公開

研 究 終 了 報 告 書

「AIと人の相互作用による技術哲学の創出」

研究期間: 2021年10月～2024年3月
研 究 者: 中尾 悠里

1. 研究のねらい

本課題のねらいは、信頼できる人工知能（AI）の実現に向けて、人間の自律性の担保に必要な技術の条件など、あるべき技術像を考える際の基盤となる技術哲学を、人と技術の相互作用を考慮した工学的に実現可能なものに刷新することである。

技術哲学は、これまでに技術決定論的に技術が人や社会に求める在り方を議論する古典的な技術哲学、実際の事例の記述・分析に重点を置く経験論的転回、倫理的な議論を基盤とする倫理的転回の三つの潮流を経て発展してきた（Verbeek 2011）。現代の技術哲学は主にAI倫理のように倫理学的議論を応用倫理として技術に適用する倫理的なアプローチが主流である。しかし、このアプローチでは、近年の個人にパーソナライズするAI技術により増大したユーザーの変化が技術に与える影響は考慮されていない。これにより技術開発の現実から技術哲学が乖離している。

そこで本課題では、人間の価値観の変化に伴ってインタラクティブに変化するAI技術を用いて人とAIのインタラクションを調査し、その結果を基に技術哲学分野を人と技術の相互作用を考慮した工学的に実現可能なものに刷新することを狙う。

本課題では、技術哲学に工学的な実現可能性をもたらすために、ユーザーの価値観に合わせて変化するAIシステムを実際に開発・評価することで工学的実証を基に技術哲学が規範を与える方法を創出することを狙う。また、技術哲学が人と技術の相互作用を考慮できるようにするために、技術の利用の中での人の価値観の変化をモデリング・技術哲学に適用可能な相互作用の図式を創出し、技術哲学に対して実証結果に基づく議論を可能とする基盤を提供する。

これらの実施を通じ、人文社会学的な議論と工学的な実践を架橋し、技術哲学を工学的な観点から刷新することで、人間と技術のインタラクションに関する新たな規範を提示することを狙う。

（参考文献：Peter-Paul Verbeek 2011: Moralizing Technology: Understanding and Designing the Morality of Things, The University of Chicago Press ；鈴木俊洋『技術の道徳化』法政大学出版会， 2015）

2. 研究成果

（1）概要

本課題の目的は、人間の価値観の変化に応じて動的に変化するAIシステムの開発と評価を通じて、技術哲学を人と技術のインタラクションによる変化を考慮する学問へと刷新することである。本課題は大きく分けて3つの研究項目に分かれ、各項目の成果は以下の通りである。

- **研究項目1：人が持ちうる価値観を包括する機械学習手法の類型化・開発：**
  本研究項目の当初の目標は、人の価値観の機械学習手法へのマッピングだったが、このアプローチが困難であると判明し、研究内容を当初の目標から修正した。結果的に、人々の技術に対する価値観を調査するワークショップ手法「Reflexive CoDesign」を開発し

た。この手法では、専門家と非専門家が共同で価値観の表明、施策の考案、技術的施策の評価を行う。特徴としては、研究ガバナンスの研究分野の一つ、「責任ある研究・イノベーション」の知見を取り入れ、ワークショップ参加者が、他のステークホルダーが持ちうる価値観を考慮し、施策の批判的評価を行うことを可能にした点である。このワークショップ手法を情報検索の技術を対象に実施し、技術に関連する多様な価値観を抽出した。

- **研究項目２：人と相互作用し価値観を反映するグラフィカルユーザーインターフェイス（GUI）の開発・評価**：
  本研究項目の目標は、Reflexive CoDesign により抽出したユーザーの価値観に基づき、インタラクティブなウェブ検索システムを開発し、その長期にわたる参加者評価を通じて技術を使う人の価値観がどのように変化するかを調査することである。開発手順には、ワークショップから得られた情報の整理、インターフェイスの詳細なイメージ作成、施策のコード化、システムの実装が含まれる。開発されたシステムを用いて長期評価が行われる予定であり、それによって情報技術の利用がユーザーの価値観に与える影響を調査する。
- **研究項目３：人の価値観の変遷のモデリングを通じた人と技術の相互作用の図式化**：
  本研究項目は、研究項目２の評価の結果を基に進められる予定で、現在は未達成であるが、2024 年度の 10 月までに完了することを計画している。

