有益であるということが明らかになった。

・上記の 2 つの目標を明確に区別すれば、人工知能が人間と同様の方法で知的課題を解決する必要はなく、むしろそうでないことが望ましい場合もあることが明らかになった。

・鈴木貴之が、日本科学哲学会ワークショップにおける提題の内容を論文化し、『科学哲学』掲載論文「深層学習の哲学的意義」として公刊した。

(4) 特記事項：とくになし。

実施項目 5：成果物の作成

(1) 目的：実施項目 1 から 4 の成果を、第 2、第 3 グループの活動成果とあわせて出版物などとして公開する。

(2) 内容・方法・活動：

(3) 結果：

・主たる成果物として、鈴木貴之の単著『人工知能の哲学入門』と鈴木貴之の編集による論文集『人工知能とどうつきあうか―哲学から考える』の出版計画をまとめ、2022 年 1 月に勁草書房によって出版計画が承認された。これら 2 冊の書籍は 2023 年春に出版を予定している。

・本研究開発プロジェクトの活動成果の一部は、プロジェクトウェブサイトで公開しており、プロジェクト終了後も公開資料を随時追加する計画である。

(4) 特記事項：とくになし。

グループ②：徳と人工知能

実施項目 6：徳にかんする哲学的議論のレビュー

(1) 目的：人工知能による徳の実現可能性を検討するための準備作業として、西洋哲学における徳研究の体系的なレビューを行う。

(2) 内容・方法・活動：

徳という考え方がどのようにして生じ、展開していったのかを、一方でその起源となる古代ギリシア哲学にまで遡り検討し、他方で人工知能をはじめとする経験科学との接点を考察した。具体的には以下のような活動を実施した。

・2019 年 3 月に東京大学駒場キャンパスで開催されたシンポジウム「人工知能の哲学 2.0 の構築に向けて」において、立花幸司が「人間らしさとしての徳と人工知能」という発表を行い、徳という考え方の由来と展開、経験的検討などを報告し、人工知能と徳の接点を考える意義を論じた。

・2019 年 5 月に開催された全体研究会で、立花幸司が徳倫理学に関する発表を行い、西洋哲学で徳がどのように位置づけられてきたのかを報告した。

・2019 年 8 月に開催された全体研究会で、植原亮が「徳認識論の規範的側面について」という発表を行い、徳認識論で議論される知的徳について整理したうえで、人工知能が知的に有徳でありうるシナリオについて考察した。

(3) 結果：

・人工知能を用いた徳の涵養というテーマを考察する上で必要となる、徳という考え方の変遷と展開について整理した。

(4) 特記事項：とくになし。

実施項目 7：経験的知見をふまえた徳理解のアップデート

(1) 目的：経験的知見に基づいた批判とそれにたいする応答・修正の検討を通じて徳理解のアップデートを行い、人間にかんする科学的知見とも整合的な徳理解を提示する。

(2) 内容・方法・活動：

実施項目 6 で得た知見をもとに、情報科学のみならず、心理学や脳神経科学などの経験的知見に基づいて、現代の科学的知見と整合的な徳のあり方を検討した。具体的には以下のような活動を実施した。

・2019 年 11 月に開催された日本科学哲学会第 52 回大会におけるワークショップ「機械学習・深層学習の哲学的意義」で、植原亮が「機械学習・深層学習と知的創造性」という発表を行い、人工知能が広義の徳の一種である創造性に対してどのような可能性をもつかを検討した。

・2019 年 11 月にスペイン・グラナダ大学で開催された国際ワークショップ Japanese-European Meeting on Artificial Intelligence and Moral Enhancement において、立花幸司が “An extended reply to Lara and Deckers” という発表を行い、アリストテレスとは異なるソクラテス的な徳の育成という考え方を批判的に検討した。

・同ワークショップにおいて、植原亮が “Could AI be a creative machine?” という発表を行い、人工知能が知的創造性という認識的徳をもちうる可能性について検討した。

・2022 年 3 月に徳と人工知能をテーマとしたワークショップをオンラインで開催し、立花幸司が「生き方を支えるものとしての徳」という発表を、植原亮が「知的に有徳であることをめぐる議論の動向」という発表を行う（予定）。

(3) 結果：

・情報科学、心理学、脳神経科学などの経験的知見に基づいて、伝統的な徳目はあくまで現象的な区分であることや、生理的な基盤や学習プロセスにもとづく別の区分の可能性を認めれば、徳目をより柔軟に設定し学習のスタイルの幅を拡げることが可能になることを確認し、現代の科学的知見と整合的な徳理解を示すことができた。

(4) 特記事項：

・国際交流のための追加予算を利用して、グラナダ大学を中心とするヨーロッパの研究チームの研究交流を実施できたのは予想外の成果である。新型コロナウイルス感染症の影響により、計画していた活動の後半部分を実施できなかったが、今後何らかの形で研究交流を継続していきたい。

実施項目 8：人工知能による徳の実現可能性

(1) 目的：人工知能による徳の実現の可能性と限界を考察し、情報テクノロジーによって人間の徳を涵養する可能性を検討する。

(2) 内容・方法・活動：

人工知能を用いた徳の強化を、倫理的徳と認識的徳の双方から多角的に検討した。その際、人工知能による人間の能力の強化が「徳」として認められるための諸条件という哲学的な検討と、社会実装するうえで今後考慮し検討すべき事柄という応用倫理的な検討の二方面から取り組んだ。具体的には以下のような活動を実施した。

・2019 年 6 月に開催された Third International Workshop On Ethics And Human Enhancement において、立花幸司が “AI-based moral enhancement and the future of human virtue” という発表を行い、人工知能を用いた道徳能力の増強は人間の徳として認められるのかを検討した。

・2019 年 9 月に、立花幸司がジェノバ大学で “Artificial Intelligence and Epistemic Virtues. Virtue, Media, and Democracy” という発表を行い、人工知能による倫理的徳の強化が認識的徳のあり方に与える影響を検討した。

・2020 年 11 月に開催された国際学会 Philosophy of Human-Technology Relations Conference 2020 (PHTR2020) におけるパネルセッション “Artificial Intelligence as a Tool” において、立花幸司が “Artificial Intelligence as A Tool for Moral Education” という発表を行い、人工知能を用いて道徳を強化することは、道徳教育として認めることができるかを検討した。

・植原亮が、単著『思考力改善ドリル―批判的思考から科学的思考へ』において、認識的徳という考え方の重要性を検討した。

・2021 年 1 月に、思考力とウェルビーイングをテーマとしたワークショップを開催し、思考力とウェルビーイングはどのような関係にあるのか、人工知能は両者にどのように役立つかを参加者と議論した。

・2021 年 3 月に、立花幸司、植原亮と第 3 グループの中澤栄輔が一般向けのワークショップを開催し、人工知能による徳の実現可能性について議論した。

・2022 年 3 月に開催する総括シンポジウムで、立花幸司がグループの活動のおもな成果を報告する (予定)。

(3) 結果：

・人工知能による徳の実現可能性を、倫理的徳と認識的徳の双方から多角的に検討することで、人工知能による人間の能力の強化が「徳」として認められるための諸条件としては、発揮された判断や行動の理由を本人が理解し説明できること、獲得した能力に状況への柔軟性があること、伝統的な手法で獲得されたほかの徳との間で一定程度の整合性が保たれていることなどが挙げられることを明らかにした。

・また、社会実装に向けて今後考慮し検討すべき事柄としては、学習のプロセスとゴールが一定程度参加者に開示されること、強制ではなく本人の参加同意があること、獲得した能力を或る程度解除できる手法があること、義務や功利性といったほかの倫理観を尊重し価値観の多様性を認める内容に制限することなどなどが挙げられることを明らかにした。

(4) 特記事項：とくになし。

実施項目 9：成果物の作成

(1) 目的：実施項目 6 から 8 の成果を、第 1、第 3 グループの活動成果とあわせて出版する。

(2) 内容・方法・活動：

(3) 結果：

・鈴木貴之の編集による論文集『人工知能とどうつきあうか−哲学から考える』において、立花・植原が寄稿することとなった。

(4) 特記事項：とくになし。

グループ③：拡張された心と人工知能

実施項目 10：拡張された心にかんする先行研究のレビュー

(1) 目的：拡張された心や身体性についてのこれまでの研究・議論の要点を明らかにする。

(2) 内容・方法・活動：

身体性認知科学（embodied cognitive science）が心の科学において市民権を得るまでに行われた論争および、4E（embodied, embedded, enactive, extended）の内実、現状、批判、論争について、関連する文献と論文の調査を実施した。拡張された心と身体性についてのこれまでの研究・議論の要点を明らかにし、1990 年代以降、現在までの心の科学の変容のサーヴェイを行った。具体的には以下のような活動を実施した。

・2019 年 1 月に開催された全体研究会において、染谷昌義が身体性認知科学および 4E と呼ばれる諸理論に関する報告を行い、第二次人工知能ブーム以降において、心の哲学や認知科学において心のはたらきの特徴としてどのような論点が提示され、議論されてきたのかを示した。

・2019 年 3 月に東京大学駒場キャンパスで開催されたシンポジウム「人工知能の哲学 2.0 の構築に向けて」において、染谷昌義が「22 世紀の AI の哲学―人間本性論から資源本正論への方向転換」という発表を行い、昨今の人工知能研究の興隆において哲学（哲学的思考）が批判的に検討すべき課題は人間知性と人工知能との原理的差異ではなく、人工知能をはじめとする情報テクノロジーを利用することで人間の知性や経験がどう変容するかであるという問題を提起した。

