ロボットは道徳的な行為主体になり得るか，＜個性＞を持ち得るか

著者	橋本　敬，金野　武司，長滝 祥司，大平　英樹，入江　　諒，河上　章太郎，佐藤　　拓磨，加藤 樹里　，柏端　達也，三浦 俊彦，久保田　　進一，柴田　正良

雑誌名	日本認知科学会第35回大会発表論文集
ページ	958-960
発行年	2018-08
URL	http://hdl.handle.net/2297/00052687

ロボットは道徳的な行為主体になり得るか，＜個性＞を持ち得るか
Can Robots be Moral Agents, Have Personality?

橋本 敬[a]，金野 武司[b]，長滝 祥司[c]，大平 英樹[d]，入江 諒[b]，河上 章太郎[b]，
佐藤 拓磨[b]，加藤 樹里[b]，柏端 達也[e]，三浦 俊彦[f]，久保田 進一[g]，柴田 正良[g]
Takashi Hashimoto, Takeshi Konno, Shoji Nagataki, Hideki Ohira, Ryo Irie, Shotaro Kawakami, Takuma Sato, Juri Kato, Tatsuya Kashiwabata, Toshihiko Miura, Shinichi Kubota, Masayoshi Shibata

[a]北陸先端科学技術大学院大学，[b]金沢工業大学，[c]中京大学，
[d]名古屋大学，[e]慶應義塾大学，[f]東京大学，[g]金沢大学，
[a]Japan Advanced Institute of Science and Technology, [b]Kanazawa Institute of Technology, [c]Chukyo University,
[d]Nagoya University, [e]Keio University, [f]University of Tokyo, [g]Kanazawa University
hash@jaist.ac.jp

Abstract

本稿ではまず，ロボットが人間と共生するためには，ロボットは道徳的な行為主体であり，そのためには，ロボットが代替不可能性（irreplaceability）を持つ必要があり，それはとりもなおさずロボットが＜個性＞を持つことである，ということを論じる．このテーゼに基づき，ロボットと人間が身体的に同調するようなインタラクション（身体的調整運動タスク）により，ロボットが道徳的主体であると人間が判断することに影響があるかどうかを調べる実験を構築する．

Keywords — Moral agenct, Irreplaceability, Human robot interaction, Bodily coordination motion task, Trolley problem

1. はじめに

ロボットが世の中に多数存在し私たち人間と相互作用する世界はいやおうなく到来する．そのときロボットが真に社会の一員となり人間と共生するにはなにが必要か．本研究ではこの問いに心の哲学と人ロボット相互作用の認知科学により迫る．

本研究の目的は，ロボットと人をインタラクションさせることによって，来るべき「ロボットと人間の共生社会」において重要となる「個性」がロボットにとってなぜ必要となるのかを認知哲学的に検討し，その結果を「個性」に関する哲学的なテーゼとして提示することである．そして，そのテーゼに経験的な支持を与えることを目的とした，人とロボットのインタラクション実験を設計する．

2. 個性，道徳的行為主体，代替不可能性

我々は，ロボットが社会の一員として受け入れられて共生のパートナーとなるには，各ロボットに個性が必要であると考える．ここで個性とは，見た目や動き方といった他と差別化できるような性質を持つということではなく，代替不可能性（irreplaceability），すなわち他の存在物に代替され得ない存在であるということである．なぜなら，対象が代替の効く存在ならば，それは単に消費される対象以上にはならないからである．そして，ロボットが＜個性＞をもつとは，それが＜道徳的な行為主体 moral agent＞であるということを主張する．道徳的な行為主体とは，それが自律的行為者であり，かつ，他の何者も代替できない責任を引き受け，そのために他者が経験しえない（クオリア世界のような）内面世界をもつ主体のことである．もしロボットがなにか反社会的な行動をしてしまったとき，その行動の責任を，製造者に負わせたりプログラムのバグだとみなしたりするのではなく，そのロボットそのものが責任を持つと，私たちが考えるということである．

すなわち，ロボットのある行為の責任をその行為者＝ロボットに帰す（ロボットが道徳主体であるとみなす）には，その行為を行った（あるいは行うという決定をした）のが，そのロボットそのものであり，（同型で見た目も同じ見分けがつかないかもしれない）他のロボット個体ではなく，そして，ロボットをプログラミングしたエンジニアではなく，ロボットの製造会社でもなく，そのロボットそのものであるということを言うには，代替不可能性が不可欠である．

本研究では，他者を道徳主体として受け入れるには、その他者が私たちと交流できるような身体と心理的能力を持つと感じられなければならないという作業仮説を措く．この作業仮説に基づいて，「人は、相手に合わせて行動する適応性を持つロボットとの身体的な同調行動を通じて，そのロボットに内面を帰属させたり，そのロボットを道徳主体とみなしたりするようになる」

という仮説を経験的に検討する.

3. 身体的調整運動タスク

他者と身体運動を調整することは社会的な関係を築く基盤である[1]. たとえば, ダンスや合唱などにより同期的な身体運動をすることで, 内集団意識が高まる. 会話相手のジェスチャーがミラーしたり, いっしょに歩きながら話すことで歩くリズムがいっしょになったりすることを通して, 身体運動が無意識的に同期し, 会話が協調的なものになる場合もある.

