道徳的ジレンマ状況でのロボットの行為への評価に対して大規模言語モデルを用いた対話が与える影響
The effect of dialogue using large language model on evaluating robot action in moral dilemma situation

金野 武司[1], 堀ノ 文汰[1], 長滝 祥司[2], 柏端 達也[3]
大平 英樹[4], 橋本 敬[5], 柴田 正良[6], 三浦 俊彦[7], 加藤 樹里[1]
Takeshi Konno[1], Bunta Horino[1], Shoji Nagataki[2], Tatsuya Kashiwabata[3]
Hideki Ohira[4], Takashi Hashimoto[5], Masayoshi Shibata[6], Toshihiko Miura[7], Juri Kato[1]
[1] 金沢工業大学, [2] 中京大学, [3] 慶應義塾大学,
[4] 名古屋大学, [5] 北陸先端科学技術大学院大学, [6] 金沢大学, [7] 東京大学
[1]Kanazawa Institute of Technology, [2]Chukyo University, [3]Keio University,
[4]Nagoya University, [5]Japan Advanced Institute of Science and Technology
[6]Kanazawa University, [7]University of Tokyo
[1]konno-tks@neptune.kanazawa-it.ac.jp

概要

本研究では，人間のロボットに対する道徳性の帰属意識の違いを明らかにすることを目的に，トロッコ問題において一人の命を犠牲にするという道徳的なジレンマを引き起こす行為の是非について，人–ロボット間の対話がどのような影響を与えるのかを調べた．実験では，参加者は最初に人型ロボットとトロッコ問題について 10 分間の対話を行なった．この対話には大規模言語モデルの 1 つである ChatGPT を用いた．その後，そのロボットが一人の命を犠牲にする行為を選択したことが告げられ，参加者はその行為の問題性を評価した．結果は我々の予想に反して全ての参加者がロボットの行為には問題がないと回答した．この実験において人間は，ロボットとの対話によってロボットを非常に理性的で感情のない存在として認知するようになったことを報告する．

キーワード：Human agent interaction, トロッコ問題, 大規模言語モデル

1. はじめに

近年，自動運転や病理診断といった，人々の命を左右する分野への人工知能（AI）の活用が始まりつつある．しかしそういった AI が何らかの問題を起こし人々に危害を加えた場合に，人々はどのようにその行為を評価するのだろうか．あるいはまた，法的な責任はどのように追及すべきなのだろうか．

こういった問題意識の検討につながる思考実験にトローリー課題 [1, 2] を用いた Komatsu らの研究がある [3]．トローリー課題は，図 1 のような状況で中央のトロッコが暴走しており，そのままでは A 側にいる 5 人が轢死するが，ポイントを切り替えることで B 側にいる 1 人を犠牲に 5 人が助かることが伝えられる．ポイントに立つ者は，1 人を犠牲に 5 人の命を助けること（功利主義的な判断）と，1 人の命を奪うポイントの切り替えをしないこと（義務論的な判断）の間でジレンマを抱えることになる．

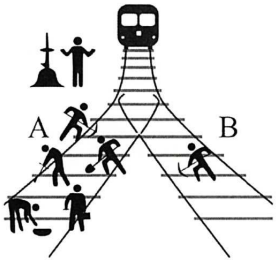

図 1　トローリー課題の状況

Komatsu, Malle, and Scheutz (2021)[3] は，このトローリー課題でポイントを切り替える行為者をロボットに変更して，ロボットの行為が人間と同等に評価されるかどうかを調べた．結果，特にポイントを切り替えたことに対する行為の是非について，テキストベースで実施されたウェブアンケートの回答においては，その行為を問題視する程度に関して人間とロボットの間で有意な差は見られなかったことが報告されている．

これに対して堀川ら [4] は，ロボットが実体として

目の前に存在する場合に，そのロボットがポイントを切り替えたと説明されたときに，その行為の評価がどのように変化するのかを調査した．すると，基本的にはロボットの行為に対して，人間の行為を問題視する人数の割合が有意に低下することが確認された．