・2019 年 3 月に高千穂大学で討論会「拡張概念をめぐってーメディア論・技術論・心の哲学」を開催した。柴田崇が「拡張概念を問う―メディア論の成果」という発表を行い、人工知能を含む人工物の議論全般で、機能や能力の「拡張」が人工物による「代替」とセットで使用されていること、その際、「代替」によって先験的に「拡張」の効果が語られる傾向があることを指摘した。上杉繁が「「拡張」のジレンマを超える―可能性を広げるデザイン」

という発表を行い、技術を使う経験によって自己が生成するという立場を示し，能力拡張によるジレンマ問題の分析方法について論じた。
・2021 年 3 月（当初予定は 2020 年 3 月。コロナ禍により 1 年遅延）に玉川大学応用脳科学研究センター「心の哲学研究部門」第 14 回研究会において、染谷昌義が「身体性と運動性―キスの制御則に示される心のはたらき―」という発表を行い、心のはたらきの身体性を、協調運動制御・知覚性運動制御・環境を友とする制御の理論と実践を紹介しながら指摘した。
(3) 結果：
・知性や認知といった心のはたらきが、身体性・環境依存性・環境内の認知的資源や道具により拡張変容する性格を有することがあらためて確認された。
・第三次人工知能ブームにおける新たな人工知能の哲学の課題として、人工知能を始めとする知能機械や情報テクノロジーを使用することで人間の経験や心のはたらきがどう変容し拡張するのかという問題が確認され、明確化された。
・技術の哲学や技術の人類学（民族誌）の研究を参照することで、心（人間の心のはたらき）の拡張性の具体像やその変容過程を参考にできることが明らかになった。
・拡張が人間の能力や生の可能性を「拡張」するといった楽観的意味だけではなく、人間の可能性を「退縮」する・閉ざすという悲観的でディストピア的な意味もあることが確認された。
・上記全体研究会における報告の一部は一般向けに書き換え、プロジェクトウェブサイトで公開した。
・身体性の重要な側面としての学習性について、身体の運動学習が入出力の連合学習でも入出力の変換学習（スキーマ形成）でもない点を、染谷昌義が「反復なき反復としてのわざ」（『わざの人類学』所収、2021 年）にて論じ公刊した。
・染谷昌義は、心のはたらきの身体性・環境依存性が示唆する心についての探究方法を「新たな自然学」と称し、身体・環境と相互作用を探究する意義を「二元論の向こう側を探る自然学のプログラム」（『現代思想』一般向け冊子）にてまとめた。
(4) 特記事項：とくになし。

実施項目 11：情報テクノロジーによる人間知性の拡張可能性の検討
(1) 目的：人間と人工知能の関係の類型化を行い、情報テクノロジーによる知能の拡張可能性という観点から、各類型の長所・短所を明らかにする。
(2) 内容・方法・活動：
　実施項目 10 の成果をふまえて、人間と人工知能を含む情報テクノロジーとの相互作用を論じている関連諸分野（認知・学習・社会的協働行為・知的道具のデザインの分野）を参照しながら、相互作用の類型と問題点（人間変容・論争・課題）の調査を実施した。具体的には以下のような活動を実施した。
・2019 年 4 月に開催された応用哲学会第 11 回年次研究大会で、上杉繁が「ロボット技術と

人間との関係におけるジレンマへのアプローチ」という発表を行い、人間―技術拡張論に基づき、ロボット利用におけるジレンマ問題の分析アプローチについて論じた。 The 21st Conference of the Society for Philosophy and Technology において、上杉繁が“Designing approaches addressing dilemma in relating to robots” という発表を行い、人間-技術拡張論に基づき、自動化と脆弱性に着目した、ロボット利用におけるジレンマ問題の分析について論じた。

・2019年8月に開催された The 22nd International Conference on Engineering Design. Delft University of Technology において、上杉繁が “Analysing and Solving the Reduced-ability and Excessive-use Dilemmas in Technology Use” という発表を行い、人間ー技術拡張論に基づいた技術利用のジレンマ問題へのアプローチ方法について論じた。

・2019 年 11 月にスペイン・グラナダ大学で開催された国際ワークショップ Japanese-European Meeting on Artificial Intelligence and Moral Enhancement において、染谷昌義が “What the 21st century's philosophy of AI should consider: Human-machine hybrid nature” という発表を行い、知性や思考能力を含めたヒトの心のはたらきがハイブリッド的であることから、人工知能についての哲学的考察がターゲットすべきなのはヒトと知能機械とのハイブリッド体である点を指摘した。

・同ワークショップで、上杉繁が “Design tool for analyzing human-AI technology relation. Japanese-European Meeting on Artificial Intelligence and Moral Enhancement” という発表を行い、人間-技術拡張論とポスト現象学の観点から、人工知能との関係を分析する概念ツールのアイディアを論じた。

・2020年8月に全体研究会を開催し、久保明教（一橋大学・人類学）による講演「ハイブリッドはいかに忘却されるか―現代将棋におけるソフトのツール化」と質疑を行った。プロ棋士が将棋ソフトと相互作用することにより生じる変容を、エスノグラフィーの成果から説明を受けた。

(3) 結果：

・技術の哲学や技術の民族誌の観点を取り入れながら、人工知能を含む情報テクノロジーと人間との相互作用を「ハイブリッド体」として理解する観点を確保し、ハイブリッド体の変容と変異を明らかにし、あわせて、課題や問題点も抽出した。

・人工知能を始め情報テクノロジーを使用することによる人間の能力の「拡張」（広がる）には、複数の意味があることが明らかになった。

・人間の能力を拡張するという問題の特徴は人間が変容する点にあり、人間と人工知能の関係を分析する上で、人間-技術拡張論とポスト現象学（とくに、アイディ、フェルベーク、クーケルバーグ）の視点が有効であることが明らかになった。

・人間の能力を拡張する問題として、技術の身体化によって人間の自己の境界が変容すること、技術による人間の代替によって他者との関係が変容することが見出された。

・技術利用における能力減退と過剰利用のジレンマ問題の分析と対応に関する上杉繁の講

演の論文は “Analysing and Solving the Reduced-Ability and Excessive-Use Dilemmas in Technology Use” (*Proceedings of the Design Society: International Conference on Engineering Design*, Cambridge University Press) に掲載された。
・染谷昌義は、心の働きが環境内のアフォーダンス資源を利用し「拡張」し変容すること自身が心のはたらきの本性であることを一般向の臨床系冊子に「「アフォーダンスからの希望」と題して公刊した。
(4) 特記事項：
・2020 年 8 月に実施した久保明教氏による講演は 2020 年 3 月に実施予定（2019 年度内実施予定）だったがコロナ禍により延期された。

実施項目 12：人間知性と人工知能の協働可能性の検討
(1) 目的：人工知能が人間とは異なる種類の自律的知的存在となりうる場合には、人間と人工知能にはどのような協働の可能性があるかを検討する。
(2) 内容・方法・活動：
実施項目 10、11 を踏まえ、人間と人工知能との協働を理解するための概念・観点と協働を設計するための問題点を抽出する考察を実施した。具体的には以下のような活動を実施した。
・2020 年 11 月に開催された国際学会 Philosophy of Human-Technology Relations Conference 2020 (PHTR2020) におけるパネルセッション “Artificial Intelligence as a Tool” において、柴田崇が AI (artificial intelligence)と IA (intelligence amplifier)という 2 つの構想を比較し、“AI vs. AI: The Real Issues Hidden in the Struggle” として発表した。
・同パネルセッションで、上杉繁が道具の 2 つのタイプという観点から人工知能の可能性を検討し、“Considerations on Analysing Relations between Humans and AI Technologies Based on Archetypes of Instruments - Club-type and Pot-type” として発表した。
・2021 年 7 月にワークショップ「AI 設計におけるリスクとその「可解性」について−「ゴリラ化問題」と「ミダス王問題」を皮切りに」を開催し、人工知能開発の立場から三枝亮（神奈川工科大学）が「ヒトと AI の共生から考える AI のリスクとリターン 」、技術哲学の立場から直江清隆（東北大学）が「多様安定性の諸様相 」という講演を行った。
・2022 年 2 月に井上悠輔（東京大学医科学研究所）を講演者として、医療における人工知能の活用をテーマとした研究会を開催する（予定)。
・2022 年 3 月に開催する総括シンポジウムで、染谷昌義がグループの活動のおもな成果を報告する（予定)。
(3) 結果：
・人工知能との協働可能性を有意義に考える観点、機械を AI（人工知能体）と見なすのか

IA（知能増幅体）と見なすのか、さらには増幅体が可能にしているのは真に増幅なのかという問題系があることが明らかになった。
・人工知能の設計は人間との協働を設計することに等しく、この設計を進めるには協働に不可避的に入り込む偶然的で多様な文脈性を、活用の個々の場面で考慮しなければならないことが明らかになった。
・技術の多様安定性の性質により、事前に設計した人間と技術の関係性がそのまま維持されることは困難である問題を明らかにした。
・人工知能との関係において、大規模、高速な情報処理の機能のみならず、偶然性などの価値も考慮した関係も設計しうることを見出した。
・柴田崇の単著（『サイボーグ―人工物を理解するための鍵』東京大学出版会、2022 年 4 月公刊予定）に、本研究課題の成果（人間と人工物の協働の歴史とその考察）を収録した。
(4) 特記事項：
・新型コロナウイルス感染症の影響により、当初予定していた医療現場における人工知能の活用についてのフィールド調査が実施できなかった。このテーマについては 2022 年度内に研究会を実施予定である。

