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

谷 辺 哲 史

Public expectations of the judgment of artificial intelligence in a situation of normative conflict

Tetsushi TANIBE

Abstract

This study conducted an online survey (n = 221) to examine how people expect artificial intelligence (AI) to make judgments in situations where explicit rules, such as laws, and actual people's behavior are in conflict. The present study focused on the speed limit of automobiles as a typical situation where two norms conflict, and as a field in which the practical application of AI is advancing. Survey participants were presented with a scenario in which one car followed the legal speed limit while many cars were going faster than the limit, and were asked about their positive and negative impressions of the car. The car following the legal speed limit was either a self-driving car or a conventional human-driven car, and participants were randomly presented with one of the two scenarios. The results showed that: (1) the impressions were not different between the scenarios, but (2) the relationship between participants' usual driving behavior and their positive impression of the car following the legal speed limit was moderated by the scenario; although those who adjust their speed to match surrounding cars tend to have a less positive impression of the car that followed the legal speed limit, this tendency weakened in the self-driving car scenario. This result indicates that people may expect AI to follow explicit rules rather than to make flexible judgments that allow for overriding those rules, as humans do. Although empirical findings such as those of this study do not directly answer normative questions of how AI should be used, they can provide a foundation for more plausible normative debates.

問題

暗黙の規範が存在する場面での AI 利用の問題

本研究の目的は、複数の規範が競合する状況において一般市民が人工知能（AI）に期待する選択が、人間同士で期待される選択と同じであるかを経験的手法によって検討することである。

2010 年代以降の人工知能研究は第 3 次 AI ブームと呼ばれる盛り上がりを見せ、医療、翻訳、創作などさまざまな用途での利用が広がっている。活用場面が拡大するのに伴って、AI を安全に活用するための制度設計などの倫理的・法的・社会的課題（ELSI）についても議論が行われている（松尾, 2015; 西垣・河島, 2019）。

AI の活用に関わる問題の中で本研究が着目するのは、複数の「正しい」判断があり得る場面での選択である。私たちの日常生活の中には、明文化された規範と、多くの人が実際に選択する行動が一致しない状況が存在する。たとえば、道路の制限速度が標識によって明示されているにもかかわらずほとんどの車が制限速度よりも速い速度で走行するような状況である。ほかにも、エスカレーターは立ち止まった状態で利用することが推奨されているにもかかわらず、実際にはエスカレーター上で歩く人も存在するし、歩くスペースを空けるために

片側に寄って乗ることがマナーだという風潮さえある。多くの人々が実際にとっている行動がその場における適切な行動の基準となることはしばしばあり、そこで成立した規範を記述的規範という（Cialdini et al., 1990; 本稿ではこれ以降、法令などの明示的な規範と対比して暗黙の規範という用語を用いる）。このような場面で人が何を「正しい」行動と見なすかは、法令などで示される正しさと必ずしも一致しない。

自動車の運転操作のように従来は人間が行っていた判断を AI が代替するようになれば、複数の規範が存在する場面でどの規範に従うかを AI の判断に委ねることになる。このような状況で AI がどのような判断を行えばよいのかという問題には、以下の 2 通りの答え方があり得る。

1 つ目の考え方は、人間と同様の柔軟性を AI に持たせるというものである。AI を導入する目的は人間の判断を代替し、人間の労力を削減することであるという観点からは、人間が実際に行っているのと同じ判断を AI もするべきだという考え方が導ける。この考え方に従えば、AI にとっての「正解」は多くの人間が実際に行っているのと同じ判断をすることである。人間は自分自身の判断の理由を常に意識できるとは限らないため、人間と同じ判断をする AI を作ろうと考えたとき、どのような場合に法令に違反してよいかの判断基準を明示的にプログラムに書き込むことはできない。しかし近年の機械学習技術の発展を前提とすれば、さまざまな状況での人間の行動データを収集し、その行動の背後にある判断の規則を AI で再現することは（技術的なハードルはあるにせよ原理的には）不可能ではないだろう。

しかしこのような考え方に対しては、法令に反する行動を可能にする AI の設計を本当に許容してもよいのかという疑問が生じる。そこで 2 つ目の考え方は、AI には人間のような柔軟な判断を許さず、法令に違反する行動を常に禁じるというものである。

どちらの考え方を採用するかが決まれば、その基準に沿って「正しい」判断をする AI を開発することは可能であるが、どちらの考え方に基づいて AI を設計するかは社会的な合意によって決められる問題である。AI 技術の実用化を円滑に進めるためには、市民がどのような AI の実現を望んでいるかを明らかにすることが重要である。

AI に対する人々の期待に関する先行研究のレビュー

それでは人々は AI に対して、人間と同じような判断をすることを望むのだろうか。倫理的判断と AI の関連を扱った先行研究は、倫理的な判断主体としての AI が人間とは異なるものと認識されていることを示唆している。以下では関連する研究を 3 つ紹介する。