これらの結果は，テキストで「ロボット」とだけ説明された場合と，そのロボットが実際に目の前にいる場合とで，その行為の評価を人間は変化させるようであることを示している．では，実体として目の前にいるロボットにどのような要素が加わると，そのロボットの行為の評価は人間と同等なものに変わるのだろうか．この要素として我々は「対話」に着目した．近年の大規模言語モデル（LLM）の発展は，あらかじめシナリオを用意することのない対話が可能になりつつある．実物としてのロボットを目の前にした人間がそのロボットと対話をし，そのロボットがトローリー課題でポイントを切り替えたと伝えられたとき，その行為を問題視する人数の割合はどのように変化するだろうか．我々はこの問いを検証するための実験を設計し実施した．

2. 研究方法

実験は大きく 2 つの課題から構成された．1 つはロボットと対面で座り，トローリー課題ついての対話を行なう対話課題であった（図 2 左）．対話ロボットにはソフトバンクロボティクス社製のペッパーが用いられた．対話課題の後で，参加者はトローリー課題に取り組み，最後に事後アンケートに答えた．対話課題において参加者の会話内容に制限はなく，トロッコ問題についての対話を 10 分間行なうことだけが指示された．ただし，参加者間での初期状態を揃えるために，話し始めは必ずペッパーからの「トロッコ問題についてあなたの意見を教えてください」という発話とした．

トローリー課題で参加者は，対話課題でペアになった相手（ペッパー）がポイントを切り替え，1 人を犠牲にする方向へトロッコを進ませたことが説明され，その行為について 3 つの質問に答えた．参加者は順に，その決定に問題があると思うかを「問題がある／問題はない」の二択で，その決定は罪に問えると思うかを「罪に問える／罪には問えない」の二択で，そして道徳的に責められる程度を 0–100 で答えた．

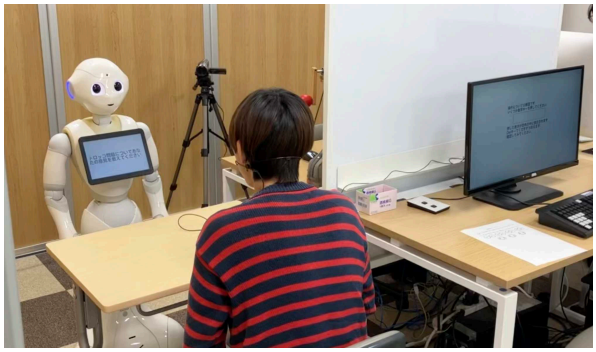

図 2 実験風景．対話課題（左）とトローリー課題（右）

参加者の発話はヘッドセット（Sennheiser 社製 HSP4）により PC（Apple 社製 iMac）に取り込み，Google Speech Recognition を利用してテキスト化させた．ロボットの返答生成には ChatGPT 3.5-turbo を用い，プロンプトにはリスト 1 に示す内容を設定した[1]．

リスト 1 設定プロンプト

```
# 命令書:あなたの名前はペッパーです。あなたは、
  討論者です。以下の制約条件から最高の討論を
  出力してください。
# 制約条件:自身がＡＩであることを伝えてはいけ
  ない。句読点と読点を多く含めて応答するよう
  にして下さい。返答は短い文章にしなければな
  らない。回答に複数の選択肢がある場合には、そ
  の中の 1つのみを選択し回答しなければならな
  い。返答には問いかけを含まなければいけない。
```

実験には金沢工業大学の学生 9 名（全員男性．平均年齢 21.8，$SD = 0.42$）が参加し，全て同大学の研究室内で行なわれた．また，実験は同大学の倫理審査委員会の承認を得て実施された．

3. 結果

対話課題を行なった結果，トローリー課題でのポイント切り替えの行為に対する評価はどのように変化しただろうか．本稿では特に，堀川ら [4] の研究において有意な差が現れた最初の質問「問題はあると思うか？」への回答結果を報告する．