（2）詳細

【本課題の目的】：人間の価値観の変化に伴ってインタラクティブに変化する AI ツールの開発・評価することにより、技術哲学の人と技術の相互作用に基づく学問へアップデートする。

【研究の方法】：本課題の研究項目は下記の三つである。

- 研究項目１：人が持ちうる価値観を包括する機械学習手法の類型化・開発
- 研究項目２：人と相互作用し価値観を反映するグラフィカルユーザーインターフェイス（GUI）の開発・評価
- 研究項目３：人の価値観の変遷のモデリングを通じた人と技術の相互作用の図式化

項目１・２では、ワークショップを行うことで人々の価値観と機械学習が内包する価値観をマッチングし、人間の多様な価値観を包括できるインタラクティブな AI ツールを開発する。項目３では、調査結果を基に一般的な人と AI のインタラクションの詳細を図式化し、技術哲学の潮流を創出する種とする。

【成果と達成状況】：

項目一：人が持ちうる価値観を包括する機械学習手法の類型化・開発

●目標との変更点：元の提案では本項目において、人が持ちうる価値観を調査すると同時に、システムの開発前に人が持ちうる価値観に対応した機械学習手法を検討・創出し、それぞれの手法が価値観をどの程度反映しているかをパラメタライズする予定を立てていた。しかし、研究を進める中で人の価値観の機械学習手法へのマッピングが予想より困難であると判明したため、機械学習の要素技術への価値観のマッピングは研究項目２の結果に基づいて行うことに変更した。従って、本項目では人々が技術（今回の場合はウェブ検索の技術）について持つ価値観を調査するためのワークショップ手法の提案と実施が主な研究成果である。

●研究成果：

i) **Reflexive CoDesign ワークショップ手法の開発：**

公開

人々の価値観が技術との相互作用でどのように変化していくかを調査するツールを開発するため、人々が技術を用いてどのような価値観を実現したいかを表現し、その価値観に合わせた技術を設計するワークショップ手法 Reflexive CoDesign ワークショップの手法を開発した。この手法は、参加型デザインや人間中心デザインの文脈から発展してできた共同デザインの手法（Sanders & Stappers 2008）と、研究開発ガバナンスの一分野である責任ある研究・イノベーション（Responsible Research and Innovation, RRI）の手法を統合したものである。

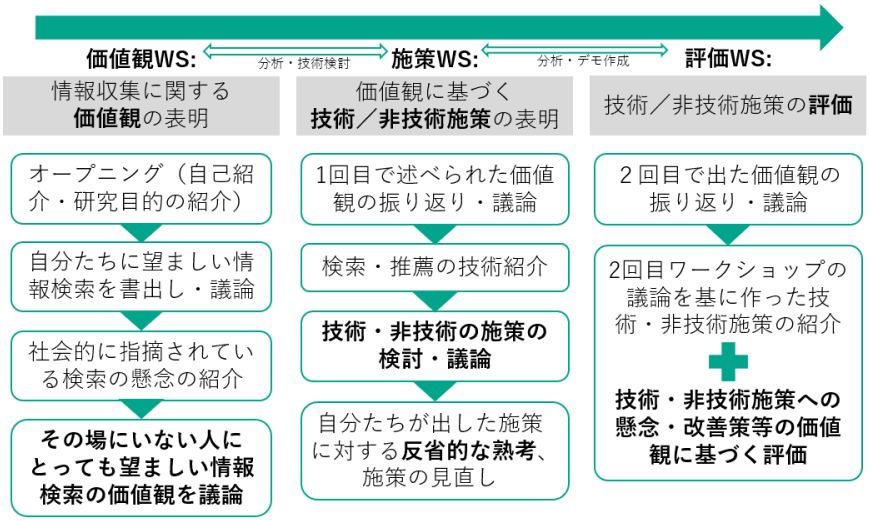

図 1　Reflexive CoDesign ワークショップ手法の概要。太字の部分が reflexive activity にあたる。

本手法では、専門家と非専門家がワークショップに参加し、価値観の表明（価値観ワークショップ（WS））、施策の考案（施策 WS）、技術的施策の評価（評価 WS）の 3 回のワークショップを通じて、非専門家が専門家の意見に誘導されずに、対等な立場で意見を表明することが可能である。

