様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月 １４ 日現在

機関番号：３２６８９
研究種目：研究活動スタート支援
研究期間：2022～2023
課題番号：２２Ｋ２０３２０

研究課題名（和文）人工知能の倫理的判断に対する期待の実証的検討

研究課題名（英文）An empirical investigation of the expectations of AI's ethical decisions

研究代表者
谷辺　哲史（Tanibe, Tetsushi）

早稲田大学・文学学術院・講師（テニュアトラック）

研究者番号：２０９６４４８０

交付決定額（研究期間全体）：（直接経費）　　2,100,000円

研究成果の概要（和文）：本研究は倫理的な問題に関わる人工知能（AI）の判断について、AIの専門家ではない一般の人々の反応を定量的に調査した。研究1では、自動車の速度に関わる矛盾した規範（法定速度の遵守と周囲の流れ（実勢速度）に合わせること）が存在する場面での自動運転車の挙動に対する評価を調査した。その結果、人間が運転する場合に比べて、自動運転では法定速度を遵守することが望ましいと判断される傾向が示された。研究2では、AIの判断が性別によるバイアスを生じさせたとき、人間が同様の判断をする場合と比べると否定的な印象が緩和される可能性が示された。ただしいずれの研究でも効果量は小さく、結果の頑健性は今後の検討課題である。

研究成果の学術的意義や社会的意義
本研究の成果は、AIの開発や社会実装を社会的に受容される形で進めていく際に参照できる経験的な知見を提供する点で社会的意義があると考えられる。人間とAIでは期待される判断内容が必ずしも同じではないことや、同じ判断をしても異なった印象を与えうるという結果は、AIの判断にどの程度の柔軟性を持たせるかという設計上の問題や、AIに判断を任せてよい範囲を見極めるという運用上の問題を検討する際に役立つことが期待できる。また学術的な意義としては、倫理的判断の内容が同じでも判断主体が人か否かによって適切さの評価が変わることを定量的手法によって確認し、倫理に関して人々が持っている判断基準を解明することに貢献した。

研究成果の概要（英文）：This study quantitatively examined the responses of the general public, who are not AI experts, to artificial intelligence (AI) judgments related to ethical issues. Study 1 investigated the evaluation of self-driving car behavior in the presence of contradictory norms related to car speed (compliance with legal speed and matching to the surrounding flow (actual speed)). The results showed a tendency to judge that compliance with legal speed is preferable in self-driving compared to driving by humans. Study 2 showed that when AI judgments were biased by gender, the negative impression may be alleviated compared to when humans make similar judgments. However, the effect sizes were small in both studies, and the robustness of the results remains to be examined.

研究分野：社会心理学

キーワード：人工知能　自動運転　規範　公正感

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

本研究は近年の人工知能（AI）研究の発展と社会的な浸透を背景として実施された。2010 年代頃からの AI 研究は第三次 AI ブームと呼ばれる進展を見せ、AI や、AI によって制御される自律ロボットが社会の広い範囲で活用されるようになっている。これまで人間が行っていた意思決定を AI の活用によって自動化できるようになり、その応用領域が拡大していくと予想されている中で、判断や行動の結果に関する責任をどのように判断するか、あるいは AI に倫理的な判断能力を備えさせることは可能かといった問題が、倫理学、法学の分野で盛んに議論されている（ウォラック・アレン，2009；久木田ほか，2017；平野，2017）。

このような新たな技術の開発やそれに関わる新たな制度が社会的に受容され、人々の利益につながる形で普及していくためには、技術自体の高度化に加えて、非専門家の人々が AI に何を期待しているかを知ることも必要である。つまり、AI 技術や倫理学・法学の専門家ではない一般市民が AI に関して抱く態度や、その態度の背後にある認知過程を明らかにすることが、AI を社会的に受容される形で活用していくために有用だと考えられる。

しかし、AI という新たな技術を人々がどのように認知し、どのような態度を抱いているのか――特に、倫理的な判断の主体としての AI にどのような役割を期待しているか――ということは、まだ明らかになっていない部分が多い。

２．研究の目的

上記の背景を踏まえ、本研究は「AI と人間では、社会的に期待される倫理判断が異なるか」という問いに取り組んだ。

AI の活用によって判断を自動化するとき、求められていることは人間と同じ判断を代替することなのかは明らかではない。私たちの日常的な行動の中では、明示的な規則に従っていなくてもある程度柔軟な判断をすることが許容されたり、むしろそれができないと「融通が利かない」といった否定的な評価を受けることさえある。しかし、人間から判断を委ねられた AI がそのような柔軟性を持つことを、私たちは受け入れられるだろうか。本研究では、複数の規範が対立しうる場面での AI の判断に対する人々の評価を定量的に調査し、人々が AI に対して抱く期待の実態を解明する。