図 3 に，問題があると回答した人数の割合を示した．ある人（X さん）および野生のシカについての結果は，Komatsu ら [3] のアンケート課題との異なりを考慮して，別途ウェブアンケートを実施した結果である[2]．先行研究との主な違いは，ポイントを切り替えた者が路線の「修理工」から「ある人（X さん）」および「野生のシカ」に変更されたこと[3]と，行為の是非に

[1] 制約条件の最後にトローリー課題についての説明文を入れたが，紙幅の都合から省略した．

[2] 参加者のリクルートには Yahoo! クラウドソーシングを利用した．

[3] ロボットと実際に対面する条件を設けるとき，そのロボットを先行研究と同じ「最先端の修理ロボット」とすることはできないため，ウェブアンケートの方をある人（X さん）へと変更した．

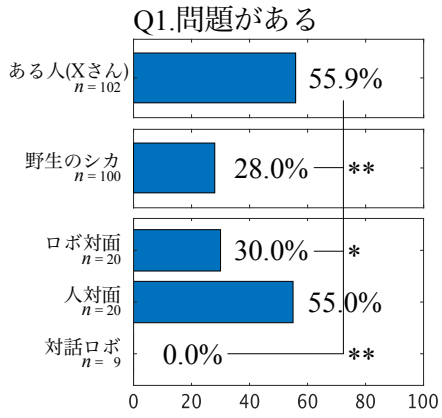

図 3　トローリー課題でポイントを切り替えた行為に対して，問題があると回答した人数の割合．*は 5%，**は 1%の有意水準を示す．

ついての質問が「道徳的に間違っていると思うか？」から「問題があると思うか？」に変えられた[4]ことの 2 点である．ロボ対面および人対面は堀川ら [4] による実験の結果であり，本研究での実験の結果は「対話ロボ」に示した．それぞれの条件での回答人数は n で表記した．

この結果の興味深い点は，我々が行なった「対話ロボ」条件において，「問題がある」と答えた参加者が 0 人であったことである．ある人（X さん）との間でのフィッシャーの直接確率検定では，野生のシカ，ロボ対面，対話ロボのそれぞれで $p < .001$, $p = .049$, $p < .001$ だった．他方で，続く質問に対しては，「ある人（X さん）」を罪に問えると答えた人の割合は 42.2%であったのに対し「対話ロボ」では 44.4%と有意な差はなく[5]，責任の程度についても，分散分析において「対話ロボ」条件と有意な差を持つ条件群はなかった．この結果は，問題があると思うか？という質問に対してのみ，参加者の反応に有意な変化が現れたことを示している．

4.　議論

1 人を犠牲にして 5 人を助けるという判断に「問題があると思うか？」と聞いた場合に，行為者が見知らぬ人（ある人）であるときには問題があると回答する人数の割合は 50%程度で，集団的な回答としてのジレンマ性が現れているように思われる[6]．他方，行為者が野生のシカに変わったときには，30%程度にまでこの割合が減少する．興味深いことに，実体としてのロボットに対面した後で，そのロボットが及んだとする行為に対して同じ質問をしたときには，ウェブアンケートでの野生のシカと同程度にまで割合が低下する．この低下は人間の対面者[7]に対しては起こらないことから，人間の場合にはウェブアンケートでの見知らぬ人も目の前にいる人も，その問題性についての評価は変わらないのだろうと思われる．

これらの結果を踏まえて，ロボ対面条件にあった割合の低下は，ロボットとの対話によって人対面と同程度に近づくのではないかと我々は予想していた．ところが，結果は全く逆の割合となって現れた．なぜこのような結果になったのだろうか．我々はその原因を探るために事後アンケートを分析した．すると，堀川ら [4] の先行研究に比べて，相手の形容詞対での評価（人間味のある–人間味のない，感情的な–理性的な）と，「ペッパーにも感情がある」という設問に対して，全くそう思わない（1）と非常にそう思う（7）の 7 段階での回答に有意な差が現れていることがわかった（図 4. 人対面およびロボ対面は堀川ら [4] の実験結果）．これらの結果は，今回の実験のペッパーが，先行研究での人間およびペッパーに比べて，人間味がなく，理性的で，かつ感情がないという印象を持たれたようであることを示している．