実施項目 13：成果物の作成
(1) 目的：実施項目 10 から 12 の成果を、第 1、第 2 グループの活動成果とあわせて出版する。
(2) 内容・方法・活動：
(3) 結果：
・鈴木貴之の編集による論文集『人工知能とどうつきあうか―哲学から考える』に染谷昌義、柴田崇、上杉繁、中澤栄輔が寄稿する（予定）。
(4) 特記事項：とくになし。

実施項目：ウェルビーイングに関する哲学的検討（全体計画書未記載項目）
(1) 目的：人工知能をはじめとする情報テクノロジーとウェルビーイングの関係、情報テクノロジーを活用してウェルビーイングを高める可能性を考察する。
(2) 内容・方法・活動：
・鈴木貴之、植原亮、染谷昌義を中心として、ウェルビーイングの哲学に関するオンライン研究会を実施した。
・2020 年 12 月にオンライン開催された第 28 回産業ストレス学会シンポジウム「これからの働き方を考える」で、鈴木貴之が「テクノロジー、幸福、朗働」という題目で発表を行った。
・JST/RISTEX 松浦プロジェクトとの共催で、2021 年 1 月にオンラインワークショップ「思考力とウェルビーイング」を開催し、批判的思考力や科学的思考力とウェルビーイングの関

係、情報テクノロジーを用いてウェルビーイングを高める可能性などについて参加者で議論した。

・上記ワークショップの内容を拡大する形で、『国際哲学研究』(東洋大学国際哲学研究センター) 第 11 号で、鈴木貴之が編集者としてウェルビーイングの哲学を特集し、植原の論文、渡邊淳司氏（NTT コミュニケーション科学基礎研究所）と鈴木の対談を含む 5 本の論文、1 つの対談を掲載した。

(3) 結果：

・情報テクノロジーはわれわれのウェルビーイングにさまざまな形で影響を与えるが、両者の関係に関しては、まだ十分な検討が行われていないことが明らかになった。

・遠方の人とのつながりの感覚を高めるなどの仕方で、情報テクノロジーがわれわれのウェルビーイングを高める可能性があることが明らかになった。

(4) 特記事項：

・2020 年度より、HITE と ERATO 池谷プロジェクトの連携プロジェクトに鈴木貴之が参加しており、本プロジェクトの成果をふまえ、脳 AI 融合テクノロジーを対象としてテクノロジーとウェルビーイングの関係を引き続き考察する予定である。

３．研究開発成果

３－１．目標の達成状況

達成目標ごとの達成状況は以下の通りである。

達成目標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を整理する

文献調査などを通じて、現在の人工知能研究の全体像を把握することができた。また、第 2 次人工知能ブーム期の哲学的議論において指摘された基本的な問題は、現在も重要性を失っていないと同時に、深層ニューラルネットワークなどの発展をふまえた再検討が必要であることが明らかになった。

達成目標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を構築する

人工知能の可能性と限界を検討する際には、課題特化型である現在の人工知能と高い汎用性をもつ生物知能の対比が有用であること、現在の人工知能研究は、生物知能とは異なるアプローチで汎用知能の実現を試みていることが明らかになった。

達成目標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を構築する

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という区別が有用であること、前者においては全面的な代替と部分的な代替を区別する必要があること、そして、部分的な代替物としての人工知能や補完物としての人工知能を利用する際には、その強みと限界を正しく理解することが重要であることが明らかになった。

達成目標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

人文諸科学には、人間の知能に関する知見を人工知能の研究開発の手がかりとして提供することと、価値主導のテクノロジー開発を行うためのビジョン作りを行うことという 2 つの重要な役割があることが明らかになった。

達成目標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

新型コロナウイルス感染症の影響により十分な活動ができなかったが、成果物としての書籍の出版やプロジェクトウェブサイトにおける資料公開を通じて活動を続ける計画である。

以上の成果については、次のセクションでくわしく説明する。

３－２．研究開発成果

本研究開発プロジェクトは具体的な技術などの研究開発を目的とするものではないため、プロジェクトの活動で得たおもな知見を、達成目標ごとに以下でくわしく記述する。

達成目標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を整理する

要点：
・現在の人工知能研究は手法、目的ともに多様であり、それらを体系的に分類するための枠組みが必要。
・第 2 次人工知能ブーム期までの人工知能の哲学における主要な問題は、意味理解の問題と関連性の問題に集約できる。
・深層ニューラルネットワークに代表される近年の人工知能研究によって、どちらの問題に関しても新たな問いが生じる。

人工知能研究の多様性

文献調査や人工知能研究者による講演を通じて明らかになったことは、現在の人工知能研究が、手法に関しても目的に関してもきわめて多様であるということである。この点で、人工知能を深層学習と同一視するような現在の一般社会における認識は、人工知能研究の現状を正しく反映したものとは言い難い。

文献調査などから明らかになった重要な理論的問題は、多様な人工知能研究のあり方を体系的に分類するための枠組みが現時点では存在しないということである。一般に、ある領域において分類を行うときには、排他的かつ網羅的な分類基準を用いて分類を行うことが理想的である。このような観点から見たとき、人工知能に関する現在の標準的な教科書のあり方は、理想的とは言い難い。理想的には、アーキテクチャ（記号計算、ニューラルネット、ベイズネットなど）、学習（学習なし、教師あり学習、教師なし学習、強化学習）、目的（画像認識、自然言語処理、運動制御など）など、相互に独立ないくつかの分類基準を組み合わせることで、多様な人工知能研究手法に関する排他的かつ網羅的な分類が可能になると考えられる。適切な分類法は、人工知能研究の可能性と限界を考察するうえでも不可欠である。

本プロジェクトの活動では、適切な分類法の提案することはできなかった。今後は、人工知能研究者からの助言を得て適切な分類法を考え、プロジェクトウェブサイトで発信することを計画している。

意味理解の問題

人工知能の可能性と限界をめぐる従来の哲学的議論に関しては、文献調査の結果、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。

意味理解の問題に関する議論としては、サールの「中国語の部屋」の思考実験にもとづくものがもっとよく知られている。しかし、サールの思考実験においては、構文論的規則によって自然な会話が実現可能であるということが前提とされている。また、この思考実験からサールが導き出す結論は、真の意味理解には脳の生物学的な特徴が不可欠であるという、多くの哲学者が同意しない主張である。意味理解に関して本質的な問題はむしろ、サールの議論を受けて認知科学者ハーナドが論じている記号接地問題だと考えられる。人間は、身体をもち、言葉が表す対象、たとえばリンゴそのものをつかんだり、食べたりすることができる。そのような世界との交渉をもたない人工知能は、言葉の相互的な定義などを与えられるだけで、言葉の意味を真に理解することができるだろうか。これがハーナドの問題提起である。意味理解の問題は、たとえば「東ロボくん」がある種の会話問題への解答において困難に直面したというような形で、現在の人工知能研究においても依然として重要な課題となっている。

意味理解の問題の本質が記号接地問題にあるとすれば、近年の人工知能研究からは、いくつかの新たな論点を見出すことができる。第一に、深層学習、とくに畳み込みネットワークを用いた画像認識においては、深層ニューラルネットワークの深い層に、イヌやネコといった抽象的な特徴の表現が生じることが明らかになっている。このような方法によれば、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かもしれない。第二に、自然言語処理、とくに機械翻訳においては、ニューラルネットワーク上で自然言語間の翻訳の媒介となる表現を得ることが有用であることが明らかになっている。このような研究もまた、ニューラルネットワークにおいては明示的な定義なしに意味の表現が可能であるということを示唆している。他方で、機械翻訳においては、単語という単位はかならずしも重要な意味をもたないということも明らかになっている。これらのことは、意味理解には記号接地は必要ないということを示唆しているのかもしれないし、翻訳には（そして言語理解一般にも）言葉の意味は重要な役割を果たしていないということを示唆しているのかもしれない。これらの研究の理論的な含意を見極めるためには、深層ニューラルネットワークを用いた自然言語処理技術の進展に今後も注目する必要がある。

関連性の問題

関連性の問題は、おもに「フレーム問題」として論じられてきた。そこで問題となるのは、ある状況において重要である要因だけを考慮し、重要でない要因は端的に無視するという

ことは、いかにして可能かということである。第 2 次人工知能ブーム期には、古典的な人工知能が用いる明示的な規則によっては、この課題を適切に解決できないということが問題となった。

現在の人工知能研究においても、本質的には同様の問題がさまざまな場面で生じている。たとえば、自動運転において、目の前の信号が青ならばそのまま進むべきだが、右の道から救急車がやってきたときには停止すべきである。自動運転においては、このような例外的な状況が無数に発生する可能性があり、それらすべてに対する適切な対応を、明示的な規則によって表現することは不可能であるように思われる。

哲学においても、フレーム問題と本質的な関連を有すると思われる問題がこれまでに論じられてきた。たとえば倫理学においては、徳という考え方が重要な役割を果たしてきた。徳とは、ある状況において、その状況をしかるべき仕方で認識し（困っている人がいる）、しかるべき情動を形成し（同情する）、しかるべき行動をとる（助ける）能力である。一般に、徳は、明示的な規則の集合として捉えることはできず、さまざまな事例を通じて学習されるものだと考えられる。徳は、フレーム問題を解決するために人工知能が身につけなければならない能力にほかならないように思われる。

同様に、美学においては、美という性質は明示的な規則の集合によって捉えることができないということがしばしば主張される。心の哲学においては、合理性に関して同様のことが主張される。このようなあり方はコード化不可能性（uncodifiability）と呼ばれる。