道徳ジレンマの一種であるトロッコ問題（トロリー問題）[(1)]を題材とした実験（Malle et al., 2015）は、人間または自律ロボットの倫理的な選択への評価を実験参加者に尋ね、判断主体の違いによって許容度が異なるかを検討した。何もしなければ 4 人が死亡するが、トロッコの進路を変えれば 4 人が助かるかわりに別の 1 人が死亡するという場面で、人間の操作者または自律ロボットが進路を変える、または変えないというシナリオを提示し、その選択が道徳的に間違っている（morally wrong）かを尋ねると、人間が判断したシナリオでは義務論的判断（介入しない）よりも功利主義的判断（進路を変える）の方が悪いと評価されたのに対して、ロボットの判断に対しては反対に義務論的判断の方が悪いと評価された。つまり、人間と AI では適切と評価される倫理的判断が異なり、AI は功利主義的な判断を期待されていた。

別の研究では、人が AI を倫理的な判断主体とは見なしておらず、倫理的な判断の責任はあくまで人間に帰属されることが示されている。自動運転車による交通事故を題材とした調査（谷辺・唐沢, 2021）では、AI に対する原因帰属と人間（自動車のユーザー、メーカー）に対する原因帰属・責任帰属がトレードオフの関係に

(1) 複数の倫理的な規則が対立する状況での選択を問う思考実験。トロッコのブレーキが故障して暴走し、このまま走行すれば複数の作業員が死亡するが、転轍機でトロッコの進路を変えると、複数の作業員が助かるかわりに変更後の進路にいる別の 1 人の作業員が轢かれて死亡するという場面で、進路を変えることは倫理的に許されるかを問う。この場面では犠牲者の数を最小化するべきという功利主義的な選択と、意図的な行為によって他者に害を与えることは許されないという義務論的な選択が対立する。トロッコ問題はもともと中絶の是非を論じるために提案された思考実験だが、本稿の題材である自動運転と関連する論点を整理した解説として笠木（2021）がある。

あるかを検討したが、実際にはAIに原因を帰属する人ほどユーザーやメーカーにも原因や責任を帰属するという正の相関関係があった。つまりAIが自律的に判断したとしても、その判断の結果に関する責任は人間にあると見なされていた。

最後に、そもそも人はAIに倫理的な判断を行わせたいのかということも調査されている。Bigman & Gray（2018）は受刑者の仮釈放の可否の判断や、医療場面でのリスクを伴う治療法の選択、自律兵器による攻撃の可否の判断といった他者の利害に影響を与える意思決定について、AIに判断を委ねることを許容できるかを回答者に尋ねた。その結果、いずれの意思決定課題でも、人間の判断者に比べてAIに判断をさせることは許容できないと評価された。この評価はAIには人間と違って心がないという信念に媒介されて生じていることも示された。

本研究の仮説

前節でレビューした研究結果を総合すると、人々はAIを人間と同等の倫理的な判断主体とは認めていないと考えられる。そして本研究の問いである明示的な規範と暗黙の規範の対立に対しては、人々はAIに対して明示的な規範に従うことを期待しているのではないかと予測できる。

暗黙の規範に従うということは、法令などで明確に定められた判断基準に従わないことをあえて選ぶということである。AIに倫理的な判断を任せることに対して人々が否定的な態度を示したり、AIの判断の結果であっても人間に責任を帰属したりするという知見からは、人はそのような柔軟な判断をAIに期待しておらず、むしろ杓子定規に規則に従うことを期待するのではないかと考えることができる。

また、道徳ジレンマ状況で少数を犠牲にするという功利主義的な判断が、AIが判断する場合には人間が判断する場合ほど悪いと見なされないという実験結果も、AIに対して期待される倫理的な判断が人間の判断者に対する場合とは異なることを示している。AIが人間とは異質な判断主体と認識されているのであれば、人間ならば周囲から否定的に評価されるような判断であっても――たとえば法定速度を誰も守っていない道路で律儀に法定速度に従って車の流れを滞らせることが、人間の場合には「融通がきかない」などと否定的な評価を受けるとしても――AIはさほど否定的に見られないだろう。

上記の議論を踏まえ、本研究は「明示的な規範と暗黙の規範が対立する状況において、人はAIに対して、人間に対する場合と比べて明示的な規範に従うことを強く期待する」という仮説を設定する。

自動車の自動運転と暗黙の規範

上記の仮説を検証するために、AIが利用される場面の中でも、自動車の自動運転を題材として調査を行う。自動車の運転は、法令に基づく制限速度という規範が明確に存在する一方で、現実には多くの車が制限速度を超えて走行する状況がしばしば起きる（実際に走行している速度を実勢速度という）。つまり、明示的な規範と暗黙の規範の対立が分かりやすく生じる状況であるため、本研究の仮説を検証するための題材とした。

