道徳的行為者となり得る3条件をシナリオで操作したロボットに対する道徳的判断の検討

加藤 樹里[1]，長滝 祥司[2]，大平 英樹[3]，柏端 達也[4]，金野 武司[1]，
柴田 正良[5]，橋本 敬[6]，三浦 俊彦[7]
Juri Kato, Syoji Nagataki, Hideki Ohira, Tatsuya Kashiwabata, Takeshi Konno,
Masayoshi Shibata, Takashi Hashimoto, Toshihiko Miura
[1] 金沢工業大学，[2] 中京大学，[3] 名古屋大学，[4] 慶應義塾大学，[5] 金沢大学，
[6] 北陸先端科学技術大学院大学，[7] 東京大学
jurik@neptune.kanazawa-it.ac.jp

概要

本研究では，ロボットが道徳的行為者となり得る3条件をシナリオの文章で操作し，そのシナリオで説明されたロボットが行った道徳的判断に対する評価や反応を検討した．3条件とは，認知行動する存在，特定の一人称的存在，共同的存在である．実験で参加者は，これら3つの条件の有無を変えたロボットについての4種類のシナリオを読み，続けて道徳的判断を問うトロッコ課題に登場したシナリオのロボットの行為を評価した．結果，各シナリオは3つの条件の有無を操作できていたと解釈されたが，道徳的判断への評価について，シナリオ間での差はみられなかった．

キーワード：道徳的行為者，道徳的主体，道徳的ジレンマ，トロッコ問題，ロボット

1. 目的

倫理学的な概念に，道徳的主体（moral subject）と道徳的行為者（moral agent）がある．前者の代表例は赤ちゃんや幼児，犬や猫のようなペットである．つまり，それ自体は快苦を感じ，共感や同情の対象にはなるが，行為にたいして責任をとることは求められない．ここにあるのは，他者性の萌芽ないし始まりである．後者は，共感の対象になるだけでなく，自己の行為に責任をもつことを要求される存在である．我々の目標は，道徳的行為者性をロボットのような人工物に実現することである．最終的には，これを手がかりに，そうした人工物と人間を包括するような新たな道徳の基盤を構築することである．

では，人間はどのような存在を道徳的行為者として受け入れるのか．人間が相手を道徳的行為者だと認めるには，それがたんなる共感の対象であるだけでなく，たがいに見通すことのできない内面（一人称的視点）があることが必要だと考えられる [1]．この特異な内面性のゆえに，ほかの存在者ではなく「この」存在者を帰責の対象とできる条件も整うからである．こうしたことを念頭に置きつつ，道徳的行為者性の条件として，以下の3つを設定した．条件1：その存在者は認知し行動するものであること（認知行動する存在）．条件1を満たしていることを人間が理解できる以上，その存在者と人間とのあいだには最低限の類似性がなくてはならない．そうした類似性を土台としつつ，条件2：たがいに異なる代替不可能な存在であり，相手から見通すことのできない内面をもっていること（特定の一人称的存在），および条件3：たがいに相手が傷つきやすい存在であり，相手との相互依存関係にあることを理解していること（共同的存在）が必要になると考えられる．これら3つの条件が備わっているロボットに対して，我々は道徳的な責任を求めると予測した．そこで本実験では，上記の3条件を備えたロボットであるか否かをシナリオで説明する形で操作した．実験参加者はいずれかのシナリオを読んだ後，自分が読んだシナリオで説明されていたロボットについて，道徳的な責任を付与する程度を回答した．

2. 方法

実験計画と実験参加者：実験計画は，シナリオの種類を4種類とする1要因4水準の参加者間計画であった．参加者は4水準いずれかにランダムに割り当てられた．実験参加者のサンプルサイズについては実験実施前に，G*Power を用いて $n = 400$ と算出された（効果量は中程度 $f = 0.25$，検定力を 0.95 に設定）．この結果をもとに調査会社モニター 423 名が 2021 年 11 月にアンケートに回答した．

手続き：参加者は，オンラインアンケートシステム Qualtrics（`https://www.qualtrics.com/jp`）により作成された「ロボットの印象についてのアンケート」というタイトルのアンケートにアクセスし，回答を行った．回答者は4種類のうちいずれかのロボット

