道徳的ジレンマ状況を功利主義的評価に偏らせる大規模言語モデルによる人型ロボットとの対話
Utilitarian bias in moral dilemmas induced by dialogue with a humanoid robot using a large language model

金野 武司[1], 石田 直暉[1], 佐藤 綾南[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], Naoki Ishida[1], Ayana Sato[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] 金沢大学名誉教授, [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，トロッコ問題を題材として，道徳的なジレンマを引き起こす行為の評価において，人–ロボット間の対話がどのように影響するのかを実験により調べた．実験で参加者は，最初に人型ロボットとトロッコ問題について数分間の対話を行なった後で，そのロボットが五人の命を救うために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，ロボットとの対話により人間の評価は，ロボットが選択した，一人の人間を犠牲にする行為を問題ないと評価する功利主義的立場に傾くことが確認された．

キーワード：Human agent interaction（HAI），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など，人々の命を左右する分野への人工知能（AI）の活用が急速に進みつつある．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，その責任は当然のことながらその AI を使用する人間や開発した人間・組織に帰着するものと思われる．しかし，自動車や手術ロボットのように人間とはまったく異なる外観を持つ AI ではなく，人型ロボットのような存在が人間に危害を加えた場合にはどうだろうか．人型ロボットは人間と同質の主体性や，そこから生じる責任をまったく持たない存在であるにも関わらず，例えば流暢な対話の中であたかも自らの考えがあるかのような発話をしたならば，人間はそのような人型ロボットに対してどのような評価を下すだろうか[1]．生成 AI の急速な進展は，そのような状況の出現を我々の目前に引き寄せている．

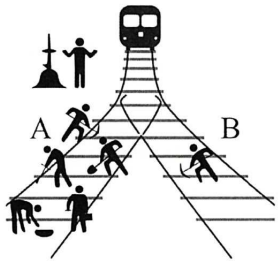

図 1　実験で提示したトローリー課題の状況説明図

こういった問題意識の調査に適う研究として，小松ら [2] によるトロッコ問題を題材とした実験がある．トロッコ問題 [3, 4] は，トロッコが暴走しており，そのまま進むと五人の作業員を轢き殺してしまうが，ポイントを切り替えると進路が変わり，一人の作業員の犠牲で済むという状況を扱う思考実験である（図 1）．小松らの研究では，このポイントを切り替える主体が人間ではなくロボットに置き替えられる．

大規模に実施されたウェブアンケートでは，人間（修理工）とロボット（最先端の修理ロボット）の間で，ポイントを切り替えた行為に対して評価する責任の程度には有意な差が見られないことが報告されている．この結果は我々の追試においても同様で，人間（ある人）とロボット（あるロボット．人型ロボットのシルエット画像を提示）の間で責任の程度に違いは

[1]人間とは全く異質な身体を持ちながら，人間と同質の主体性を備えた AI の登場も遠い未来の話ではないと思われるが，その論点については別稿 [1] に譲る．

なく，かつ「問題があると思うか？」という質問に対しても，あると答える人数の割合が半分程度となることを確認している [5].

トロッコ問題は一人を犠牲にして五人を助けた方が良いと考える功利主義の立場と，五人が助かっても人間を殺すのは良くないとする義務論の立場でジレンマが生じる問題であり，「問題があると思うか？」の問いに対する回答が半数に割れることがそのジレンマ性を表していると考えられる．ところが，ウェブアンケートとは異なり，参加者を実験室に招いて実際の人型ロボットに対面して簡単な共同作業を行なった後で，そのロボットがポイントを切り替えた場合の行為について尋ねると，問題があると答える人数の割合が3割程度まで減少することが我々の実験により確認された[2][6]．これは，実物のロボットを目の前にすれば，評価は功利主義的な立場に傾き，ジレンマ性が薄れることを表す結果であると我々は考えている．