本手法の特徴は、ワークショップ参加者以外のステークホルダーが持ちうる価値観を参加者に考えてもらう活動や、参加者自身が考えた施策の批判的な評価を促すための活動を省察的活動（reflexive activity）を称して設けた点である。Reflexive activity では、参加者に対して、ある施策を考えたきっかけ・その際に考慮した要因・ありうる代替案を聞く活動といった活動も行った。これは、科学研究者の行った意思決定に対して、学術研究よりも広い文脈を意識し、再考を促す Midstream Modulation（Fischer 2007）という RRI 領域の考え方基づいて設計されている。これらの活動を通じ、reflexive CoDesign では、単純にワークショップに参加した参加者だけでなく、多様な人が持ちうる価値観に配慮できるように工夫されている。

ii)　**Reflexive CoDesign ワークショップの実施：**

上述の Reflexive CoDesign ワークショップを AI 技術の一つのアプリケーションとして情報検索を対象に実施した。

参加者として、日本全国より、ウェブの習熟度が高い参加者と低い参加者 5 名ずつの計 10 名に参加してもらい、価値観 WS、施策 WS、評価 WS の計 3 回のワークショップを実施した。各々のワークショップの間に一月ほどの間を設けるため、全体として 3 か月間にわたり同一の参加者が参加した。

結果として、価値観 WS では情報収集や情報検索に関する価値観ついて次の 11 カテゴリの価値観が抽出され、施策 WS では技術施策 21 カテゴリと非技術施策 16 カテゴリが価値観についての 11 カテゴリとの対応を含めて抽出され、施策 WS と評価 WS で施策についての改善

点やトレードオフを含む評価を収集することができた。

評価 WS では価値観 WS と施策 WS の結果を基に GUI のプロトタイプも作成した。

項目２：人と相互作用し価値観を反映するグラフィカルユーザーインターフェイス（GUI）の開発・評価

●研究成果：

項目１で収集した価値観・施策に基づいてインタラクティブなウェブ検索システムの開発を行い、その評価実験を行う。2024 年 3 月現在、2024 年 7 月～10 月の参加者実験を目指してシステムを開発中であるため、以下には開発の手順と設計した評価手法を記載する。

i) **Reflexive CoDesign ワークショップ手法に基づく技術開発手順：**

Reflexive CoDesign ワークショップで得た知見を、下記のステップに従って実際のアプリケーションとして実装している。

●**Step 1. ワークショップのコード・発言をまとめる**：施策 WS で抽出された施策にコードをマッピングし、同時に評価 WS のときに得られた批判的評価を施策ごとにまとめる。

●**Step 2. 施策をインターフェイス上で実現した際のラフイメージの作成**：技術施策を画面上に表示した際にどのように表示できるかのラフスケッチを描く。

●**Step 3. 施策に関するコードの重複を整理**：異なる施策の中で似た機能が求められている場合、それらを列挙し、技術的な実現可能性に基づいて整理する。

●**Step 4. 機能整理**： 未発表部分のため省略。

●**Step 5. 実装**：Step 4 で作成した構成に基づき、開発を行う。

ii) **開発システムの長期評価（計画）：**

上記ステップで開発した検索システムについて長期評価を試みる。長期評価は、情報技術を長期間利用する際にユーザーが持つ技術に対する価値観がどのように変化するかを調査する目的で行う。詳細は未発表部分のため省略。

項目３：人の価値観の変遷のモデリングを通じた人と技術の相互作用の図式化

本項目については項目２で行った評価の結果行われるため、現状未達成だが、2024 年 10 月中を目標に完了することを計画している。

【参考文献】：Fisher, E. (2007). Ethnographic invention: Probing the capacity of laboratory decisions. *NanoEthics*, 1(2), 155-165.

Sanders, Elizabeth B N., and Pieter Jan Stappers. (2008). Co-creation and the new landscapes of design. *Co-design,* 4(1), 5-18.