コード化不可能性に関しては、その一つの要因が、ある種の全体論的性格あるいは文脈依存性にあるということが指摘されている。たとえば、ある絵画においては、ある色使いがその絵画の美しさに貢献しているとしても、主題が異なる別の絵画においては、同じ色使いが美しさを損なうことも考えられる。美に貢献する要素の働き方は、他の要素の働き方によって大きく変化するのである。別の言い方をすれば、絵画の美しさを、その要素の特徴の総和として単純に理解することはできないのである。

コード化不可能性の本質がこのようなことにあるとすれば、近年の人工知能研究からは、この問題に対処するための手がかりを得ることができる。上の分析によれば、美がコード化不可能性であるとは、美に関係する各要素は相互に独立に美に貢献するわけではないということである。そうだとすれば、第一の可能性として、要素と全体のより複雑な関係、すなわち、要素間の（場合によっては非常に高次の）相互作用を認めるようなあり方が考えられる。深層ニューラルネットワークは、そのような複雑な関係を表現する関数を高い精度で近似することが可能である。そうだとすれば、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（近似的に）捉えることが可能かもしれない。

第二に、コード化不可能な性質をノンパラメトリックモデルで理解するという可能性も考えられる。たとえば、多様な絵画は、主題、構図、色使いといった多様な次元から構成される高次元空間に配置することが可能だろう。ある絵画の美しさは、この高次元空間上でその絵画に近接した他の絵画の美しさに基づいて評価することが可能かもしれない。

ただし、これらの戦略がどの程度うまくいくかにはさらなる検討が必要である。第一に、絵画の美しさにはこれらのモデルが適用可能かもしれないが、意思決定の合理性は、論理的妥当性、経済的な最適性、道徳的な正しさなど、複数の性質を含むより包括的な判断である。特定の問題領域を超えてこれらのモデルが適用可能かどうかには、さらなる検討が必要である。第二に、美や善は、あらかじめ正解が存在する問題ではなく、新たな作品や事例に直面するたびにわれわれが基準を改訂するという性格をもつ現象かもしれない。コード化不可能性について考える際には、このようなダイナミックな構造も考慮に入れる必要がある。

意味理解の問題と関連性の問題は、いずれも身体と関連付けられて論じられてきたという点も注目に値する。記号接地問題においては、真の意味理解を達成するためには、人工知能は身体をもち、現実世界と相互作用をすることが不可欠なのではないかということが論じられてきた。関連性の問題に関して、ドレイファスなどの哲学者は、人間は身体をもち現実世界において行為することで、明示的な規則で捉えることのできない常識や暗黙的な知識を身につけることが可能となり、それがフレーム問題の解決を可能にすると論じている。このように、古典的人工知能の問題は、知能における身体の重要性という問題に収斂するように思われる。

とはいえ、身体がどのようにしてこれらの問題を解決するのかという点に関して、哲学者はこれまで具体的な説明を与えてこなかった。深層ニューラルネットワークを用いた人工知能研究や身体性認知科学の研究などを参照することで、身体と意味理解の関係や身体と関連性理解の関係を明らかにすること（そして身体が本当に不可欠なのかどうかを明らかにすること）は、今後さらに分析を進めるべき課題である。

達成目標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を構築する

要点：
・人工知能研究の目的として、汎用知能の実現と、課題特化型の人工知能の開発を区別することが重要。
・生物進化とは異なる過程を通じて汎用人工知能を実現する方法は明らかではない。
・応用という観点からは、汎用人工知能よりも課題特化型人工知能の方が短期的には有用である。

人工知能研究の目的

人工知能研究は、人間がもつ知能のように、特定の課題に限定されない汎用知能を人工的に実現することを究極目標としてきた。また、この目標を追求する際には、人間の知能と人工知能は基本原理を共有しうるということが前提とされてきた。人間の知能の原理が明らかになれば、それをコンピュータに実装することで人工知能が実現でき、ある知的課題をコ

ンピュータが行うことができるとすれば、そこでコンピュータが行っていることが、その課題に関する人間の知能のメカニズムにほかならないと考えられてきたのである。古典的な人工知能研究は、両者の基本原理は記号計算だと考え、それをコンピュータに実装しようとしてきた。しかし、今日の人工知能研究の多くは、汎用人工知能の開発を直接的な目的とするものではなく、特定の課題に特化された人工知能システムの開発を目的としている。このような経緯をふまえれば、人工知能研究のいくつかの目標や問題関心を区別することで、人工知能研究の可能性と限界に関して、人工知能は可能かというような抽象度の高い問いだけでなく、より具体的な問いを立てることが可能になる。

深層学習を例に考えてみよう。まず、課題特化型の人工知能の一手法としての深層学習には、莫大なパラメータをもつモデルである深層ニューラルネットワークはなぜ過学習を回避できるのか、(3 層ネットワークと深層ネットワークの表現力に違いはないはずであるにも関わらず) なぜ深層化することで性能が向上するのか、といった理論的な問いが考えられるだろう。

汎用人工知能の実現という観点からは、転移学習やメタ学習を組み込むことで深層ニューラルネットワークを基礎とする汎用人工知能は実現できるか、汎用人工知能を実現するためにはニューラルネットワークと記号計算的人工知能の統合が不可欠か、不可欠だとすればそれはいかにして可能か、といった問いが考えられるだろう。

人間の知能を理解するという観点からは、生物の脳と深層ニューラルネットワークには、層の深さや再帰的結合の量といった構造の違い、パラメータ調整の方法の違い、学習に必要な事例数の違い、誤りのタイプの違いなどが問題となる。これらの違いにもかかわらず、深層学習は知能の一般原理と言えるのだろうか。これは、知能の本質を考える上で興味深い問いである。

汎用人工知能の実現

近年における人工知能研究の急速な進展にもかかわらず、汎用人工知能実現の見通しはまだ立っていない。

汎用人工知能の実現可能性を考えるうえでは、現在の人工知能がうまくできる課題とうまくできない課題の本質的な違いを分析することが必要である。深層ニューラルネットワークによる機械翻訳は飛躍的に進展しているが、人工知能がチューリングテストに合格できるような自然な日常会話を行うことは依然として困難である。また、強化学習を用いた人工知能がレーシングゲームを人間よりもはるかにうまくプレーすることは可能だが、実世界における運転で直面するさまざまな例外的状況に対して、人間と同程度にうまく対処することは困難である。これらの事例の本質的な違いはどこにあるのだろうか。

直観的には、両者の違いは、条件や目標を明確に記述することが可能な課題であるかどうかという点にあるように思われる。この違いをより明確に定式化することは可能だろうか。

両者の違いは、たとえば、課題に対して適切な評価関数が設定できるかどうかという違いや、演繹的推論やベクトルの変換といった形式的な操作として問題を表現できるかどうかという違いとして考えることができるかもしれない。この点に関しては、さらなる分析が必要である。

汎用人工知能の実現可能性を考える上では、生物知能への理解を深めることも不可欠である。第 2 次人工知能ブーム期以降の 30 年ほどのあいだに、生物知能に関する理解も大きく進展した。そこで明らかになったことは、生物知能は進化の歴史と認知的な資源による制約を受けたものであり、ある状況で解決すべき問題に対して理想的な解答を与えるものというよりは、課された制約の下で最善の解答を与えるもの、すなわち限定的合理性を実現するものだということである。

理想的知能がすべての課題に対して高い能力を発揮できるものだとすれば、生物知能は、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知能だと言うことができる。進化の過程で汎用性と個々の課題における性能をともに高めていくことによって、生物知能は複雑化してきたのである。生物知能を手本とするならば、人工汎用知能を実現する一つの方法は、単純な自律的人工エージェントを作成し、それを徐々に複雑化することだということになる。

これに対して、現在存在する多くの人工知能システムは、特定の課題に特化して人間よりも高いパフォーマンスを示すものであり、この点で生物知能とは大きく異なるあり方をしている。そうだとすれば、一つの重要な問いは、特化型人工知能の適用範囲を徐々に拡張することや、複数の特化型人工知能を組み合わせることによって汎用人工知能を実現することは可能か、というものである。これは、生物知能の成立過程とはまったく異なるルートによって汎用知能を実現するということであり、それが可能であることは、けっして自明ではないのである。

他方で、生物知能との対比という観点から考えるならば、人工知能の社会的応用においては、汎用知能の実現という目標は重要なものではないかもしれない。生物知能と同様、人工知能も計算速度や記憶容量などの認知的資源の制約を受ける。したがって、人工知能によって実現しうるものもまた、ある種の限定的合理性だということになる。そうだとすれば、個々の課題における性能は理想にはるか及ばない（つまり人間と大きく異ならない）汎用人工知能よりも、利用可能な課題はかぎられているが、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の方が、われわれにとって有用だろう。

達成目標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を構築する

要点：
・人間の知能の代替物としての人工知能と補完物としての人工知能を区別することが重要。

・代替物としての人工知能においては、全面的代替と部分的代替を区別する必要がある。
・人間の知能の部分的代替物としての人工知能や補完物としての人工知能を利用する際には、その強みと限界を理解することが重要。

達成目標②で見たように、汎用型の人工知能と領域特化型の人工知能は本質的に異なるものである。そして、それぞれの背景には、2 つの異なる人工知能観を見出すことができる。それは、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という 2 つの見方である。

人間の知能の代替物としての人工知能

人間の代替物としての人工知能には、さらに 2 つの可能性が考えられる。人間の知能を全面的に代替するものとしての人工知能と、部分的に代替するものとしての人工知能である。人工知能研究が究極目標としてきたのは、人間の知能を全面的に代替するものとしての人工知能の実現である。これに対して、これまでの人工知能研究によって実際に開発されてきたのは、人間の知能の一部を代替するものとしての人工知能である。その例としては、たとえば自動運転車や画像診断システムなどを挙げることができる。