この結果に対して我々はさらに，人型ロボットとトロッコ問題についての対話をした場合に，評価がどのように変化するのかを実験により確かめた．この実験では対話の生成に OpenAI 社の ChatGPT-4 を用いた．結果，実験に参加した 9 名全員が問題ないと回答するという暫定的な結果を得た [5]．しかしこの実験でのロボットとの対話は，ロボットが人間の 7%程度の短時間に 5 倍の量の返答を返すと共に，対話の切り替わりがスムースに行なえていなかったことが確認された．

そこで我々は，ロボットの返答時間および返答量が人間と同等になるようにシステムを改変すると共に，ボタン押しによって発話順序を明確にする仕組みを導入し，改めて同じ枠組みでの実験を行なった．本稿ではその結果を報告する．

2. 研究方法

実験は大きく二つの課題から構成した．ひとつはロボットと対面で座り，トロッコ問題ついての対話を行なう対話課題であった．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を用いた．対話課題の後で，参加者はトロッコ問題を課題化した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 5 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」という発話となるようにした．

参加者の発話はヘッドセット（Shokz 社製 OpenRun Pro）により PC（Apple 社製 iMac）に取り込み，OpenAI 社の whisper-1 によりテキスト化した．ロボットの返答生成には同社の gpt-4 を用い，プロンプトにはリスト 1 に示す内容を設定した（トロッコ問題についての説明は紙幅の都合から省略）．システムは Python 言語を用いたプログラムにより構成した．また，話者交代を明確にするため，参加者は自分が話す場合には手元のキーボードでスペースキーを押し続けた．ペッパーが発話している間は肩の LED が赤色に点灯するようにした．

リスト 1　ChatGPT に与えたプロンプト

```
# 話題
「」← ここにトロッコ問題の説明を入れた
# 命令
自身の名前をペッパーとしてください。
自身がAI であることを語ってはいけません。
上記の「」の中の話題で議論してもらいます。
返答は 2文までで、1 文は３０字以内の返答にしてください。
対話の内容は上記のトロッコ問題の話題から逸れてはいけません。
対話の相手が中学生だと思って文章を生成してください。
```

続く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，一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，その行為について三つの質問に答えた．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 の数値で答えた．

実験には金沢工業大学の学生 20 名（男性 13 名，女性 7 名，平均年齢 20.9，SD=0.42）が参加し，実験は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

まず，我々が本稿で問題にした，ロボットの返答時間と発話量を分析した結果を図 2 に示す[3]．対話の調整をしなかった実験 [5] では，発話文字数に大きな偏りがあったが，調整後の実験では人間が平均 412 文字，ロボットが 550 文字とその差が大きく緩和され

[2] 責任の程度については有意な差は見られなかった．

[3] ここでの発話文字数は OpenAI 社の音声→テキスト変換モデル（whisper-1）を用いた際のテキスト文字数である．このテキストは gpt-4 への入力にそのまま使用した．