さらに、AI の判断が社会や個人に及ぼす影響についても、それを人々がどのように認識するかは重要な問題である。人間がこれまで行っていた判断を AI によって代替したとき、判断の内容ではなく、判断主体が人間か AI かという違いによって、その判断の適切さは異なるように評価されるかもしれない。

本研究では上記の 2 つの観点（複数の規範が存在する場面での選択、選択によって生じた結果の適切さ）から、人間の判断と AI の判断に対して一般の人々が示す反応を定量的手法を用いて調査した。

３．研究の方法

(1) 規範が対立する場面での選択への期待

対立する規範が存在する場面として自動車の運転における速度制限を取り上げ、判断の適切さに対する人々の評価を調査した。2023 年 1 月にウェブ調査を実施した（N = 221、平均年齢 42.1 歳）。クラウドソーシングサービスを通じて参加者を募集し、日本国内に在住し日常的に自動車を利用する人（週に 1 回以上を目安とした）が調査に回答した。

具体的には、実勢速度（多くの自動車が実際に走っている速度）が法定速度を超えている中で、自分の前方の車（自動運転車または人間が運転する自動車）が法定速度で走行しているために車の流れが滞るという状況を文章で提示し、法定速度を守る自動車に対する評価を求めた。加えて、暗黙の規範（法定速度に違反する実勢速度）に対する態度の個人差を統制するため、普段の回答者自身の行動傾向（法定速度と実勢速度のどちらに合わせるか）も尋ねた。

(2) バイアスのある判断に対する反応

統計データに基づく判断が不平等な結果を生む場面として、金融機関の与信審査における男女差を取り上げ、判断の適切さに対する人々の評価を調査した。2023 年 11 月にウェブ調査を実施した（N = 1000、平均年齢 46.3 歳）。調査会社の回答者パネル登録者を対象として調査を行い、日本国内に在住する 1000 名が回答した。

調査では、2019 年に報道された判断バイアスの事例を基にした架空の事例を文章で提示し、それに対する評価を尋ねた。具体的には、個人向けの融資審査において、性別以外の情報が同じ

である場合に男性の方が有利な条件で融資を受けやすいという男女差が生じるというものである。ただし提示する文章には 4 つのパターンがある。審査が AI を利用した自動審査システムによるものか（AI あり vs なし）、審査結果に差が生じる理由が詳述されているか否か（離職率の高さなどから将来の平均的な収入に差が生じることが審査に反映されるという説明を加える、または説明を行わない）という 2 つの要因が操作されており、各回答者は 4 条件のうち 1 つのみをランダムに提示された。文章を提示した後、この銀行の審査に対する印象（信頼できる、公正だ、など 7 項目）を尋ねた。

４．研究成果

(1) 規範が対立する場面での選択への期待

法定速度を守る車への評価をポジティブ評価（適切な走行をしている、安心するなど 4 項目；α＝.89）、ネガティブ評価（迷惑だ、融通がきかないなど 4 項目；α＝.92）に分けて、提示したシナリオ（自動運転 vs.人間）と回答者の運転の傾向によって評価が異なるかを分析した。

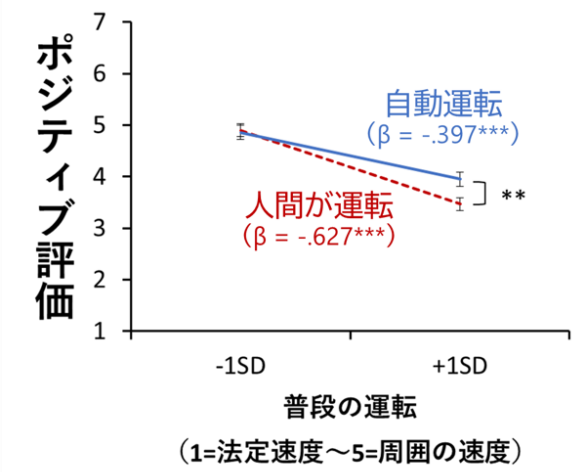
分析の結果、普段から実勢速度（周囲の車の速度）に合わせる傾向が強い人ほど、法定速度を守る車へのポジティブ評価は低く、ネガティブ評価は高いという想定通りの相関関係が確認された。しかし、ポジティブ評価についてはシナリオの違いの効果があり、自動運転のシナリオを読んだ場合には、自身の運転の傾向が評価に与える影響が小さくなった（図 1）。言い換えると、普段から実勢速度に合わせて走行する人は、他者が実勢速度よりも法定速度を優先していると否定的な態度を示すが、その相手が人間ではなく自動運転車だった場合にはそれほど評価を下げなかった。

