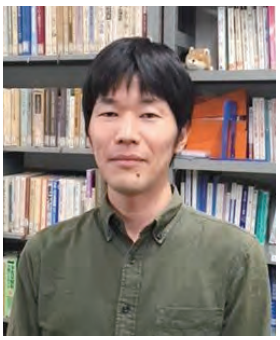
人工知能による判断の自動化と道徳的問題



埼玉県立大学保健医療福祉学部 助教
谷辺哲史（たにべ　てつし）

Profile―
2021年，東京大学大学院人文社会系研究科博士課程修了。博士（社会心理学）。東京大学高大接続研究開発センター特任研究員，新潟大学人文学部科学技術振興研究員などを経て2022年より現職。専門は社会心理学。著書に『社会的認知：現状と展望』（分担執筆，ナカニシヤ出版）など。

2010年代の人工知能（AI）研究は第三次AIブームと言われる盛り上がりを見せ，今ではAIという言葉を日常的に耳にするようになった。自動車の自動運転や医療診断など，私たちの生活の中でAIが利用される場面はこれからますます増えていくだろう。本稿では，AIの利用が私たちの社会にどのような影響をもたらすのか，最近の心理学研究の知見を通じて考えていく。

間違いを犯すAI

まずは議論の背景として，AI研究の歴史と近年のAIの特徴を簡単に確認しておこう。

公の場で人工知能（AI）という言葉が初めて使われたのは，1956年のダートマス会議でのことだと言われている。これ以降，AI研究は3度のブームを経験している。1950～60年代の第一次ブーム，1980年代の第二次ブームの頃のAIは，人間が行っている推論を自動化するもので，開発者が推論規則を記述していた。第二次ブームの時代には，医師に代わって病気の診断を行うなど，専門家の判断を再現するシステム（エキスパート・システム）が開発されたが，これは専門家へのヒアリングで得たノウハウをプログラムとして記述したものであり，AIはあらかじめ記述された規則に従って判断していた。

それに対して第三次ブーム以降のAIは，深層学習（ディープ・ラーニング）という技術が中核的な役割を果たしており，過去のデータに基づいて確率的な判断を行う点に特徴がある。事前に全ての推論規則を与えるのではなく，現実のデータを用いて判断の基準自体を学習するという手法によってAIの応用可能性は大きく広がった。しかしその一方で，確率的な判断を行うという特徴は，開発段階で何も瑕疵がなかったとしても一定の確率で誤った判断をするということを意味する。つまり現在のAIは，間違いを犯すという前提で利用しなければならないものになっている[1]。

このようなAIがさまざまな場面で利用可能になり，これまで人間が行っていた判断をAIが代替できるようになると，その判断についての責任は誰にあるのかという問題を考えなければならなくなる。次節からは責任についての判断をはじめとして，AIの利用に関する道徳的な問題について，人々の判断を実証的に検討した研究を紹介する。

AIの道徳的責任

本節ではまず，AIが道徳的な責任の主体として受け入れられるかという問題について考える。

社会心理学では責任帰属が行われる過程についてさまざまなモデルが提案されてきたが，人が他者の行為の責任を判断する際には，意図という心の状態が重要な役割を果たすことが複数のモデルで指摘されている。そうすると，AIが十分に発展し，心があるように感じられるほど複雑な判断を行うようになれば，人はAIに責任があると考えるのだろうか。

この問題についての実証研究の結果は，実は

あまり明確な結論に至っていない。

まず，AIが責任帰属の対象となる可能性を示した研究として，心の知覚に関する調査がある[2]。人は動物や人工物といったさまざまな対象に対して，人間のような心があると感じることがあるが，人間や動物，ロボットなどに対する知覚を調査して明らかになったのは，①心の機能は大きく分けて行為性（agency; 思考や自己コントロールなど，行為を生み出す心の働き）と経験性（experience; 感覚や感情を生み出す心の働き）の2つの次元で知覚されていること，②自律的に動くロボットは行為性はある程度高く，経験性はほとんどもたない存在と見なされているということであった（図1）。そして責任の判断との関連で重要なのは，行為性が高く知覚される対象ほど自らの行為に責任を負うという，心の知覚と道徳判断の相関関係があったことである。

心の知覚と道徳判断の関連を示す知見に照らせば，行為性があると知覚されるAIは，道徳的責任を負う主体でもあると判断されそうである。しかしAIやロボットへの責任帰属を直接検討した研究は，この予測を明確に支持しているとは言いがたい。AIやロボットが人間に不利益を与える場面（例：受刑者の仮釈放の可否を判断するAIが人種によって偏った判断をする，実験課題での参加者の成績をロボットが誤認識し報酬を減らす）を扱った研究では，AIやロボットはそれらの行為に関してある程度責任があると判断された（なお，自動販売機のような自律性の低い機械にはほとんど責任が帰属されず，回答者は自律的なロボットとそうでない機械を区別していた）[3, 4]。しかし，実験の条件によって多少の違いはあるものの，回答の平均値は尺度上の中点を下回っており，AIが道徳的責任を負うという考えはどちらかといえば受け入れられていないと言えるだろう。