全面的代替と部分的代替を区別することからは、いくつかの示唆が得られる。第一に、これまでの人工知能研究の歴史が明らかにしてきたように、人間の知能の全面的代替物である汎用人工知能を実現することは容易ではない。それゆえ、短期的に重要となるのは、人間の知能の部分的代替物としての人工知能だということになる。第二に、部分的代替物としての人工知能は、人間と同じ方法で課題を解決する必要はない。人間と異なる方法で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が可能になるかもしれないからである。第三に、部分的代替物としての人工知能にとっては、全面的代替物、すなわち汎用人工知能に拡張可能であるかどうかは重要ではない。

他方で、人間の知能の部分的代替物として人工知能を利用する際には、両者のメカニズムの違いに注意が必要である。ある課題特化型人工知能が人間とは異なる方法によって課題を解決するときには、人間とは異なるタイプの誤りを犯す可能性が生じる。深層ニューラルネットワークによる画像認識における敵対的事例は、まさにこのような例だと考えられる。特化型人工知能が人間と異なるメカニズムにもとづくものであるときには、そのパフォーマンスがどれだけ優れているとしても、敵対的事例に類似した現象の可能性はつねに残されていると考えられる。

近年、説明可能な人工知能(XAI)の重要性が唱えられている背景にあるのは、このような事情である。しかし、特化型人工知能と人間の知能のメカニズムが異なるとすれば、このような試みにも一定の限界が生じるということに注意が必要である。「説明」が人間にとって理解可能なものとなるほど、人工知能において実際に生じている過程との乖離は大きくな

るかもしれないからである。

達成目標②で論じたことをふまえれば、人間の知能のどの部分を人工知能が代替するかも重要な問題になる。現在では、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。条件や目標を正確に定式化できなくなればなるほど、人工知能がフレーム問題や類似の問題に直面する可能性は高まる。この点で、将棋を指すことと現実世界で自動車を運転することのあいだには、大きな違いがある。自動運転は、人工知能による代替が困難な課題の一つかもしれないのである。一般的に、広義の徳と呼びうる能力を人工知能によって代替することは困難であり、そのような試みには注意が必要である。

人間の知能の補完物としての人工知能

人工知能の第二の可能性は、人間の知能を補完し、拡張し、増強するものとしての人工知能である。企業によるビッグデータの分析は、その一例である。たとえばスーパーマーケットの経営者は、どの時期にどの商品をどのような価格で販売するかを決定する際、これまでは、個人の経験かごく小規模な統計的データにもとづいて判断を下すほかなかった。しかし、今日では、全店舗の購買データを分析することで、店舗ごとに詳細な販売戦略を決定することができる。このような場面において、人工知能によるビッグデータの分析は、望遠鏡や顕微鏡と類比的な仕方で、われわれの認識能力を拡張してくれるのである。

人工知能による人間の知能の補完には、さまざまな可能性がある。たとえば、ある種のナッジエージェントを用いることで、われわれは、ある意思決定状況において検討すべき選択肢をすべて把握し、それぞれの選択肢から予想される帰結を具体的に思い描くことが可能になり、より適切な意思決定が可能になるかもしれない。また、人工知能を利用したバーチャルリアリティ上の経験を重ねることによって、われわれは倫理的徳や認識的徳を涵養することができるかもしれない。

他方で、このような目的への人工知能の利用にはさまざまな倫理的問題が生じうるという点に注意が必要である。たとえば、ナッジエージェントを利用することで、われわれの意思決定は特定の方向に誘導され、われわれの自律は損なわれるかもしれない。バーチャルリアリティによる訓練は、戦場において躊躇なく相手を攻撃する能力を身につけるというような、倫理的に問題のある目的に用いられるかもしれない。人工知能を用いた人間の知能の増強が、どのような場面でわれわれの生をよりよいものにし、どのような場面でより悪いものにするのかということを考えるためには、人工知能による能力増強テクノロジーの可能性に関する技術的な分析と、よい生とは何かという哲学的な問題の分析の両者が不可欠である。

補完物としての人工知能においては、インターフェースも重要となる。たとえば、深層ニューラルネットワークにおける情報処理は、そのままでは人間には理解が困難である。した

がって、深層ニューラルネットワークによる情報処理の強みを損なうことなく、その入出力を人間にとって扱いやすいものにすることが重要となる。具体的には、（機械翻訳におけるトンランスフォーマのような）自然言語による入力を深層ニューラルネットワークで扱いやすい形に変換するための技術などが重要になるだろう。

人間の知能と人工知能それぞれの強みを生かすような協働のあり方も、重要な課題となる。上でも述べたように、人工知能は条件や目標を量的な仕方で明確に表現できる課題において優れたパフォーマンスを発揮する。他方で、人間の知能は、そのような限定が困難な領域においても、柔軟な問題解決を行うことができる。両者の特徴を適切に組み合わせることで、人間あるいは人工知能単独の場合よりも高いパフォーマンスを発揮することが可能になると考えられる。たとえば、医療においては、画像診断のような課題は人工知能が行い、最終的な治療方針の選択は人間の医師が患者の意向をふまえて行うといった形が望ましいだろう。逆に、治療方針の選択を人工知能に任せてしまえば、フレーム問題やそれに類似した問題が生じることになる。人工知能の各応用領域において、望ましい協働や望ましくない協働のあり方を具体的に明らかにしていくこともまた、重要な課題である。

達成目標④：情報テクノロジーの研究開発において哲学をはじめとする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

要点：
・人文科学における人間の知能に関する知見は、人工知能の研究開発に重要な手がかりを提供する。
・人工知能の社会実装においては、価値主導のテクノロジー開発が必要。

かつて、哲学者は人工知能にはできないことがあると主張してきた。このような哲学者からの挑戦が、人工知能研究を進展させる一つの原動力になったことは間違いない。しかし、人文科学研究者が人工知能の研究開発に対してなしうる貢献は、このような挑戦を投げかけることだけではない。

第一に、哲学や心理学といった分野の研究者は、人間の知能に関してさまざまな知見を有している。上でも述べたように、人間の知能は進化の歴史と認知的資源の制約の産物である。人間の知能に固有の制約を明らかにすることは、より高いパフォーマンスを発揮できる人工知能を開発するための手がかりを与えてくれるだろう。また、人間の知能においては、身体や環境も重要な役割を果たしている。身体や環境に関する知見も、高い汎用性を持つ人工知能の開発には不可欠だろう。

第二に、達成目標③で確認したように、当面は、部分的代替物または補完物としての人工知能の利用が中心となると考えられる。したがって、その望ましい利用法を考えることが重要になる。この問題を考えるうえでは、人文科学者を交えたビジョン作りが重要になる。工

学としての人工知能研究では、設定された目標に対して、それを実現できる人工知能を具体的に開発することを目指す。しかし、情報テクノロジーと社会の関係をより大きな視点で見れば、われわれは人工知能の活用方法としてどのような目標を設定すべきなのかということが問題となる。これが問題となるのは、人工知能をはじめとする情報テクノロジーは、われわれにできることを増やし、われわれの能力を高めるものだが、それがわれわれにとってよいことであるとはかぎらないからである。たとえば、ナッジエージェントは、ある意思決定状況において利用可能な選択肢すべての詳細な検討を可能にしてくれるかもしれない。しかし、その結果われわれは情報過多に陥り、適切な意思決定を行うことができなくなってしまうかもしれない。このような事態を避けるためには、われわれは、望ましい社会のあり方を思い描き、そこに到達する助けとなるような仕方で情報テクノロジーを社会に実装する必要がある。ここで重要になるのは、われわれはどのような社会を望んでいるのかという問いである。これは、われわれは何を大切だと考えているのか、どのような生をよい生だと考えているのかという、価値に関する問いにほかならない。端的に言えば、これからの社会では価値主導のテクノロジー開発が必要なのである。この問題を考える上で、人文科学研究者が重要な貢献をなしうることは明らかである。

他方で、人文科学者がこれらの貢献をするうえでは課題もある。現在の人工知能研究に関して当を得た考察を展開するためには、当然のことながら、人工知能研究の現状に関してある程度の実質的な理解を有する必要がある。しかし、深層ニューラルネットワークをはじめとする現在の人工知能は複雑な数値計算を原理としており、数学に関する一定の知識なしに実質的な理解を得ることは困難である。本研究開発プロジェクトでは、メンバーによる試行錯誤を通じて、文系の非専門家でも人工知能研究の現状に関して一定程度の実質的な理解が可能であることや、どのような順序で学習を進めていけば一定の理解に到達できるかということを明らかにしてきた。今後は、人工知能の哲学に関する概説書の出版や、プロジェクトウェブサイト上での資料公開を通じて、これらの成果の共有を図りたい。

達成目標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

要点：
・人工知能研究者と人文科学研究者の交流を再び活性化させる必要がある。
・人工知能の問題は、テクノロジーとウェルビーイングの問題というより広い文脈でも考察すべき。

第 2 次人工知能ブーム期から研究を行っている人工知能研究者や、当時から人工知能について論じている哲学者へのインタビューによって明らかになったのは、交流の場の重要性である。当時は、人工知能と哲学や心理学といった関連領域の若い研究者が参加する研究

会などの場がいくつか存在したことによって、異分野間の交流が促進され、それぞれの分野の研究も促進された。その後、人工知能研究が専門領域として確立された結果、若手研究者はより具体的な研究に多くの時間を割くことになり、このような場は失われてしまった。このような場を再び作ることは容易ではないが、人工知能の社会的影響が広く論じられるようになった現在では、他分野の研究者との対話を望む若い人工知能研究者は再び増加していると考えられる。本研究開発プロジェクトでは、若手研究者との連携を具体的に進めることはほとんどできなかったが、オンラインでの研究活動が一般化した今日の状況を生かして、引き続き連携の可能性を探っていきたい。