研究の実践的な意義という点でも、自動運転はAIの活用が進んでいる代表的な領域であり、AIに対する一般市民の期待を調査することの意義は大きい。2023年には、限定された領域ではあるが人間の運転者による操作を必要としないレベル4の自動運転が法的に可能になり、市民が実際に自動運転車に乗車することが現実のものになっている。また、自分自身がユーザーとして乗車する以外にも、自動運転車が普及する過渡期には人間の運転する車と自動運転車が道路上で混在することになるため、交通の流れを乱すことなく2種類の車が走行できるかという問題もある。そこで本研究では、自動運転車が周囲の車の運転者に受容されるかという点に焦点を当て、人々が自動運転車にどのような走行を期待しているかを調査する。

したがって前節で述べた仮説をより具体的に言い換えると、本研究で検証する仮説は「法定速度と実勢速度が乖離する状況において、人は自動運転車に対して、人間の運転者の場合と比べて法定速度に従って走行することを期待する」となる。

方法[2]

参加者

2023 年 1 月にウェブ調査を実施した。インターネットを介したクラウドソーシングサービスのクラウドワークスを通じて参加者を募集した。募集条件は日本国内に居住し、自動車を週 1 回以上の頻度で運転する人とした。参加者はクラウドワークスのウェブサイトから調査ページにアクセスし、各自の所有するパソコン、スマートフォン等の端末を用いて調査に回答した。

調査ページにアクセスした 300 名のうち回答に不備があった 79 名を除外したため[3]、有効回答は 221 件となった。回答者の性別の内訳は男性 112 名、女性 107 名、無回答 2 名で、年齢は 20 歳から 68 歳（$M = 42.09$, $SD = 9.60$；無回答 3 名）だった。

調査票の構成

自動車運転の場面想定　冒頭で自動車の利用状況を尋ねた後、自動車の運転に関する架空のシナリオを提示した。提示したシナリオは回答者自身が自動車を運転している状況で、周囲のほとんどの車が法定速度の 50km/h を超える 60km/h 〜 65km/h で走行しているが、自分の前にいる車だけが 50km/h で走行しているというものだった。

シナリオの内容は 2 通りあり、どちらか 1 つが無作為に提示された。一方のシナリオは前方の車が人工知能を用いた完全な自動運転で（AI 条件）、他方のシナリオは運転者に関する情報を明示しなかった（人間条件[4]）（シナリオの全文を付録に掲載した）。

法定速度を守る運転者の評価　シナリオを読んだ後、前方を走る車に対する考えを尋ねた。質問項目は 8 項目あり、いずれも「1 = 全く当てはまらない」から「7 = 非常に当てはまる」の 7 件法で回答を求めた。「適切な走行をしている」、「安心する」、「信頼できる」、「今の速度のまま走るべきだ」の 4 項目（$\alpha = .89$）の回答を加算平均したものをポジティブ評価、「迷惑だ」、「融通がきかない」、「いらいらする」、「もっと速度を上げるべきだ」の 4 項目（$\alpha = .92$）の回答を加算平均したものをネガティブ評価とした。順序効果を避けるため、これら 8 項目をランダムな順序で表示した。

暗黙の規範に対する態度　次に、提示したシナリオの内容とは無関係に、自動車の運転に関する規範に対する考え方を尋ねた。ただし、AI 条件の参加者には「自動運転車の設計について、あなたの考えをお聞きします」、人間条件の参加者には「自動車の運転について、あなたの考えをお聞きします」という文を質問項目の前に提示した。質問項目は「周囲の状況に合わせることが最も重要だ」、「状況によっては、法律で決められた制限速度を超えるほうが適切だ」、「法律で決められた制限速度が実態に合っていない場合は、制限速度を超えてもかまわない」、「法律で決められた制限速度を守ることが最も重要だ」（逆転項目）の 4 項目で、「1 = 全く当てはまらない」から「7 = 非常に当てはまる」の 7 件法で回答を求めた。これら 4 項目はランダムな順序で提示された。値が大きいほど暗黙の規範（実勢速度）に従うことに肯定的な態度を表すように逆転項目の回答を変換し、4 項目の回答（$\alpha = .85$）を加算平均したものを暗黙の規範への支持とした。

普段の運転行動と適切さの判断　回答者自身の普段の運転について、実勢速度（周囲の車の速度）が制限速

---

(2)　本研究の実施にあたって、調査実施時に著者が所属していた埼玉県立大学の研究倫理委員会にて承認を受けた（受付番号 22060）。

(3)　まず、自動車を運転しない、または頻度が週 1 回未満の回答者 41 名を除外した（クラウドワークスの画面上で募集条件を記載したが、対象外の者が参加することをシステム上で妨げることができないため、募集対象に該当するか確認するための質問項目を設けた）。次に、不注意回答者のチェック項目（暗黙の規範に対する態度を尋ねるページで「読み取り確認のため、この質問には『当てはまる』と答えてください」という質問を設けた）に指示通り回答しなかった者 4 名、提示したシナリオの内容に関する正誤問題で誤答した者 34 名を除外した。