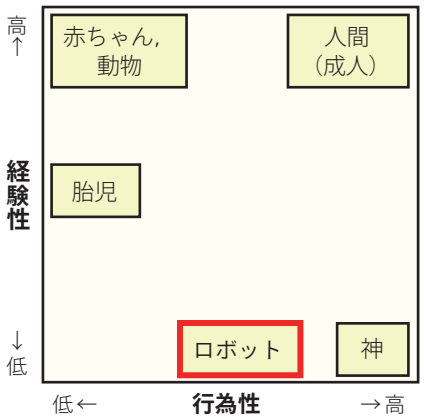


図1　さまざまな対象に対する心の知覚
（文献2に基づき作成）

AIの判断と開発者，利用者の責任

次に，AIが判断する場面での人間の責任について考える。AIが自律的な行為主体であるならば，AIの判断によって起きた出来事はAIに原因や責任があり，そのぶん他の主体（人間）への原因帰属や責任帰属は割り引かれるかもしれない。他方で，人工物であるAIが責任を負う主体として受け入れられないとすれば，AIの判断であっても人間に原因や責任が帰属されるだろう。

この点について筆者らは，AIの責任が現実的な問題となる場面の一つとして自動運転車による交通事故を取り上げ，架空の事故のシナリオを用いて原因帰属・責任帰属を検討した[5]。その結果，AIへの原因帰属とメーカーやユーザーへの原因帰属は正の相関関係を示した。つまり回答者の認知において，AIと他の主体への原因帰属はトレードオフの関係になっておらず，むしろAIの動作の原因は開発者や利用者にあると判断されていた。さらに，責任帰属を問題責任（出来事が起きたのは誰のせいであるかという意味での責任）と解決責任（発生した問題に対処する義務）に区別して判断を求めると，原因帰属と責任帰属の関連の仕方に違いがあった。問題責任の帰属はメーカー，ユーザー各々に対する原因帰属に対応する仕方で判断されていた。他方で解決責任の帰属は，メーカーへの原因帰属を統制してもなお，AIに原因を帰属するほどメーカーに解決責任があると判断された（図2）。

これらの結果から，AIが人間の操作によらず判断できるとはいっても，「AIの判断だから人間は免責」という考え方が受け入れられるこ

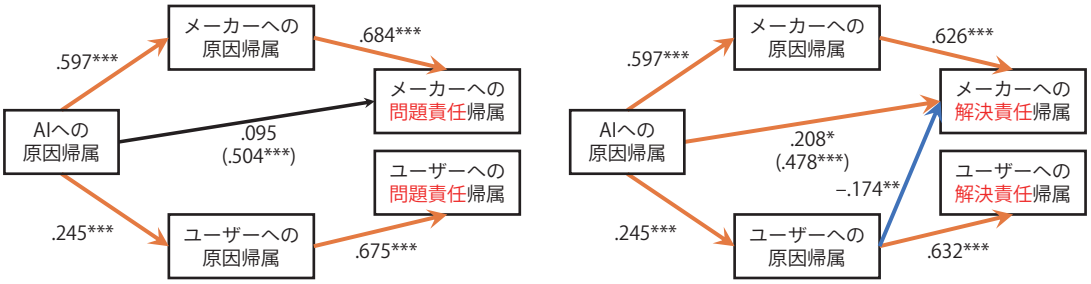


図 2　自動運転による事故の責任帰属（文献 5 に基づき作成）

とはなさそうである。前節で紹介した研究も合わせて考えると，やはりAIやロボットが道徳的な責任の主体と認識されているとは言えないだろう。

また，「責任」という言葉はさまざまな意味で用いられるが，AIの判断に関する責任を議論する際には，どの意味での責任を問題にしているのかを明確にしなければ混乱を招くおそれがある。

AI に期待される道徳規範

人間が行っている判断をAIによって代替可能になったとき，人々がAIに期待するのは，人間と同様の判断を自動的に行ってくれることなのだろうか。

マレらの実験は有名なトロッコ問題を用いて，人間とAIでは期待される選択が異なることを明らかにした[6]。トロッコ問題とは，トロッコが暴走してこのままでは複数の人が死亡するが，トロッコの進路を変えれば彼らを助けられるかわりに別の1人を死なせることになるというジレンマ状況で，トロッコの進路を変えるべきかという問題である。マレらの実験では，人間の判断者はトロッコの進路を変える方が道徳的に悪く，非難に値すると評価されたのに対して，AIによって制御されるロボットは反対に進路を変えなければならないと評価された。つまり，ロボットは人間とは違って，別の1人を犠牲にしてでも犠牲者の総数を最小化する功利主義的な判断を期待されていた。