本プロジェクトの活動からは、より広い文脈で人間と情報テクノロジーの関係を考える枠組みとして、テクノロジーとウェルビーイングの関係が重要な主題となることも明らかとなった。古代ギリシア時代から、哲学者は、幸福とは何か、よい生とは何かを論じてきた。その伝統は、現在ではウェルビーイングの哲学と呼ばれる領域に受け継がれている。しかし、ウェルビーイングに関する従来の哲学研究は、ウェルビーイングとは何かに関する概念的な分析をおもな主題としており、われわれの生において具体的にどのような要因がウェルビーイングを高めるのかという問題には、十分な検討が加えられてこなかった。他方、情報テクノロジーの社会的影響に関しては、人工知能の倫理やロボット倫理という形で個別の問題が検討されつつある。しかし、情報テクノロジーはわれわれの生をよりよいものにするかというより大きな問いについては、十分な検討は行われていない。情報テクノロジーをめぐる人文科学的な考察は、ELSI の枠組みを超えて、ウェルビーイングの問題と結びつけられる必要がある。さらに、当然のことながら、幸福やよい生は哲学者だけの問題ではない。情報テクノロジーとウェルビーイングの問題を考えるうえでは、一般市民の関与もまた不可欠である。情報テクノロジーとウェルビーイングの問題は、情報テクノロジー開発研究者、人文科学研究者、一般市民の三者が協働して取り組むべき問題として、今後重要な主題となると考えられる。

３－３．今後の成果の活用・展開に向けた状況

本研究開発プロジェクトの活動を通して得られた人工知能に関する非専門家としての理解は、広く共有する価値があるものである。成果物としての書籍の出版やプロジェクトウェブサイトにおける資料公開などを通じて、引き続き成果の共有を行っていくことを計画している。成果の共有を通じて、人文科学研究者による人工知能研究への理解を促進し、人工知能の哲学を含む、人工知能に関する人文科学的な研究を活性化することが、本プロジェクトの中期的な影響として期待されることである。

本研究開発プロジェクトの活動からは、テクノロジーとウェルビーイングというより広い問題設定の重要性も明らかになった。このテーマに関しては、すでに HITE 松浦プロジェクトとの連携によってワークショップや紀要における特集などを実施しているが、今後も

同様の活動の可能性が考えられる。また、2020 年度より鈴木貴之が HITE と ERATO 池谷プロジェクトとの連携プロジェクトに参加しており、脳 AI 融合テクノロジーを対象として、価値やウェルビーイングの観点からテクノロジーの望ましいありかたを検討している。このプロジェクトでも、テクノロジーとウェルビーイングという問題設定をさらに展開していくことが可能である。そのほかにも、鈴木は情報テクノロジーに関連するいくつかの大規模プロジェクトに参加を予定しており、これらのプロジェクトが採択された場合には、情報テクノロジーと労働や情報テクノロジーとメンタルヘルスといった問題に関して、テクノロジーとウェルビーイングの関係に関する考察を続けていくことが可能となる。

４．領域目標達成への貢献

全体計画書に記載した本研究開発プロジェクトの主たるアウトプットは、「技術と社会の対話の共通基盤となる概念の構築」である。このアウトプットを通じて、具体的には以下の形で領域目標の達成に貢献することが目標であった。

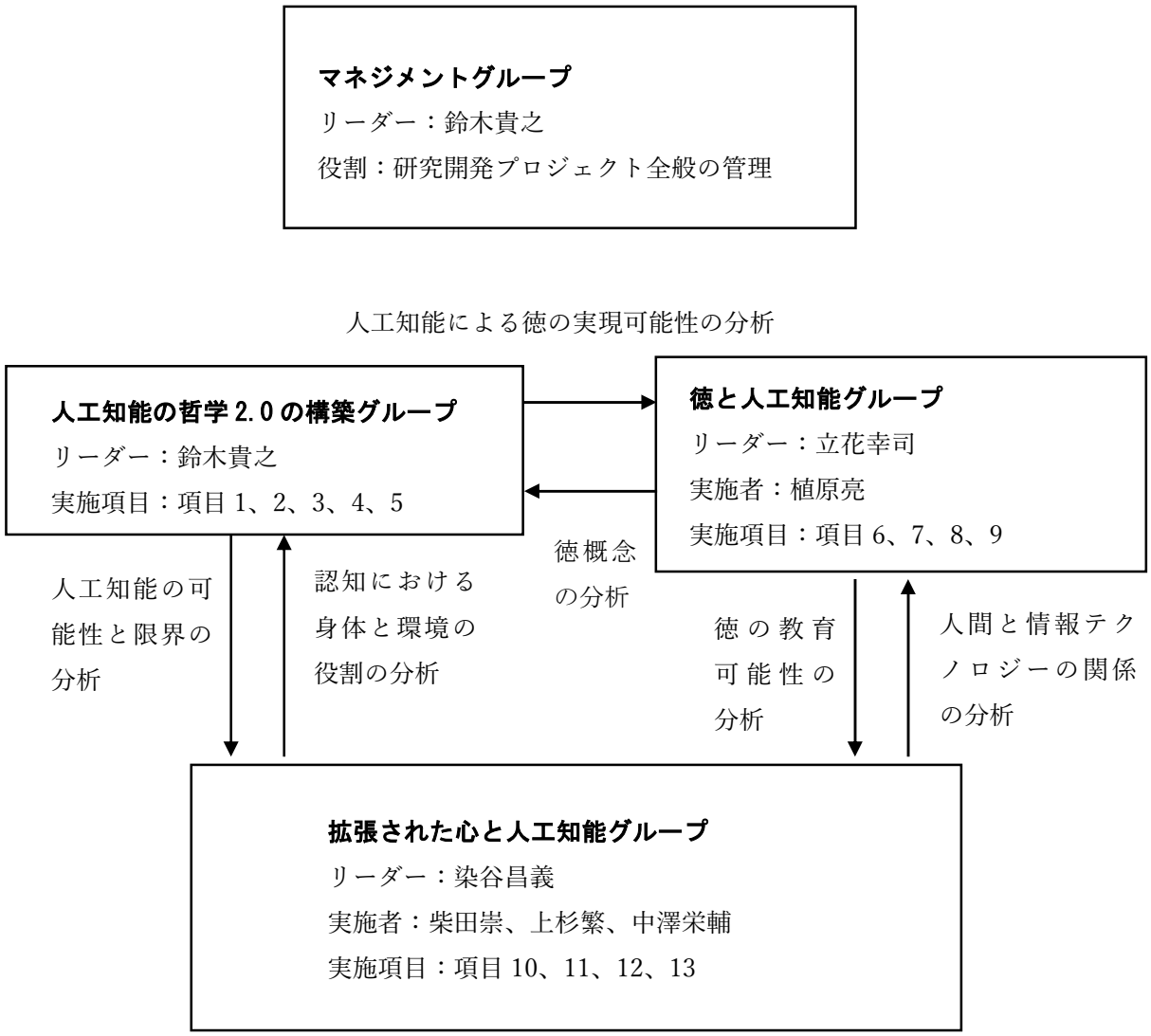
①人と情報テクノロジーの異同、それぞれの長所・短所、代替可能性と不可能性を明らかにする。
②人と情報テクノロジーの関係を考えるうえでの類型を提示し、各類型の長所・短所を明らかにする。
③哲学をはじめとする人文諸科学の研究者による情報テクノロジー研究開発への貢献の形を示す。
④若手研究者に、領域横断的・多分野融合型の研究プロジェクトの実施を経験する場を提供することで、「技術と社会の対話のプラットフォーム構築」に貢献する。
⑤研究開発の成果を一般の人々にもアクセス可能な形で提示することによって、人と情報テクノロジーの関係をめぐる諸問題についてのリテラシー向上をうながす。

本研究開発プロジェクトの活動からは、これらの各項目に関して、以下のような知見・成果が得られた。
①人間の知能は限定的合理性を実現する汎用型知能であるのに対して、現在の人工知能は課題特化型の知能である。この違いゆえに、人工知能は特定の課題においては人間よりも優れたパフォーマンスを発揮するが、より包括的な課題への対応には限界がある。人間の知能を人工知能によって代替する際には、両者のこのような違いに注意する必要がある。
②人工知能に関して人と情報テクノロジーの関係を分類すれば、人間の代替物としての人工知能と、人間の補完物としての人工知能に大分できる。前者はさらに、全面的代替物としての人工知能と部分的代替物としての人工知能に分類できる。人間の全面的代替物としての汎用人工知能を実現することは短期的には困難であるため、他の 2 つの目的を追求することが社会的に有用である。ただし、人工知能を人間の部分的代替物とする際には、両者のメカニズムの違いに注意が必要であり、人工知能を人間の知能の補完物として用いる際には、両者の長所が相互に補完し合うようにすることが重要である。
③人工知能の研究開発に対する人文科学研究者の貢献可能性としては、第一に、人間の知能に関する知見を提供することによって、その弱みを克服できる人工知能を開発するための手がかりとすることが可能であり、第二に、研究開発者とともに、望ましい社会のあり方、望ましい生のあり方に関するビジョン作りを行うことによって、どのような人工知能技術を開発し、社会実装していくべきかに関する指針を提供することが可能である。
④文献調査や読書会に東京大学大学院の大学院生を参加させることを通じて、人文科学研