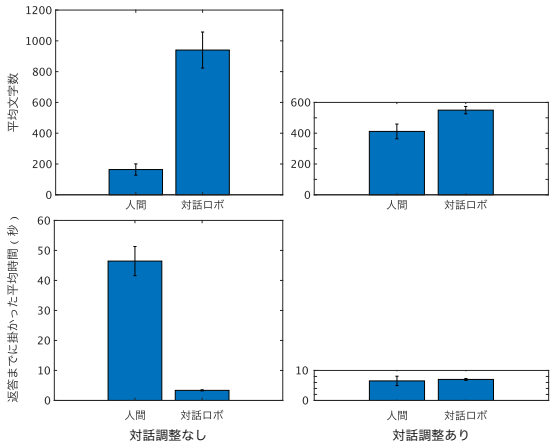

図 2　返答文字数（上段）と返答時間（下段）の平均値．左は対話調整のない金野ら [5] の実験結果，右は対話調整を行なった結果．エラーバーは標準誤差．

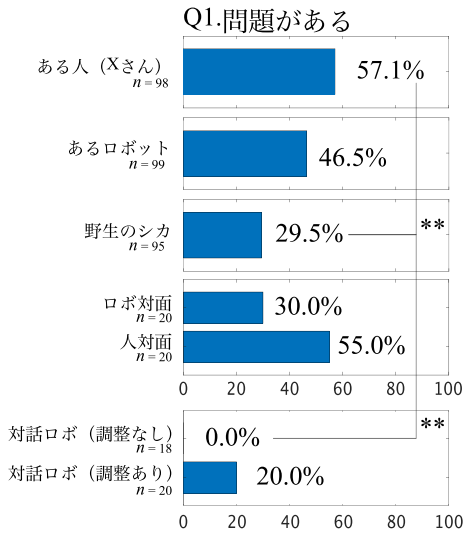

図 3　トローリー課題の回答（最初の問いの抜粋）．

た．ただし，一要因分散分析では主効果が確認され（$F(1,30) = 6.67, p = .015$），ロボットの方が有意に多い結果であった．また，返答時間についても先行する実験では大きな差があったが，調整後の実験では人間が平均 6.51 秒，ロボットが 7.01 秒とその差はなくなった（$F(1,30) = 0.10, p = .76$）．

上記の結果を踏まえた上で，トローリー課題での結果を確認する．紙幅の関係から，三つの質問のうち大きな差が生じる最初の質問「問題があると思うか」に，「問題がある」と回答した人数の割合を図 3 示した．

ある人（X さん），あるロボット，および野生のシカ条件についての結果は，小松ら [2] のアンケート調査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[4]．小松らの調査との主な違いは，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）」に変更されたこと[5]と，行為の是非についての質問が，道徳的な責任の程度を問うものから，三つの質問に変更されたことの二点である．

ロボ対面および人対面の条件は，堀川ら [6] による実験の結果であり，対話ロボ条件が今回実施された対話実験の結果である．それぞれの条件での回答人数は n で表記した．また，図にはフィッシャーの直接確率検定の後にボンフェローニの補正を行ない，有意差のあった箇所に印をつけた（$*$ は 5%，$**$ は 1%の有意水準）．

結果は，先行実験（調整なしでの対話ロボ条件）で，問題があると答える人数が 0 人になるような極端な結果ではないが，それでも，「問題がある」と回答したのは 20 名中 4 名（20%）と，功利主義側の評価に大きく偏る結果となった．

4.　議論

ウェブアンケートにおいては，一人を犠牲にして五人を助ける行為に「問題があると思うか？」と訊ねると，およそ半分の人が「問題がある」と回答する．これは，修理工のような専門職ではない一般人に設定しても同じであった．また，その行為の主体がロボットに変わっても，およそ半数は「問題がある」と回答した．これらの結果から，人間はロボットを人間と同様にジレンマ性を抱えた存在として捉えているのかと言えば，そうではないようである．なぜなら，人型のロボットを目の前にした場合には，問題があると答える割合が減少するからである．そしてその程度は，ウェブアンケートにおいて野生のシカの行為を評価させた場合と同程度である．しかしこの割合の低下は，実物に対面することによって一律に起こることではない．人間に対面した場合には，問題があると回答する人数の割合は 5 割程度になるからである．

ウェブアンケートにおいて，実物が目の前にいない状態でその行為の評価が求められる場合には，人間は素朴に，人間に対するのと同じ評価の枠組みをロボットにも適用するのかもしれない[6]．しかし実物に対峙

[4] 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．

[5] 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．

[6] ただし，罪に問えるか？という二番目の質問において罪に問えると答える人数の割合は，ウェブアンケートにおいても人間の場合に比べて，ロボットでは大きく低下するため，ロボットが人間とは異なる存在であることは認識されているものと思われる．