(4)　現在の自動車の実態を踏まえると、特に情報がなければ人間が運転していると考えることが自然であり、その点に言及する方が不自然な印象を与える可能性があると判断したため、運転者を明示しなかった。なお、人間条件では調査全体を通じて「人工知能」や「自動運転」といった文言は一度も使われていない。

度を超えている状況でどのように行動しているか、およびどのような行動が適切だと思うかを尋ねた（いずれも「1＝法律上の制限速度に合わせる」から「5＝周囲の車の速度に合わせる」の5件法）[5]。

結果

データ分析は統計分析ソフトのR version 4.5.0（R Core Team, 2025）とHAD version 18.008（清水, 2016）を用いて行った。

暗黙の規範に従わない車への評価・暗黙の規範に対する態度の条件間比較

法定速度に従う車へのポジティブ評価、ネガティブ評価と暗黙の規範に対する態度の分布は図1のようになった。AI条件と人間条件で回答に差があるかをそれぞれWelchのt検定によって検討したが、いずれの指標も条件間で有意な差はなかった（ポジティブ評価：$t(212.47) = 1.30$, $p = .196$；ネガティブ評価：$t(207.82) = 0.98$, $p = .329$；暗黙の規範への支持：$t(212.07) = 0.31$, $p = .760$）。

また、これらの変数どうしの相関関係は表1のようになった。いずれの変数間でも中程度から強い相関関係があったが、AI条件におけるポジティブ評価とネガティブ評価の相関関係（$r = -.67$, $p < .001$, 95% CI [-.76, -.54]）は人間条件（$r = -.82$, $p < .001$, 95% CI [-.87, -.76]）よりも弱くなっていた。

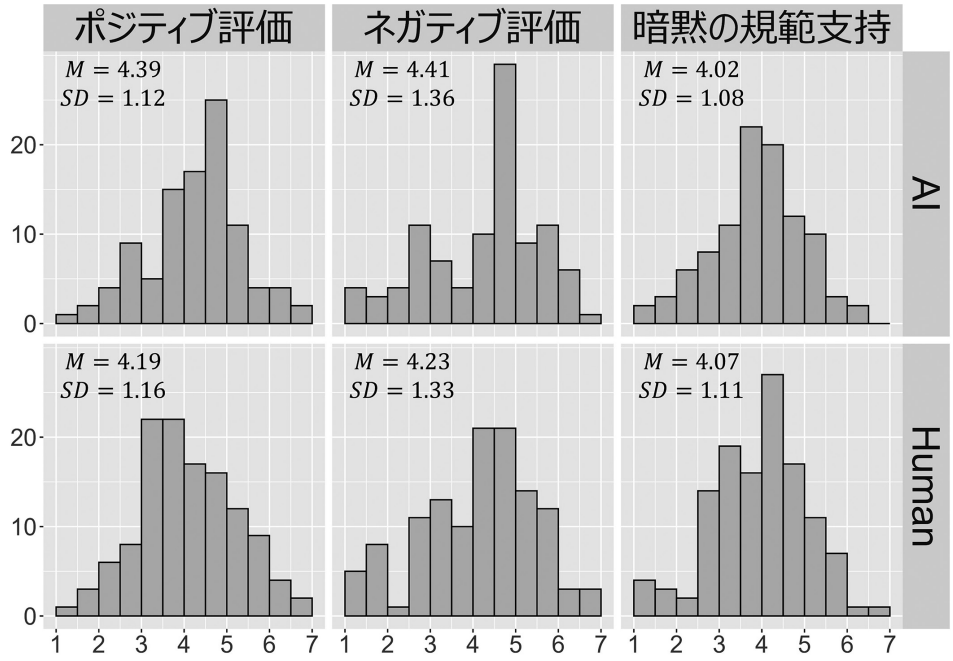

図1　法定速度を守る車への評価、暗黙の規範に対する態度の分布

表1　法定速度を守る車への評価、暗黙の規範に対する態度の相関関係

	ポジティブ評価	ネガティブ評価	暗黙の規範支持
ポジティブ評価		-.67***	-.59***
ネガティブ評価	-.82***		.67***
暗黙の規範支持	-.69***	.69***	

（注）　対角線の右上はAI条件、左下は人間条件での相関係数を表す。***$p < .001$

(5)　適切さについての回答はこれ以降の分析には用いていないが、行動傾向との相関関係があり（$r = .78$, $p < .001$）、適切さに関する回答と実際の行動傾向は概ね一致していた。

規範遵守傾向による調整効果

次に探索的な検討として、回答者自身が普段の運転で暗黙の規範（実勢速度）と明示的な規範（法定速度）のどちらに従うかという行動傾向の個人差と、法定速度に従う車への評価や暗黙の規範に対する態度との関連を分析した。