究者が最新の情報テクノロジーについて学び、そこに人文科学的な問題を見出していくという過程を経験させることができた。
⑤本研究開発プロジェクトの主たる成果物である書籍（2023 年春出版予定）とプロジェクトウェブサイトにおける資料公開を通じて、人工知能研究の現状や社会実装に際しての注意点などについて、一般市民に向けた情報発信を行う計画である。

５．研究開発の実施体制

５－１．研究開発実施体制の構成図



５－２．研究開発実施者

（1）人工知能の哲学 2.0 の構築グループ（リーダー：鈴木貴之）

役割：第 2 次人工知能ブーム期までの議論とその後の人工知能研究の発展をふまえ、新たな人工知能の哲学の枠組みを構築する。

氏名	フリガナ	所属機関	所属部署	役職（身分）
鈴木貴之	スズキ　タカユキ	東京大学	大学院総合文化研究科	准教授

（2）徳と人工知能グループ（リーダー：立花幸司）

役割：人間の知能の本質を徳という観点から分析し、人工知能によるその実現可能性を検討する。

氏名	フリガナ	所属機関	所属部署	役職（身分）
立花幸司	タチバナ　コウジ	千葉大学	大学院人文科学研究院	助教
植原亮	ウエハラ　リョウ	関西大学	総合情報学部	教授

（3）拡張された心と人工知能グループ（リーダー：染谷昌義）

役割：人間の認知は脳だけでなく身体・環境を利用した活動であるという観点から、知的道具としての人工知能の可能性を検討する。

氏名	フリガナ	所属機関	所属部署	役職（身分）
染谷昌義	ソメヤ　マサヨシ	高千穂大学	人間科学部	教授
柴田崇	シバタ　タカシ	北海学園大学	人文学部	教授
上杉繁	ウエスギ　シゲル	早稲田大学	創造理工学部	教授
中澤栄輔	ナカザワ　エイスケ	東京大学	大学院医学研究科	講師

５－３．研究開発の協力者

氏　名	フリガナ	所　属	役職（身分）	協力内容
柴田正良	シバタ　マサヨシ	金沢大学	理事	人工知能の哲学に関する専門的知識の提供
黒崎政男	クロサキ　マサオ	東京女子大学	教授	人工知能の哲学に関する専門的知識の提供
土屋俊	ツチヤ　シュン	大学改革支援・学位授与機構	特任教授	人工知能の哲学に関する専門的知識の提供

西垣通	ニシガキ　トオル	東京経済大学	教授	人工知能研究に関する専門的知識の提供
松原仁	マツバラ　ヒトシ	東京大学	教授	人工知能研究に関する専門的知識の提供
尾形哲也	オガタ　テツヤ	早稲田大学	教授	人工知能研究に関する専門的知識の提供
堀浩一	ホリ　コウイチ	東京大学	教授	人工知能研究に関する専門的知識の提供
小野哲雄	オノ　テツオ	北海道大学	教授	人工知能研究に関する専門的知識の提供
飯塚博幸	イイヅカ　ヒロユキ	北海道大学	准教授	人工知能研究に関する専門的知識の提供
久保明教	クボ　アキノリ	一橋大学	准教授	人と情報テクノロジーの関係に関する専門的知識の提供
渡邊淳司	ワタナベ　ジュンジ	NTT コミュニケーション科学基礎研究所	上席特別研究員	人と情報テクノロジーの関係に関する専門的知識の提供
松浦和也	マツウラ　カズヤ	東洋大学	准教授	ウェルビーイングの哲学に関する専門的知識の提供
信原幸弘	ノブハラ　ユキヒロ	東京大学	名誉教授	ウェルビーイングの哲学に関する専門的知識の提供
三宅陽一郎	ミヤケ　ヨウイチロウ	スクエアエニックス		人工知能研究に関する専門的知識の提供
中島秀之	ナカシマ　ヒデユキ	札幌市立大学	学長	人工知能研究に関する専門的知識の提供
古宮嘉那子	コミヤ　カナコ	東京農工大学	准教授	人工知能研究に関する専門的知識の提供
直江清隆	ナオエ　キヨタカ	東北大学	教授	人と情報テクノロジーの関係に関する専門的知識の提供

三枝亮	サエグサ　リョウ	神奈川工科大学	准教授	人と情報テクノロジーの関係に関する専門的知識の提供
鈴木真	スズキ　マコト	名古屋大学	准教授	ウェルビーイングの哲学に関する専門的知識の提供
飯塚理恵	イイヅカ　リエ	関西大学	日本学術振興会特別研究員(PD)	ウェルビーイングの哲学に関する専門的知識の提供
古賀高雄	コガ　タカオ	東北大学	特任助教	技術哲学に関する専門的知識の提供
井上悠輔	イノウエ　ユウスケ	東京大学	准教授	医療における人工知能の活用に関する専門的知識の提供

６．研究開発成果の発表・発信状況、アウトリーチ活動など

６－１．社会に向けた情報発信状況、アウトリーチ活動など

６―１―１．プロジェクトで主催したイベント（シンポジウム・ワークショップなど）

年月日	名　称	場　所	概要・反響など	参加人数
2018/10/27	2018 年度第 1 回全体研究会	東京大学駒場キャンパス	鈴木貴之による発表	7 名
2018/11/24	2018 年度第 2 回全体研究会	東京大学駒場キャンパス	鈴木貴之による発表	7 名
2019/1/26	2018 年度第 3 回全体研究会	東京大学駒場キャンパス	染谷昌義による発表	7 名
2019/2/23	2018 年度第 4 回全体研究会	東京大学駒場キャンパス	尾形哲也氏による発表	8 名
2019/3/23	シンポジウム「人工知能の哲学 2.0 の構築に向けて」	東京大学駒場キャンパス	柴田正良氏・黒崎政男氏・松原仁氏および鈴木貴之・立花幸司・染谷昌義の講演と総合討論	約 80 名
2019/3/24	討論会「拡張概念をめぐって–メディア論・技術論・心の哲学」	高千穂大学	柴田崇・上杉繁の講演と総合討論	10 名
2019/5/10	2019 年度第 1 回全体研究会	東京大学駒場キャンパス	立花幸司による発表	7 名
2019/8/29,30	2019 年度第 2 回全体研究会	北海学園大学	飯塚博幸氏・小野哲雄氏による講演と、植原亮・鈴木貴之による発表	9 名
2019/11/10	ワークショップ「機械学習・深層学習の哲学的意義」	慶應義塾大学三田キャンパス	日本科学哲学会第 52 回大会におけるワークショップ。鈴木貴之・植原亮および大塚淳氏による提題。	約 80 名

2019/11/26, 27	Japanese-European Meeting on Artificial Intelligence and Moral Enhancement	グラナダ大学（スペイン）	立花幸司・鈴木貴之・植原亮・染谷昌義・上杉繁による講演。	約 15 名
2020/8/8	オンライン講演会	オンライン	久保昭教氏による講演	約 30 名
2020/10/12	2020 年度第 1 回全体研究会	オンライン	鈴木貴之・立花幸司・柴田崇・上杉繁による発表	7 名
2020/11/6	Panel Session: Artificial Intelligence as a Tool	オンライン（主催：トゥウェンテ大学）	Philosophy of Human-Technology Relations Conference 2020 におけるパネルセッション。鈴木・立花・柴田・上杉による提題。	約 25 名
2021/1/23	ワークショップ「思考力とウェルビーイング」	オンライン	植原亮と信原幸弘氏・松浦和也氏・渡邊淳司氏による提題	約 30 名
2021/3/20	Shiawase2021 ワークショップ「哲学者とともに考える AI 時代のわたしたちのウェルビーイング」	オンライン	立花幸司・植原亮・中澤栄輔による提題	20 名
2021/6/29	2021 年度第 1 回全体研究会	オンライン	古宮嘉那子氏による講演	12 名
2021/7/25	ワークショップ「AI 設計におけるリスクとその「可解性」について–「ゴリラ化問題」と「ミダス王問題」を皮切りに」	オンライン	直江清隆氏・三枝亮氏による提題	約 25 名
2022/2/18	2021 年度第 2 回全体研究会（予定）	オンライン	古賀高雄氏による講演	

2022/2（日程調整中）	2021 年度第 3 回全体研究会（予定）	オンライン	井上悠輔氏による講演	
2022/3/11	ワークショップ「とくと人工知能」（予定）	オンライン	鈴木貴之・立花幸司・植原亮による提題	
2022/3（日程調整中）	総括シンポジウム（予定）	オンライン	鈴木貴之・立花幸司・染谷昌義による提題	

６―１―２．書籍、DVD など論文以外に発行したもの

(1) 植原亮、『思考力改善ドリル―批判的思考から科学的思考へ』、勁草書房、2020 年 10 月
(2) 染谷昌義、「反復なき反復としてのわざ―動作の哲学から浮かび上がるわざの本性」、床呂郁哉（編著）『わざの人類学』、京都大学出版会、2021 年 11 月