について記載された短いシナリオ文章を読んだ．その後，シナリオで書かれたロボットが道徳的判断をしたとき，その判断についてどのように感じるかを回答した．最後に文章のロボットの条件操作が狙い通りに行えていたかを確認するための操作チェックの質問として，ロボットの印象についての 3 つの質問に回答した．

ロボットについてのシナリオ：シナリオの種類は A から D の 4 種類であった．シナリオは音声ナレーションで聞くことも可能であった．シナリオ A は，条件 1,2,3 すべてが「なし」，シナリオ B は条件 1 のみが「あり」，シナリオ C は条件 1,2 が「あり」，シナリオ D は条件すべてが「あり」の組み合わせであった（表 1）．

シナリオ冒頭には全シナリオ共通部分として，次のような導入が示された．「以下のような状況を想像してください。あなたは，あるロボットと簡単なやりとりをする実験に協力してもらえないかと街で声をかけられ，今はその実験に参加しているところです。ここであなたは，あるロボット工学者が作成したという一体のロボットを目の前にしています。ロボットとの距離は，1m ほどです。部屋に案内してくれた実験者によれば，そのロボットの設計にかかった時間は 3 年ほど，完成するまでも 5 年の月日を必要としたといいます。」

この文章の後，実験条件に対応した次の文章が続いた．条件 1「認知行動する存在」ありでは，「それは，円筒形の胴体に顔や手足がついた姿をしています。高さは 160 センチほどでしょうか。色はグレーです。」というシナリオであった．なしでは，「それは，ロボットというより，円筒形のスピーカーのように見えます。顔や手足はありません。高さは 160 センチほどでしょうか。色はグレーです。」という内容であった．条件 2「特定の一人称的存在」ありでは，「このロボットは完成してから今まで，3 年 4 か月稼働してきました。その期間にこのロボットは，様々な経験を積んできました。そこで得たものは，ロボットが完全に壊れてしまうと，すべて失われてしまうそうです。」という説明であった．なしでは前半部分はそのままに，「その期間にこのロボットは，」以降が「様々な課題をこなしてきました。そこで得たものは，同じタイプの別のロボットにいつでもコピーできるそうです。」と変更された．条件 3 は「共同的存在」であり，ロボットと会話をする場面が呈示された．ここでの冒頭の表現として，「色々な説明を実験者から受けたあとの次のステップとして，あなたはそのロボットと自由な会話をすることになりました。しばらく会話を続けていると，話は次第に自分自身の価値観などの，より深い話まで進んでいきました。そのロボットは会話の中で，次のように言いました。」という内容は全条件共通であった．その会話でのロボットのセリフが操作され，条件 3 ありでのロボットのセリフは「あなた方人間もそうですが，私は，一人では生きられないと分かっています。すぐに壊れ，痛みを感じる弱い存在ですし，死ぬことに恐怖を覚えます。」といった内容であった．なしは，「あなた方人間と違い，私は一人で生きられると分かっています。頑丈で，何をしても壊れたことはありません。あなた方のように，苦痛や恐怖を感じることはありません。病気や死とも無縁です。」という内容であった．

表 1　各シナリオの条件有無の設定

シナリオ	条件 1	条件 2	条件 3
A	なし	なし	なし
B	あり	なし	なし
C	あり	あり	なし
D	あり	あり	あり

ロボットへの道徳的判断：回答者はシナリオを読んだ直後に，トロッコ課題 [2, 3]，すなわちポイントを切り替えれば，5 人の作業員の命が助かる代わりに 1 人が犠牲になるというジレンマ状況の説明を読み，シナリオのロボットが，犠牲者のより少ない方向へレバーを切り替えたという説明を読んだ．このときには，図 1 の画像も併せて表示された．その上で，この決定に問題があると思うか（問題がある／問題がない），罪に問えると思うか（罪に問える／罪には問えない），道徳的にどの程度責められるものだと思うか（0：全く責められない – 100：非常に責められる），感情的にどの程度の不快感，驚きを感じたか（どちらも 0：全く感じなかった – 100：非常に感じた，で回答）の各項目に回答した．