法定速度に従う車へのポジティブ評価と回答者の行動傾向の関係を図 2 に示した。シナリオ（0 = 人間条件, 1 = AI 条件）と行動傾向（値が大きいほど実勢速度に従うことを表す）、さらにこれら 2 変数の交互作用を説明変数とする重回帰分析を行った。その結果、行動傾向の主効果（$\beta = -.512$, $p < .001$）、行動傾向とシナリオの交互作用（$\beta = .115$, $p = .046$）が有意だった。シナリオの主効果は、AI 条件の方がポジティブ評価が高い傾向があったが、統計的に有意ではなかった（$\beta = .095$, $p = .095$）。このモデルの決定係数は $R^2 = .298$ だった（$F(3, 217) = 30.69$, $p < .001$；自由度調整済み決定係数 $Adjust\ R^2 = .288$）。

交互作用が有意だったため単純主効果の分析を行った。どちらのシナリオでも行動傾向の単純主効果が有意であり、実勢速度に従う傾向が強い人ほど法定速度に従う車に対するポジティブ評価が低かったが、その効果は AI 条件（$\beta = -.397$, $p < .001$）よりも人間条件（$\beta = -.627$, $p < .001$）でより大きかった。また、運転行動の傾向によってシナリオの効果が異なるかを分析すると、法定速度に従う傾向が強い場合(6)（平均−1SD）の平均値は人間条件では 4.90、AI 条件では 4.86 となり、シナリオの単純主効果がなかった（$\beta = -.019$, $p = .812$）。一方、実勢速度に従う傾向が強い場合（平均＋1SD）には人間条件で 3.47、AI 条件で 3.95 となり、AI 条件でポジティブ評価が高かった（$\beta = .210$, $p = .010$）。

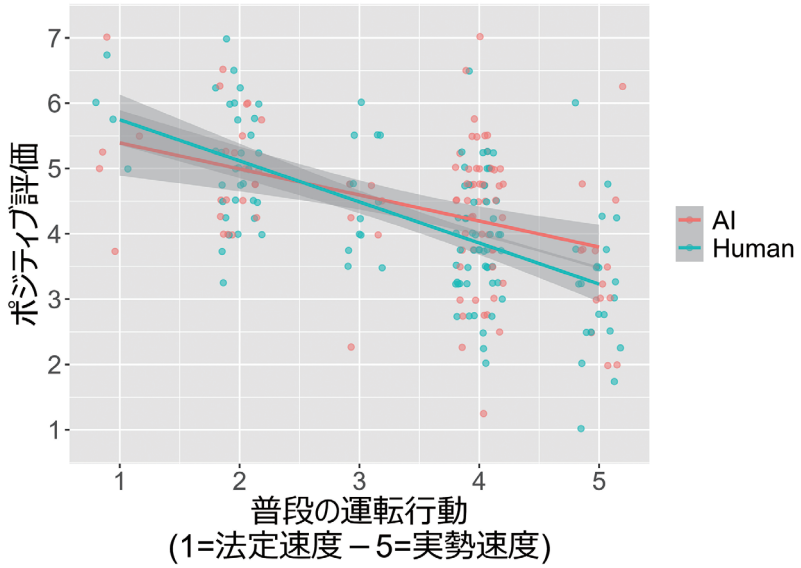

図 2　運転行動の傾向と法定速度に従う車へのポジティブ評価の関連

ネガティブ評価、暗黙の規範に対する態度についても同様のモデルで重回帰分析を行ったが、行動傾向との関連はあったものの、シナリオとの関連は見られなかった。

ネガティブ評価を目的変数とした重回帰分析（$R^2 = .332$, $Adjust\ R^2 = .323$, $F(3, 217) = 35.94$, $p < .001$）では行動傾向の主効果があり、実勢速度に従う人ほどネガティブ評価が高かった（$\beta = .571$, $p < .001$）。シナリオの主効果（$\beta = .057$, $p = .308$）、行動傾向とシナリオの交互作用（$\beta = -.013$, $p = .817$）は有意ではなかった。

同様に、暗黙の規範に対する態度を目的変数とした重回帰分析（$R^2 = .399$, $Adjust\ R^2 = .391$, $F(3, 217) = 48.00$, $p < .001$）でも行動傾向の主効果があり、実勢速度に従う人ほど暗黙の規範に肯定的な態度を示した（β

(6)　普段の運転行動は連続量として分析に用いたため、条件ごとに推定された回帰式に基づいて、普段の運転行動の値が平均値から標準偏差 1 つ分だけ離れた場合のポジティブ評価を推定した。本研究では、普段の運転行動の回答は平均 3.48、標準偏差 1.14 だったので、この値が 2.34 と 4.62 の場合の推定値によって条件間の比較を行った。

= .616, p < .001）。シナリオの主効果（β = -.031, p = .560）、行動傾向とシナリオの交互作用（β = -.081, p = .129）は有意ではなかった。