６―１―３．ウェブメディア開設・運営

(1) プロジェクトウェブサイト（URL：http://updatingphilosophyofai.net/、2019 年 2 月開設）

６―１―４．学会以外のシンポジウムなどでの招へい講演　など

(1) 鈴木貴之、「人工知能と自然知能―代用品？上位互換？それとも新種？」、第 28 回自然科学研究機構シンポジウム「SF／未来／科学技術」、2019 年 8 月 24 日、国際交流会議場
(2) 柴田崇、「サイボーグ論の転回：『拡張』思想の射程の考察から」、早稲田大学創造理工学研究科総合機械工学科専攻主催講演会、2019 年 10 月 3 日、早稲田大学
(3) 上杉 繁、「人間と技術の「間」をデザインする―人文学的知見を踏まえた工学デザイン・工学教育の可能性―」、北陸信越工学教育協会石川県支部金沢工大部会　学術講演会、2019 年 12 月、金沢工業大学
(4) 柴田崇、「サイボーグ思想における不便益の意義」、第 31 回不便益システム研究会、2020 年 12 月 7 日
(5) 中澤栄輔、「脳科学分野における ELSI」、応用脳科学アカデミーベーシックコース 3「ELSI」第 1 回、NTT データ経営研究所、2020 年 10 月 30 日
(6) 上杉 繁、「経験拡張技術のための「拡張」概念の整理と設計指針としての拡張性の検討」、第 19 回認知的コミュニケーションワークショップ、2020 年 11 月、オンライン
(7) 上杉 繁、「人間拡張技術におけるジレンマの分析と解決方法の検討」、自動車技術会マルチメディア部門委員会、2020 年 12 月、オンライン

(8) 染谷昌義、「身体性と運動性―キスの制御則に示される心のはたらき―」、玉川大学応用脳科学研究センター「心の哲学研究研究部門」第 14 回研究会、2021 年 3 月 6 日、オンライン
(9) 上杉 繁、「意図せざる使用へ備えるための観察・分析・実践におけるデザインの拡張, デザインと意図せざる使用 ―工学・経済学・哲学からの考察―」、第 2 回 科学技術倫理セミナー、2021 年 12 月、オンライン
(10) 鈴木貴之、「人工知能の哲学 2.0 に向けて」、東京大学次世代知能科学研究センター連続シンポジウム第 6 回「AI 時代の哲学を考える」、2022 年 1 月 21 日、オンライン

６－２．論文発表

６―２―１．査読付き（2 件）

(1) Shigeru Wesugi. (2019) “Analysing and Solving the Reduced-ability and Excessive-use Dilemmas in Technology Use.” *Proceedings of the 22nd International Conference on Engineering Design*. 1393-1402. (DOI:10.1017/dsi.2019.145)
(2) 柴田崇、「メディア研究と心理学の接点:『探索モデル』」、『新人文学部』、北海学園大学大学院文学研究科、第 17 巻、50-67 頁、2020 年 12 月

６―２―２．査読なし（6 件）

(1) 染谷昌義、「アフォーダンスからの希望」、『臨床心理学』、第 20 巻第 2 号、136-141 頁、2020 年 3 月
(2) 染谷昌義、「二元論の向こう側を探る自然学のプログラム」、『現代思想』、第 48 巻 8 号、187-195 頁、2020 年 6 月
(3) 植原亮、「作り物の徳認識論の規範性」、『情報研究』、関西大学総合情報学部、第 51 号、1-20 頁、2020 年 8 月
(4) 柴田崇、「AI vs. IA： 論争に隠れた真の課題」、『人文論集』、北海学園大学人文学部、第 70 巻、115-126 頁、2021 年 3 月
(5) 鈴木貴之、「深層学習の哲学的意義」、『科学哲学』、日本科学哲学会、53 巻 2 号、151-167 頁、2021 年 3 月（依頼論文）
(6) 植原亮、「人工知能は科学を人間から切り離してしまうのか？」、『セミナー年報 2020』、関西大学経済・政治研究所、69-82 頁、2021 年

６－３．口頭発表（国際学会発表及び主要な国内学会発表）

６―３―１．招待講演（国内会議 3 件、国際会議 0 件）

(1) 柴田崇、「メディア研究と心理学の接点：『探索モデル』」、公開シンポジウム「ネットメディアの生態心理学」、日本心理学会第 84 回大会、公開期間 2020 年 9 月 8 日～11 月 2 日、オンライン（録画）
(2) 鈴木貴之、「テクノロジー、幸福、朗働」、第 28 回日本産業ストレス学会シンポジウム「これからの働き方を考える」、2020 年 12 月 5 日、オンライン
(3) Takayuki Suzuki. “Transparency in AI: Identifying the Real Issue.” Symposium: Fairness, Integrity and Transparency of Formal Systems: Challenges for a Society Increasingly Dominated by Technology. 科学基礎論学会 2022 年度総会と講演会、2021 年 6 月 19 日、オンライン

６―３―２．口頭発表（国内会議 3 件、国際会議 15 件）

(1) 上杉繁（早稲田大学)、「ロボット技術と人間との関係におけるジレンマへのアプローチ」、応用哲学会第 11 回年次研究大会、京都大学吉田キャンパス、2019 年 4 月 21 日
(2) 鈴木貴之（東京大学)、「深層学習の哲学的意義：認知科学の哲学と人工知能の哲学の場合」、日本科学哲学会第 52 回大会ワークショップ「機械学習・深層学習の哲学的意義」、慶應義塾大学三田キャンパス、2019 年 11 月 10 日
(3) 植原亮（関西大学)、「機械学習・深層学習と知的創造性」、日本科学哲学会第 52 回大会ワークショップ「機械学習・深層学習の哲学的意義」、慶應義塾大学三田キャンパス、2019 年 11 月 10 日
(4) Shigeru Wesugi. Waseda University. “Designing approaches addressing dilemma in relating to robots.” The 21st Conference of the Society for Philosophy and Technology. Texas A&M University in College Station. 2019/5/20-22.
(5) Koji Tachibana. Kumamoto University. “AI-based moral enhancement and the future of human virtue.” Third International Workshop On Ethics And Human Enhancement. University of Granada, Spain. 2019/6/3.
(6) Shigeru Wesugi. Waseda University. “Analysing and Solving the Reduced-ability and Excessive-use Dilemmas in Technology Use.” The 22nd International Conference on Engineering Design. Delft University of Technology, The Netherlands. 2019/8/5-8.
(7) Koji Tachibana. Kumamoto University. “Artificial Intelligence and Epistemic Virtues. Virtue, Media, and Democracy.” University of Genova, Italy. 2019/9/26.
(8) Koji Tachibana. Kumamoto University. “An extended reply to Lara and Deckers

(2019).” Japanese-European Meeting on Artificial Intelligence and Moral Enhancement. University of Granada, Spain. 2019/11/26.

(9) Takayuki Suzuki. The University of Tokyo. “Toward an Update of Philosophy of Artificial Intelligence.” Japanese-European Meeting on Artificial Intelligence and Moral Enhancement. University of Granada, Spain. 2019/11/27.

(10) Ryo Uehara. Kansai University. “Could AI be a creative machine?” Japanese-European Meeting on Artificial Intelligence and Moral Enhancement. University of Granada, Spain. 2019/11/27.

(11) Someya Masayoshi. Takachiho University. “What the 21st century's philosophy of AI should consider: Human-machine hybrid nature.” Japanese-European Meeting on Artificial Intelligence and Moral Enhancement. University of Granada, Spain. 2019/11/27.

(12) Shigeru Wesugi. Waseda University. “Design tool for analyzing human-AI technology relation.” Japanese-European Meeting on Artificial Intelligence and Moral Enhancement. University of Granada, Spain. 2019/11/27.

(13) Koji Tachibana. Kumamoto University. “AI and the cultivation of human moral emotion. Budapest Workshop on Philosophy of Technology.” Budapest University of Technology and Economics, Hungary. 2019/12/12

(14) Shigeru Wesugi. Waseda University. “Conceptual Tool for Designing Human Extension Technologies.” Panel Session: Technology and Human Body: An Interdisciplinary Approach. Philosophy of Human-Technology Relations Conference 2020. Online (University of Twente). 2020/11/4.

(15) Takayuki Suzuki. The University of Tokyo. “Two Concepts of Artificial Intelligence.” Panel Session: Artificial Intelligence as a Tool. Philosophy of Human-Technology Relations Conference 2020. Online (University of Twente). 2020/11/6.

(16) Koji Tachibana. Kumamoto University. “Artificial Intelligence as A Tool for Moral Education.” Panel Session: Artificial Intelligence as a Tool. Philosophy of Human-Technology Relations Conference 2020. Online (University of Twente). 2020/11/6.

(17) Takashi Shibata. Hokkai Gakuen University. “AI vs. AI: The Real issues hidden in the struggle.” Panel Session: Artificial Intelligence as a Tool. Philosophy of Human-Technology Relations Conference 2020. Online (University of Twente). 2020/11/6.

(18) Shigeru Wesughi. Waseda University. “Considerations on Analysing Relations between Humans and AI Technologies Based on Archetypes of Instruments - Club-type

and Pot-type.” Panel Session: Artificial Intelligence as a Tool. Philosophy of Human-Technology Relations Conference 2020. Online (University of Twente). 2020/11/6.

６―３―３．ポスター発表（国内会議 0 件、国際会議 0 件）

６－４．新聞/TV 報道・投稿、受賞など

６―４―１．新聞/TV 報道・投稿

(1) 読売新聞、2019 年 2 月 21 日、東京夕刊（人工知能の将来的可能性に関する鈴木貴之のコメントが掲載された。インタビュー時には本プロジェクトへの取り組みについても説明したが、記事には言及なし。）

６―４―２．受賞

なし

６―４―３．その他

(1) 植原亮、「人工知能(1)不気味さの諸相と「汝自身を知れ」」、『文部科学教育通信』、499 号、22-23 頁、2021 年（解説記事）
(2) 植原亮、「人工知能(2)ＡＩと人間にとっての創造性の意味」、『文部科学教育通信』、500 号、22-23 頁、2021 年（解説記事）

６－５．特許出願

６―５―１．国内出願（0 件）

６―５―２．海外出願（0 件）