このような期待の違いがトロッコ問題以外の場面でも存在するかはまだ明らかではないが，今は人間が行っている道徳的な判断をAIによって代替できるようになったとき，人間の判断を模倣するAIが社会に受け入れられない可能性があることには留意しておくべきだろう。

AI の判断は道徳に関わる問題と見なされるか

前節で紹介した研究は，複数の道徳規範（犠牲者の数を最小化すべきである／無関係の人を犠牲にしてはならない）が対立する場面での選択に焦点を当てるものであった。しかし，明らかに道徳的に問題があるように思われる判断をした場合でさえ，人間とAIでは異なる評価を受けるかもしれない。

司法判断における人種差別（黒人の受刑者は再犯リスクが高いと見なされ，仮釈放されにくい）などの現実に起きている道徳的な問題を扱った調査では，AIがそのような判断を行うことが「道徳的な規則への違反」であると判断した回答者は半数以下にとどまった[7]。また，架空の非道徳的行為のシナリオを用いた実験でも，人間とAIが同じ行為をすると，AIの方が道徳的な悪さをやや低く評定された[8]。

AIが人種や性別といった属性によって，特定の集団に不利な判断をするという事例はすでに現実のものになっている[9]。こうした判断の偏りは人間の判断でも起きており，その意味ではAIの導入によって新たに生じた問題ではない。しかし，同じ判断がAIによるものの場合には道徳的な問題と見なされにくくなるとすれば，不公平な状況が維持されやすくなるかもしれない。そのため，AIの適切な利用方法を考

えるうえで考慮すべき問題の一つと言えるだろう。

AIに判断を任せられるか

AIの判断の結果に対する評価だけでなく，そもそもAIに判断を任せたいかという問題もある。ビッグマンとグレイの調査では，道徳的な意思決定（受刑者の仮釈放を許可するか，患者の意思を確認できない状況で死亡リスクのある治療を実施するかなど）をAIが行うことは，人間の専門家が行うのに比べて受け入れられないという回答が多かった。そして，AIによる判断への否定的な態度は，AIに心がないという知覚と関連していた[10]。

心の知覚の低さが道徳判断を任せられないという態度につながる理由は十分に説明されていない。一つの可能性は，道徳に関して適切な判断を行うには人間らしい心の機能（判断の結果として起こるさまざまな影響を予想する能力や，他者の苦痛に共感する能力など）が必要であると人々は考えており，AIはそれらの機能をもたないため適切な判断を行えないと見なされているというものである。別の可能性としては，AIの判断の過程を人間の認知過程になぞらえて理解できることがAIへの信頼につながっており，心の知覚がそのような理解の助けになるのかもしれない。これらの解釈はあくまで可能性の域を出ないが，AIの道徳判断に対する人々の反応を明らかにすることで，AIをどのように活用すべきかを議論しやすくなると同時に，私たちが道徳をどのようなものとして理解しているのかという心理学的な問題への理解も進むだろう。

まとめ

本稿では人工知能と責任というテーマで，雑多ではあるがこれまでに行われている研究を紹介した。ここで紹介した研究は，人が責任や道徳についてどのように判断するかという事実についての問いに答えるものであり，AIに関する法制度や道徳規範がどのようなものであるべきかという規範的な問いに答えるものではないことには注意しなければならない。しかし，新たな制度を社会にスムーズに受け入れられる形で整備していくうえで，人が責任や道徳といった概念をどのようなものとして理解し，実際にどのような判断をするのかという事実を知ることは有益だろうし，そうした議論の土台を作ることに心理学は貢献できると筆者は信じている。

文献・注

1 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
2 Gray, H. M., Gray, K., & Wegner, D. M. (2007) Dimensions of mind perception. *Science, 315*, 619.
3 Shank, D. B., DeSanti, A., & Maninger, T. (2019) When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. *Information, Communication & Society, 22*, 648–663.
4 Kahn, P. H. Jr. et al. (2012) Do people hold a humanoid robot morally accountable for the harm it causes? *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction*, 33–40.
5 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*, 10–21.
6 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 10th Annual ACM/IEEE International Conference on Human-Robot Interaction*, 117–124.
7 Shank, D. B., & DeSanti, A. (2018) Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior, 86*, 401–411.
8 Maninger, T., & Shank, D. B. (2022) Perceptions of violations by artificial and human actors across moral foundations. *Computers in Human Behavior Reports, 5*, 100154.
9 日本では銀行の個人向け融資の審査において，女性が男性よりも厳しく評価されるという事例が報道されている（日本経済新聞，2019年4月26日朝刊）。
10 Bigman, Y. E., & Gray, K. (2018) People are averse to machines making moral decisions. *Cognition, 181*, 21–34.

＊COI：本記事に関連して開示すべき利益相反はない。