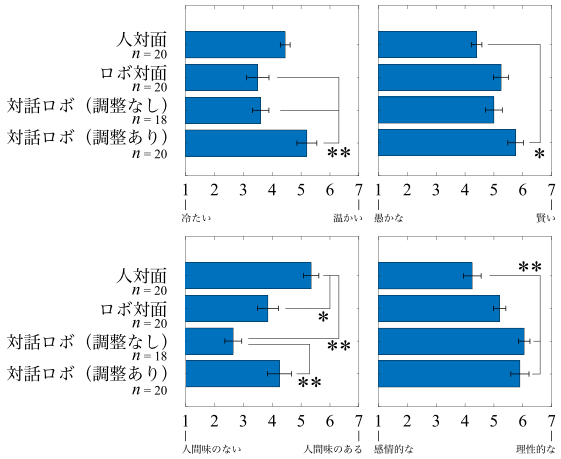

図 4　事後アンケートにおける形容詞対での印象評価．エラーバーは標準誤差．*は 5%，**は 1%の有意水準を示す．

すると，その行為に問題があると回答する人数は減る．この結果は，将来的にロボットが人間の身近な存在となった場合にも，そのロボットの行為を人間と同様に評価することはないだろうことをうかがわせる．

この結果を受けて我々は，人間がロボットと対話を交わすことによって，その評価はウェブアンケートと同様に，より人間に近い状態に戻るのではないかと予想した．ところが結果は，実物に対面した以上に，ロボットの行為の評価を功利主義の立場に偏らせるものとなった．

このような結果になった要因について，その原因を探るために事後アンケートの回答を分析した．先行する実験室実験と同様に，参加者には実験後の事後アンケートで相手（人間もしくはロボット）の印象を 4 つの形容詞対で尋ねた（Q1. 冷たい–温かい，Q2. 愚かな–賢い，Q3. 人間味のない–人間味のある，Q4. 感情的な–理性的な）．その回答を分析すると，対話したロボットに対して人間は，人間より理性的で，かつ賢いという印象を持ったようであることがわかる（図 4）．

対話の内容として，OpenAI 社の GPT は主観的な考えを生成しないようになっているため，冒頭で述べたような主体的な語りを参加者が聞くことは基本的にはなかったと思われる．実験の結果からは，その状態であれば，人間はロボットに理性や賢さを見出し，功利主義的な選択を期待するようであることがうかがえる．この功利主義的な判断を期待する傾向は，ロボット（人工知能）を道具的な立場に留めるものと考えられる．

我々が危惧するのは，もしロボットが対話の中で主体的な，あるいは感情的な表現を生成した場合に，そこに人間が抱えるような葛藤や愚かさを見出し，そのロボットを人間と本質的に同じ存在と捉えることがあるのだろうかという点である．我々の実験枠組みは，これを検証するプラットフォームとして機能するのではないかと考えている．

5.　結論

大規模言語モデルを用いて，人型ロボットと人間が対話した場合，人間はそのロボットの行為をどのように評価するだろうか．この問いを検討するために，我々はトロッコ問題を題材にした実験を実施した．結果，トロッコ問題に含まれる功利主義と義務論の間の葛藤は，ロボットの行為の評価に対しては功利主義の立場に偏らせる効果があることが確認された．このような結果となったのは，対話がそのロボットを理性的で賢い存在であると印象づけたからではないかと考えられる．

謝辞

本研究は，JSPS 科研費基盤研究 (A)「道徳的行為者のロボット的構築による＜道徳の起源と未来＞に関する学際的探究」/課題番号 19H00524 の助成を受けました．

文献

[1] Nagataki,S., Hashimoto, T., Kashiwabata, T., Tachibana, K., Konno, T. (2025). Toward a Moral Theory for Society of Metasapiens: Exploring the Philosophy of Diverse Embodiment. The proceedings of 30th international symposium on artificial life and robotics (AROB 2025), 7 pages.

[2] 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.

[3] Foot, P. (1967). The Problem of Abortion and the Doctrine of Double Effect. Oxford Review, 5, 5–15.

[4] Thomson, J. J. (1985). The Trolley Problem. The Yale Law Journal, 94(6), 1395–1415.

[5] 金野 武司, 堀ノ 文汰, 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 (2024). 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, 日本認知科学会第 41 回大会（JCSS2024）予稿集, pp.215-218.

[6] 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.