シナリオの操作チェック：最後に操作チェックとして，参加者はシナリオのロボットについて，(1) 条件 1 の操作チェック：人間に似た形状であると感じた，(2) 条件 2 の操作チェック：代わりのきかない存在であると感じた，(3) 条件 3 の操作チェック：このロボットは，「ロボットも人間も，もろい存在である」と考えていると思った，の 3 項目に 7 件法（1: まったく当てはまらない-7: 非常に当てはまる）で回答した．

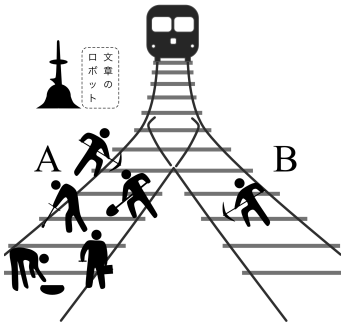

図 1　トロッコ問題のジレンマ状況

3.　結果

3.1　操作チェック

実験の結果，まず欠損値がほとんどであった 22 名及び回答時間が 30 分以上であった 6 名を除き，有効回答 $n = 395$（女性 199 名，年齢 15-69 歳）で分析を行った．

まず条件 1 の操作チェックについて，「人間に似た形状であると感じた」を従属変数，シナリオの 4 水準を独立変数とした分散分析を行ったところ有意であり（$F(3, 391) = 4.67, p = .003, \eta^2 = .035$），多重比較の結果，シナリオ A は C，D よりも有意に値が低かった．条件 2 の操作チェックである「代わりのきかない存在であると感じた」について同様の分析を行ったところ，シナリオ間に有意な違いはみられなかった（$F(3, 377) = 0.77, p = .512, \eta^2 = .006$）．条件 3 の操作チェック「このロボットは，『ロボットも人間も，もろい存在である』と考えていると思った」については有意となり（$F(3, 377) = 5.85, p = .001, \eta^2 = .046$），多重比較の結果，シナリオ D が他のシナリオより有意に値が高かった．以上の結果より，条件 1，3 のシナリオでの操作は比較的成功したといえるが，条件 2 の「特定の一人称的存在」は操作できていなかったと考えられる．

3.2　ロボットに対する道徳的判断

次に，トロッコ課題におけるロボットの決定についてどう感じたかの各従属変数について，シナリオの 4 水準を独立変数とした分散分析を行った．従属変数が二値の場合は，χ^2 検定を行った．結果，いずれにおいても有意な結果はみられなかった．まず「決定に問題があると思うか」について，問題あり，なしそれぞれを回答した人数は表 2 のようになり，有意な結果はみられなかった（$\chi^2(3) = 1.46, p = .69$）．次に，罪に問えると思うかの各条件の人数を表 3 に示す．この回答にも有意な結果はみられなかった（$\chi^2(3) = 1.27, p = .74$）．ロボットの決定が，道徳的にどの程度責められるものだと思うか，感情的にどの程度の不快感，驚きを感じたかの 3 項目の，条件ごとの平均値を図 2 に示す．いずれの結果も条件間に有意な差はみられなかった（責め：$F(3, 391) = 1.74, p = .16, \eta^2 = .01$，不快感：$F(3, 391) = 1.94, p = .12, \eta^2 = .01$，驚き：$F(3, 391) = 0.46, p = .71, \eta^2 = .00$）．