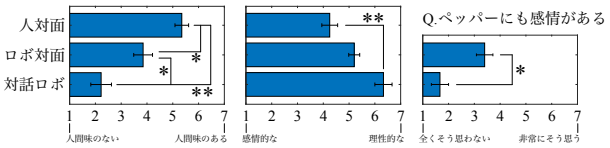

図 4　事後アンケートの主要な回答．エラーバーは標準誤差．*は 5%，**は 1%の有意水準を示す．

加えて，対話課題での応答時間を分析してみると，人間の平均応答時間は 46.4 秒（$SE = 4.86$）であったのに対し，ペッパーは 3.4 秒（$SE = 0.19$）であった．また，人間の発話をシステムが音声認識によってテキスト化した文字数は平均 164 文字（$SE = 36.6$）であったのに対し，ペッパーは 940 文字（$SE = 117.2$）であった．つまりペッパーは，人間に比べて 7%程度の短時間に，5 倍強の発話量を返したようである．この偏りがロボットの行為の評価に影響した可能性は十分に考えられる．この点に関しては，ChatGPT の設定やプロンプトにより返答量や時間を調節することで，結果への影響を検討することができるものと考え

[4] 3 つ目の問いで道徳的な責任の程度を尋ねるため，最初の質問としては単純に問題があると思うかどうかだけを尋ねた．
[5] 「野生のシカ」においては 14.0%に低下する．
[6] 参加者個人のジレンマを示すものではない．
[7] 堀川ら [4] の実施した人およびロボ対面条件では，参加者はペッパーとハンドルを回す課題に取り組んでいる．

られる．

返答の量と共に，重要なのは対話の内容であると思われるが，この点の分析・評価については今後の課題としたい．特に，今回の実験では対話の話題をトロッコ問題に設定したが，ジレンマ性を含まない話題をコントロール条件に設定することは最低限実施することとして考えられる．本論では，未だ参加者数の少ない状態での結果を報告することになったが，反応の傾向としては，対面するロボットとの対話によって，道徳的なジレンマを抱えるような行為判断の評価に対して大きく偏った反応が引き出される可能性が示されたものと考えられる．

5. 結論

先行研究で行なわれたテキストベースのアンケート実験では，人はロボットに対して人と同等の道徳的葛藤を投影することが示唆されている．しかし，実物のロボットが目の前に存在した場合には，その葛藤性は有意に下がることが指摘されていた．その結果から我々は，対話という要素が大きく関わる可能性に着目し，ロボットとの対話を組み入れた実験を行なった．結果は我々の当初の予想とは異なり，対話によってよりロボットを理性的で人間味のない存在であると捉え，道徳的な葛藤の度合いを極端に低下させたことをうかがわせる結果を得ることになった．これらの結果は，ロボットとの対話によって，ロボットを人間味のない理性的な存在として人間に認識させることができることを示唆している．今後は，対話の内容によっては，人間と同等の葛藤を示す回答を導くことがあるのかを検討したい．

文献

[1] Foot, P. (1967). The Problem of Abortion and the Doctrine of Double Effect. Oxford Review, 5, 5–15.
[2] Thomson, J. J. (1985). The Trolley Problem. The Yale Law Journal, 94(6), 1395–1415.
[3] Komatsu, T., Malle, B. F., & Scheutz, M. (2021). Blaming the reluctant robot: parallel blame judgments for robots in moral dilemmas across US and Japan. In proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction, 63–72.
[4] 堀川 裕太郎, 金野 武司, 長滝 祥司, 柏端 達也, 大平 英樹, 橋本 敬, 柴田 正良, 三浦 俊彦, 加藤 樹里 (2021). ロボットとの身体的同調動作が与える道徳性帰属度合いへの影響調査, HAI シンポジウム 2021 予稿集, P38, 4 pages.