考察

AI と人間の判断に対する評価の差異

本研究では自動車の運転における法定速度と実勢速度の乖離を題材として、一般市民が AI の判断に対して期待する倫理的な判断を明らかにするために調査を行った。AI は人間の判断者と比べて、明示的な規則に従うことを期待されているという仮説を設定し、仮想的な場面で明示的な規範に従った判断者に対するポジティブ・ネガティブな評価と、暗黙の規範を優先することに対する態度という 3 つの指標を用いて仮説を検証した。しかし、いずれの指標でも判断主体が AI の場合と人間の場合で有意な差は見られず、仮説は支持されなかった。

仮説を支持する結果が得られなかった理由として、調査内で提示したシナリオの内容が参加者の反応に意図しない影響を与えていた可能性がある。AI 条件では、車に乗っている人はいるが操作をしていないという状況を提示して判断を求めた。この状況は、たとえば「速さと法令遵守のどちらを優先するか」などの優先順位を設定することで、搭乗者が運転操作にある程度の影響を及ぼせる立場にあったと解釈することも可能である。そのように解釈したとしたら、AI 条件でも搭乗者が倫理的な行為主体と見なすことができる。そのため AI 条件でも実際には人間の意図的な判断に対する評価が行われており、人間条件と AI 条件の差異が生じなくなったと考えられる。本研究では探索的な分析の一部でのみシナリオの効果が見られたが、AI 条件の搭乗者の関与についての想定が参加者の解釈に委ねられていたことで条件内の誤差が大きくなり、条件間の差を統計的に検出できなかったために、部分的に差が見られるという曖昧な結果になったのかもしれない。

2 つ目の要因として、調査の題材として想定した暗黙の規範が、実際には想定ほど強い規範として受け入れられていなかったと考えられる。今回の調査では、いずれの指標も平均値が尺度の中点付近となり、明示的規範と暗黙の規範という 2 つの規範の選択について中庸な態度が示された。つまり、自動車の速度に関して、法令よりも周囲の流れに合わせることを優先すべきだという規範が広く共有されているとはいえない状況であり、本研究が前提として想定していた「法令などの明示的な規範と、人々が実際に従っている規範が対立している」という状況が成り立っていないため、本研究の仮説を検証する題材として不適切だった可能性がある。現実には実勢速度が法定速度を超えることがあるが(7)、自動車を運転する当事者がその状況を肯定的に受け入れて周囲の流れに合わせているのではなく、多元的無知（集合的無知；Miller, 2023; 神, 2009）の状況が生じているのかもしれない。つまり、多くの運転者は法定速度を超過することを望ましくないと考えているが、周囲の他者の真の態度を正確に知ることができないため、周囲の他者が法定速度を超過しているという事実から「他の人は、法定速度を超えても構わないと考えているのだろう」と推測し、その推測された「規範」に従っているだけということである。

ただし、今回の調査への回答は「本来、法令には従うべきだ」という社会的望ましさの影響を受けたものである可能性もある。自己報告型の質問紙調査では、回答者は社会的に望ましいと見なされる性質を自分自身も持っていると回答する傾向がある（Edwards, 1953）。本研究で測定した態度は Edwards（1953）が用いたパーソナリティ特性のような心理尺度とは異なるが、法令に違反することの是非という、公的に望ましいとされる回答が存在する質問であった。そのため、匿名の調査とはいえ法令違反を容認するような態度を表出することを避けて回答が歪んだ可能性を否定できない。

状況の解釈の曖昧さ、暗黙の規範の存在の確からしさへの疑問、社会的望ましさによる回答の歪みという 3 つの可能性を考慮すると、調査の題材とする具体的な場面の内容や態度の測定方法を改善し、妥当性を高めた手続きで仮説を検証する必要がある。

---

(7) 本研究では主要な変数の測定後に、シナリオで示したような実勢速度が法定速度を超える状況を普段どのくらい経験するかを尋ねた。回答者のうち 122 名（55.2%）が「運転するときはほとんどいつも」、86 名（38.9%）が「ときどきある」と回答しており、実勢速度と法定速度の乖離が実際に起きていることは確認できた。

行動傾向の個人差との関連

探索的な検討として、回答者自身が明示的な規範と暗黙の規範のどちらに従っているかという行動傾向の個人差が、AI または人間の判断に対する期待と関連するかを分析した。その結果、明示的な規範に従う判断主体へのポジティブ評価とネガティブ評価、暗黙の規範に対する態度の 3 つの指標で一貫して得られた結果として、回答者自身の行動傾向を肯定する回答のパターンが確認できた。すなわち、普段から実勢速度に従う人ほど、法定速度を守る車に対するポジティブ評価は低く、ネガティブ評価は高く、自動車の運転や自動運転車の設計において実勢速度に従うことに肯定的な判断を示した。