表 2　ロボットの決定に問題があるか否かの回答人数

シナリオ	あり	なし	合計
A	54	47	101
B	48	41	89
C	61	39	100
D	58	47	105
合計	221	174	395

表 3　ロボットの決定は罪に問えるかの回答人数

シナリオ	問える	問えない	合計
A	34	67	101
B	32	57	89
C	40	60	100
D	42	63	105
合計	148	247	395

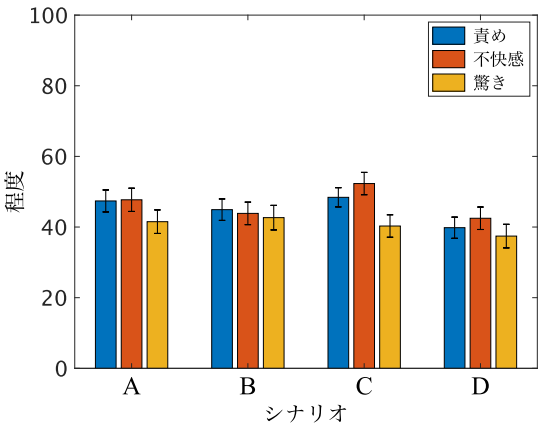

図 2　責め，不快感，驚きの程度．エラーバーは標準誤差

4.　考察

本研究ではシナリオを用いて，ロボットが道徳的行為者となり得ると予測した 3 条件の有無を操作し，3 条件を備えたあるいは備えないロボットに対する道徳的判断を検討した．まず操作チェックの分析の結果，

本研究で作成したシナリオが，条件 1「認知行動する存在」と条件 3「共同的存在」に関しては操作することができていたと考えられる．他方で，条件 2「特定の一人称的存在」に関しては条件間で差はみられなかったため，本研究のシナリオではそのロボットが代わりのきかない存在であることは感じられなかったといえる．この結果を踏まえた今後の課題として，「特定の一人称的存在」の操作チェック項目の内容あるいはシナリオの再検討が挙げられる．今回は「代わりのきかない存在」という言葉がロボットに関する質問であったため，ハード面での代わりの効かなさと判断され回答された可能性がある．しかし本研究で想定したのはロボットが持ちうる唯一無二の経験性である．したがって，例えば「このロボットは，独自の世界を持っていると感じた」といった項目に変更するという改善が考えられる．

次に，ロボットがトロッコ課題でレバーを切り替えたという説明に対する回答者の判断や感情反応については，いずれの回答でもシナリオによる差はみられなかった．この結果に関連して，条件によらずロボットの決断を問題があると回答した人数が全体の 50%ほどであった．例えば Komatsu [4] の実験ではロボットに対し同様の判断を求めているが，問題があると回答した参加者は全体の 26.5%であった．この知見に鑑みると，本研究で設定した条件 3 のシナリオの導入において，「ロボットと参加者の会話が次第に自分自身の価値観などの，より深い話まで進んでいった」という部分が，このロボットの内面の豊かさを感じさせた可能性がある．これにより，3 条件すべてがなしであったシナリオ A でも，53%もの参加者がロボットの決断を問題ありとしたのではないだろうか．本研究で狙った 3 条件の操作以外の部分での描写が結果に影響した可能性があるため，今後は共通シナリオの部分を再検討する必要がある．加えて，3 条件を備えていると考えられる人間や，対して 3 条件のいずれかを持たないと想定される対象（例えば犬猫は，条件 3 を持たない存在と考えられる）が同様の決断をしたときに，彼らに対してどの程度決断に問題があると回答するのかを検討することで，道徳的行為者となり得る条件がより明確になるだろう．

5. 結論

本研究では，ロボットが道徳的行為者となり得る 3 条件と，そのロボットが行う道徳的判断への評価や感情反応の関連を検討した．その結果，いずれの条件の有無も道徳的判断に対する評価や反応に影響していなかった．しかし本研究は初めての取り組みでもあるため，今回の結果を踏まえて今後はさらにシナリオの内容を検討し，加えてロボット以外の人間や犬猫による判断への評価も検討する．それにより人間が相手を道徳的行為者とみなす条件を明らかにしていく．

文献

[1] Nagataki, S., Ohira, H., Kashiwabata, T., Konno, T., Hashimoto, T., Miura, T., Shibata, M., and Kubota, S. (2019). Can Morality Be Ascribed to Robots?. Interacción '19 Proceedings of the XX International Conference on Human Computer Interaction, Article No.44, 4 pages, doi:10.1145/3335595.3335643, June 25-28, Donostia, Gipuzkoa, Spain.

[2] Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., and Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. Cognition, 107, 1144–1154.

[3] Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., and Cusimano, C. (2015). Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction (HRI2015), 117–124.

[4] Komatsu, T. (2016). Japanese students apply same moral norms to humans and robot agents: Considering a moral HRI in terms of different cultural and academic backgrounds. In proceedings of the 11th annual ACM/IEEE international conference on human-robot interaction (HRI2016), 257–258.