３． 今後の展開

本課題では、上述の通り、Reflexive CoDesign と技術開発、長期評価からなる「人の価値観を反映した技術開発手法」を開発し、その手法に基づいてインタラクティブなウェブ検索システムを開発している。

この成果を踏まえた今後の研究開発の展開として、**1～2 年の短期的**には、本課題で創出した人の価値観を反映した技術開発手法の一般への公開と他技術への適用による汎用化、

公開

開発したインタラクティブなウェブ検索システムの一般への公開、また、得られた知見の技術哲学としての整理が挙げられる。特に、本課題の主たる成果である人の価値観を反映した技術開発手法は、ウェブ検索だけでなく多様な技術に適用できるべきものであり、他の技術に適用することでより汎用的なものにすることができる。

また、**今後 5 年程度の中期的**には、人の価値観を反映した技術開発手法を用いて、人々が多様な技術開発に取り組めるような状態を確保していきたい。本手法の普及によって、人々が多様な技術を開発する際に人々の価値観を考慮することができ、また技術への適用結果を技術開発手法に対してフィードバックしてもらうことで、より多くの人に使いやすい手法へと発展させていきたい。

最後に、**今後10年程度の長期的**には、人々の価値観を考慮した技術を長期で変革していくというループを社会に根付かせ、人間の自律性を担保できる AI 技術によって人間がサポートされる社会を実現する。多様な人の価値観が入り込んだ AI 技術が、多様な技術や意思決定の中に利用され、様々な人にとって技術を信頼することのできる状態になっていることを狙いたい。

4. 自己評価

研究目的の達成状況としては、最終的な目標である技術哲学の刷新にはまだ届いていないが、人間の価値観の変化に伴ってインタラクティブに変化する AI ツールの開発・評価に関しては、人の価値観の抽出は完了し、AI ツール開発のための機能整理が完了した状態である。そのため、研究項目の前半部は達成したという状態である。本課題での技術開発・評価を通じて、多様な人々の技術に関する価値観を知る方法の創出・実行を完了し、技術実装に繋がる手順を高い透明性を持って整理することができる見込みであり、技術哲学の刷新の基盤の創出に確実に着手できた。人文系の学問領域である技術哲学と、技術を実装する工学領域を結び付けるもの他に例を見ない活動ができたと考えている。

研究の進め方（研究実施体制及び研究費執行状況）に関しては、全体を通して完全に研究代表者単独の研究体制であったため、実施期間の最中で当初予定していなかった職務・学務が入るなどした場合、予定通りに進捗させることが難しい場合があった。これに関しては、予測外の事象に備えて予定より早めに進捗させることを心がけ、外部研究者との共同研究の実施など、より効果的に進める方法を考えることで改善すると考えている。

研究成果の科学技術及び社会・経済への波及効果としては、人工知能学会での本課題の発表、日本ロボット学会誌上での本課題の解説記事の出版の機会等を頂くことができ、工学の分野にも技術哲学や責任ある研究・イノベーションの方法論を用いた技術開発の可能性をアピールできた。しかしながら、もちろんまだ科学技術についても、社会・経済についても波及効果は十分とは言えない。今後、この課題で生み出した知見を広め、生成 AI の普及により今までより更に一般社会に普及していく AI 技術が人の価値観を踏まえた機能を提供し、人間が AI 技術を信頼できる社会を作ることで、科学技術の発展につなげ、また科学技術を用いた社会に人と技術の良好な関係をもたらし、それが経済発展の源泉となる社会になることを見込んでいる。

5. 主な研究成果リスト

公開

（1）代表的な論文（原著論文）発表

研究期間累積件数：0件

（2）特許出願

研究期間全出願件数：0 件（特許公開前のものも含む）

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

●学会発表（国内）：

1. 中尾悠里. 人と共に変化する AI 倫理のための共同デザインワークショップ. 人工能学会全国大会論文集. 2023 年, 37 巻, 4B3GS1102, 1-4.
2. 中尾悠里, 藤垣裕子. 技術設計過程における「反射性」の組み込み ～ユーザーとの共同デザイン方法論の検討. 科学技術社会論学会第 21 回年次研究大会. 2022 年 11 月 27 日.
3. 中尾悠里, 藤垣裕子. RRI の省察性／Reflexivity が研究開発実践にとって持つ意味：異分野との比較. 科学技術社会論学会第 22 回年次研究大会. 2023 年 12 月 10 日.

●総説

1. 中尾悠里. AI と人の相互作用に基づく技術哲学を導く設計手法. 日本ロボット学会誌. 2023 年, 41 巻 8 号, 696－699.