自分自身の行動を肯定するような態度を表明すること自体はごく自然に予測できる反応である。もともと法定速度よりも実勢速度に従うことに肯定的な態度を持っている人は、実際に両者が乖離する場面では実勢速度に合わせて走行するだろう。反対に、行動と信念を一致させることで認知的不協和を低減する（Festinger & Carlsmith, 1959）という認知的な傾向ゆえに、実勢速度に合わせて走行するという経験を通じて、自らの行動を肯定するような態度が強くなったということも考えられる。いずれにせよ普段の行動傾向と質問に対して表明する態度が相関することは特に驚く結果ではない。

しかし本研究で興味深いのは、実勢速度に従う傾向が強い人々において、法定速度に従う自動運転車のポジティブ評価が人間の運転者に対する場合よりも高かったことである。自己の行動傾向とは異なる規範に従う他者に対して評価が否定的になるのは自然なことだが、その反応が AI の判断に対しては弱くなっていたということであり、人々は人間の他者と AI に対して異なる基準で評価を行っていたことになる。暗黙の規範に従う人であっても、AI が明示的な規範に従うことを人間の他者の場合ほど否定的には評価しないという結果は、「AI は人間と比べて明示的な規範に従うことを期待される」という本研究の仮説と一致するものであった。

一方で、法定速度に従う傾向が強い人々は、判断主体の違いにかかわらず法定速度に従う車をポジティブに評価した。普段から法定速度に従う人々は、自己の行動傾向と同じ行動を肯定的に評価することと、AI に明示的な規範に従った判断を期待することとがどちらも法定速度に従う車への肯定的な態度につながるため、シナリオ間で差がなく同程度に肯定的な反応になったと考えられる。

ただし、普段の行動傾向とシナリオとの交互作用があったのはポジティブ評価のみだった。実勢速度に従う傾向が強い人に着目して分析結果を整理すると、彼らは「人間の運転でも自動運転でも実勢速度に従うことが適切だと考えるし、法定速度に従う自動運転車に対しては相手が人間の場合と同じようにネガティブに評価する。しかし同時に、人間の場合と比べるとポジティブにも評価する」という両面的な態度を示していることになる。この両面性は、ポジティブ評価とネガティブ評価の相関関係が AI 条件では比較的弱くなっていたことにも表れている。つまり、人間の他者に対してはポジティブに評価することとネガティブに評価しないことが表裏一体になっているのに対して、AI に対してはネガティブとポジティブ両面の評価がある程度区別されたものとして同時に生じているといえる。本研究で提示したシナリオに即して解釈すると、普段から実勢速度に合わせている人々は自動運転車も実勢速度に合わせるべきだと考え、流れに乗らない走行をする自動運転車に対しては「迷惑だ」などの否定的な反応が生じるが、一方で AI が明示的な規則に従っていることに対して「安心する」「信頼できる」といった感覚もある程度生じるということである。これは仮説通りではなかった結果に対する事後的な解釈であることに注意が必要だが、人々は AI の判断の結果（杓子定規な判断によって不利益を被ること）への評価と、AI の判断過程の適切さへの評価を切り分けて判断しているのかもしれない。

本研究の実践的な意義

AI と人間に対する評価の違いはポジティブ評価という 1 つの指標のみで得られたものではあるが、将来の AI の活用のあり方について議論するための材料を与えてくれる。AI は従来人間が行っていた判断を自動化することで人間にとって利益をもたらす技術である。しかしユーザーとなる一般市民が AI に期待することはあらゆる場面で人間と同じ判断を再現できることではなく、むしろ杓子定規に規則に従う AI の方が求められているのかもしれない。前節で述べた、判断の結果への評価と判断過程への評価の区別という観点も踏まえると、AI に判断を任せたことで多少の不利益があったとしても、適切な過程を経て判断した結果であれば納得する

という受容のあり方も考えることができる。AIの判断が現在の社会における人々の実際の判断のあり方とは異なっていたとしても、AIにはAIに期待される判断のしかたがあるという前提で、社会的に受容されるAIの開発を進めることが重要になる。

また、先行研究（Malle et al., 2015; Bigman & Gray, 2018; 谷辺・唐沢, 2021）の知見と合わせて考えると、AIと人間に対する期待が異なることの背景には、AIはあくまで機械であるため柔軟な判断を行うには限界があるといったAIに対する一般的なイメージや、AIがどれだけ自律的に振る舞うように見えても必ず人間の開発者が存在し、判断に伴う責任は人間に帰属されるためAI自体に高い柔軟性を持たせるべきではないといった責任の所在に関する判断があると考えられる。しかし近年の生成AIの性能向上と普及に鑑みると、現在以上に人間らしく振る舞えるAIが実現し、AIが機械であるという感覚は薄れていくかもしれない。そうすると、対立する規範の中でどれを優先するかといった価値判断までAIが行ってくれるというようなAIに対する過信が生じるおそれがある。市民が正しい知識に基づいてAIを利用し、その恩恵を十分に受けられるように、開発者やサービス提供者も適切な情報提供を続けることが求められるだろう。