一方で、ネガティブ評価についてはシナリオの効果は有意ではなかった。

この結果は、人が AI に対して期待する倫理的な判断基準は、他者に対する期待と比べると暗黙の規範に従うことを期待されておらず、法律などの明示的な規範に従うべきだと見なされていることを示唆する。すなわち、AI は明示的な規則に従った杓子定規な判断をしても人間ほど否定的には評価されず、むしろ柔軟な判断を期待されていないのかもしれない。ただし本研究ではポジティブ評価の分析でのみこのような効果が見られ、ネガティブ評価についてはシナリオの効果が見られなかった。結果の頑健性については今後の研究を通じて引き続き検討していく必要がある。

なお、本研究の結果は 2023 年度の日本グループ・ダイナミックス学会大会で報告された。

図 1 法定速度を守る車への評価



(2) バイアスのある判断に対する反応

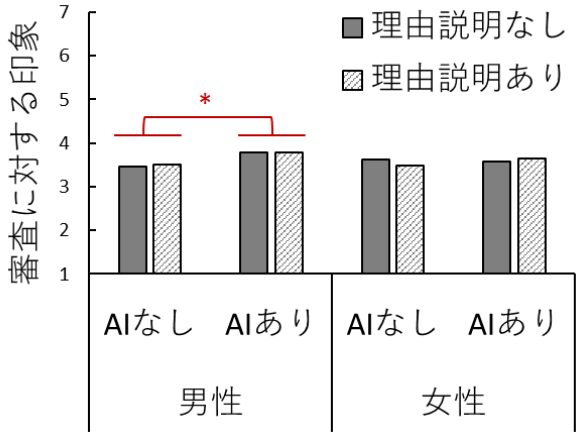
審査の主体（AI 利用の有無）と理由説明の有無、回答者の性別の 3 要因分散分析によって審査に対する印象を比較した。審査主体の主効果が有意で、AI で審査するというシナリオの方が肯定的に評価されていた。ただし、いずれの条件でも平均値が尺度上の中点（4＝ どちらともいえない）を下回っており、基本的には否定的な評価を受けていたことには注意が必要である。一方で、理由説明の有無は主効果、交互作用ともに有意ではなく、印象に影響を与えるとはいえない結果だった。

また、審査の主体と回答者の性別の交互作用は統計的に有意ではなかったものの（$p < .10$）、AI の判断の方が肯定的に評価されるという単純主効果が男性において顕著になる傾向が見られた（図 2）。

上記の結果は条件間の差が小さいことに留意が必要だが、AI の活用が不平等な結果をもたらすことに対する社会の反応について、倫理的な問題が生じうることを示唆するものである。過去の統計情報に基づいて何らかの社会的カテゴリー（性別、人種など）に属す人が不利な扱いを受けることは、人間の判断でも起きていることであり、それ自体は AI の導入によって新たに起きる問題とはいえない。しかし本研究では、結果として生じる判断のバイアスは同じであっても、AI の判断によってそれが生じたという情報を付加することで否定的な印象がやや緩和されることが示された。責任帰属に関する心理学研究の知見は、行為者の意図などの心的状態が責任帰属の大きさに影響することが知られているが、AI の判断には人間のような意図が存在しないため、そもそも道徳的な問題として認識されにくくなるのかもしれない。AI の判断の結果が道徳的な問題と認識しにくくなるとしたら、結果として生じている不平等な状態が看過され、問題の解消が遅れるおそれもある。本研究では金融機関での審査の男女差という具体例に焦点を当てたが、今後は研究結果の頑健性や一般化可能性を幅広く検討しつつ、そうした経験的証拠と規範的な議論とのつながりについても考察していくことが課題となる。

なお、本研究は研究期間の終盤に調査を実施したため成果の発表に至らなかったが、2024 年度の学会で結果を報告する予定である。

図 2 シナリオ別の審査に対する印象評定



５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 谷辺哲史
２．発表標題 規範が競合する場面での人工知能の判断に対する市民の期待
３．学会等名 日本グループ・ダイナミックス学会第69回大会
４．発表年 2023年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関