本研究の限界と今後の展望

本研究の結果は部分的に仮説と一致するものであったが、この結果が得られたのは3つの指標のうち1つだけであり、それも探索的な分析の中で見出されたものであった。今回の分析結果から直ちに仮説が支持されたとはいえず、今後の研究によって経験的な証拠を蓄積していく必要がある。また、本研究で扱った指標は全般的なポジティブ、ネガティブな印象や、どのように行動することが適切だと思うかという規範意識であったが、AIを適切に利用するための制度設計の議論を視野に入れるならば、より具体的な規制への賛否などに焦点を当てて市民の態度を明らかにすることも重要な課題となる。

次に、前節の最後で述べた通り、AIに対する期待が形成される過程にはAIに対する知識やイメージが影響していると考えられるが、この心理過程を実証的な証拠に基づいて明らかにすることも今後の研究課題となる。期待の形成に影響する要因を特定できれば、専門家から市民への情報提供をどのように行うべきか、またAIを用いたサービスを提供する事業者はどのような情報を開示すればよいかといった、情報の適切な扱い方についての指針を得られる。

最後に、経験的な問いと規範的な問いの区別には注意が必要である。本研究は一般市民の意識に関する経験的な研究である。社会的課題を議論する際に市民の期待は判断材料となるが、市民の期待を明らかにすることで「AIにどのような判断をさせるべきか」あるいは「AIの開発、利用に対してどのような規制を設けるべきか」という問いへの答えが得られるわけではない。倫理的な観点からどのような判断が正しいといえるか、またAIに関わる法制度がどのような内容になるべきかといった課題は、倫理学や法学といった諸領域で議論されることになるだろう。その際に本研究のような経験的な手法を通じて得られた知見は、規範的議論の前提として人々の直観的な倫理的判断や社会通念といった事実に関する情報が必要になる場面で、より妥当性の高い証拠に基づいた議論を可能にすることに貢献できるだろう(8)。

文献

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34.

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58(6)*, 1015–1026.

Edwards, A. L. (1953). The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *Journal of Applied Psychology*, *37(2)*, 90–93.

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology*, *58(2)*, 203–210.

神信人（2009）．集合的無知　日本社会心理学会（編）社会心理学事典（pp. 300–301）丸善出版

笠木雅史（2021）．自動運転の応用倫理学の現状と課題――自動運転車とトロリー問題――　日本ロボット学会誌, *39(1)*, 22–27.

(8)　規範的な問いと経験的証拠の関係については鈴木（2020）を参照。

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pp. 117–124.

松尾豊（2015）．人工知能は人間を超えるか――ディープラーニングの先にあるもの――　KADOKAWA

Miller, D. T. (2023). A century of pluralistic ignorance: What we have learned about its origins, forms, and consequences. *Frontiers in Social Psychology*, *1*, 1260896.

西垣通・河島茂生（2019）．AI 倫理――人工知能は「責任」をとれるのか――　中央公論新社

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/（2025 年 5 月 28 日閲覧）

清水裕士（2016）．フリーの統計分析ソフト HAD ――機能の紹介と統計学習・教育，研究実践における利用方法の提案――　メディア・情報・コミュニケーション研究, *1*, 59–73.

鈴木貴之（編）（2020）．実験哲学入門　勁草書房

谷辺哲史・唐沢かおり（2021）．自動運転による事故とメーカー、ユーザーに対する責任帰属　実験社会心理学研究, *61(1)*, 10–21.

付録

○ AI 条件で提示したシナリオ

近年、**人間の運転操作をまったく必要としない、完全な自動運転**の開発が進んでいます。近い将来には、完全な自動運転車と通常の自動車が同じ道路を走るようになるかもしれません。このような状況での自動車の運転について、あなたの考えをお聞きします。

あなたが自動車を運転しているところを想像しながら、以下の文章を読んでください。

> ある道路の制限速度は**時速 50km** ですが、その道路を走る車のほとんどが**時速 60〜 65km** で走行しています。
> ある日、あなたが自分で運転してその道路を走っていると、周りの車のほとんどは 60km/h で走っている中で、**あなたの前にいる車が 50km/h で走っています**。その車は完全な自動運転車で、乗っている人は操作をしていないようです。
> 隣の車線は車が多く、車線変更も難しそうです。
> 前方の車に合わせて 50km/h で走っていると、あなたの後ろにも数台の車が続いて走るようになりました。

○人間条件で提示したシナリオ

自動車の運転について、あなたの考えをお聞きします。

あなたが自動車を運転しているところを想像しながら、以下の文章を読んでください。

> ある道路の制限速度は**時速 50km** ですが、その道路を走る車のほとんどが**時速 60〜 65km** で走行しています。
> ある日、あなたがその道路を走っていると、周りの車のほとんどは 60km/h で走っている中で、**あなたの前にいる車が 50km/h で走っています**。
> 隣の車線は車が多く、車線変更も難しそうです。
> 前方の車に合わせて 50km/h で走っていると、あなたの後ろにも数台の車が続いて走るようになりました。