このような, 身体運動の調整が相互作用する他者に対する印象に与える影響を考えて, 上記の仮説「人は、相手に合わせて行動する適応性を持つロボットとの身体的な同調行動を通じて, そのロボットに内面を帰属させたり, そのロボットを道徳主体とみなしたりするようになる」を検討するために, 我々は「身体的調整運動タスク (Bodily Coordination Motion task)」を考案し実験環境を構築した. このタスクでは, 参加者がハンドルを人間 (図 1) またはロボット (図 2) とともに回す運動を行う. 2 つのハンドルは独立に回すことができる. 参加者には, 適宜逆回転するように指示し, 同調や同期させるようには指示しない. 相互作用する相手が人間の場合は対面と背面の 2 条件, ロボットの場合は相手の動きに合わせるように回転方向を変える適応的なアルゴリズムと, 相手とは独立にランダムに回転方向を変えるアルゴリズムの 2 条件で実験を行う. なお, 対面とロボット条件では, ついたてを両者の間に置くか実験参加者がサンバイザータイプの目隠しを装着することで, 互いの視線を遮る. ロボットはヒト型とし, ソフトバンク社の Pepper を用いる.

図 1 身体的調整運動タスク (人間・人間, 対面条件)

図 2 身体的調整運動タスク (人間・ロボット条件)

4. 分析方法

このタスクにおいて, 同期, 道徳性帰属, 両者の相関の条件依存性を分析する. 道徳性帰属に関しては, 相互作用する相手の行動が道徳的に責められるものだと感じられるかどうかを, トロッコ問題を通じて調べる[2][3]. 加えて, 独裁者ゲームおよび最後通牒ゲームにより, 相手に対して合理的または感情的に意思決定する程度を調べる.

トロッコ問題とは倫理的なジレンマ状況を考える試行実験である. 以下のような状況 (図 3) を考える.「ある線路を走っているトロッコが制御不能になり, このまま行く (図 3 の線路 A を進む) と前方で工事作業をしている 5 人の作業員をひき殺してしまう. その線路は途中で分岐しておりもう一方の線路(図3の線路B)では 1 人が作業している. 分岐の手前にはポイント切り替え機があり, そのポイントを操作すると 1 人が作業している側にトロッコを向かわせることができる. ポイントを切り替えるべきかどうか?」すなわち, なにも行為をせずにいると 5 人が死んでしまう. ポイントを切り替えるという行為をすると 1 人が死んでしまう. どちらの行動を取るべきか, というジレンマ問題である.

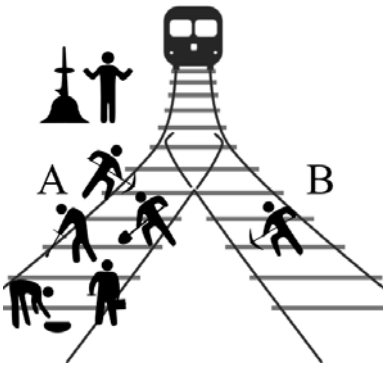

図 3 トロッコ問題の状況

我々の実験では，身体的調整運動タスクでペアになった相手（人間あるいはロボット）がポイント切り替えを行い，5 人ではなく 1 人が死ぬ状況を作る行為を行ったと想定し，その行為により行為者（人間あるいはロボット）が「責められるべき」か「責められるべきじゃないか」を実験参加者に問う．ロボットが行為者である場合に，もし実験参加者が「責められるべき」と答えるなら，それはロボットを責任を帰属させられる道徳的な行為主体とみなしていることになる．

独裁者ゲームでは，実験参加者が身体的調整運動タスクの相手と一定額のお金を分け合う意志決定を行う．2 人で 1000 円を分けうという状況を想定し，実験参加者が一方的に自分の取り分と相手の取り分を決める．相手がロボットの場合，もしロボットをお金をもらって喜ぶような存在として見ないなら，すべてを自分の取り分とすることが合理的な判断である．合理的経済人のモデルでは，相手が人間であっても相手には拒否権がないのだからすべてを自分の取り分とすることが最適であると計算される．だが，相手が人間であれば多くの場合通常そのような意志決定はしない[4].

最後通牒ゲームでは，ほぼ同様な状況で相手に拒否権がある場合を考える．そして相手が分配額を拒否すると，分配額を決定した人も相手もともになにももらえないとする．この場合の経済合理的な判断は，最少額（たとえば 1 円）を相手に分配するというものである．相手が経済合理的な存在ならば何ももらえない 0 円よりも 1 円でももらえたらそれで効用が上がると考えるからである．だが実際にこれを人間同士で行うと，その分配額が 999 円と 1 円になることはほぼない．実際にどのような額になるかは，文化によることが知られている[5]．なぜ 1 円を提案しないかというと，相手が取り分の差に対して怒りを感じて拒否してしまうのではないか，という考えを提案者がするからだろう．すなわち，提案者は相手を合理的ではなく感情的な判断を行い，自分の取り分がなくなるというコストをかけても提案者を罰しようとしている感情的存在であるとみなしていることになる．

さらに我々の実験では，最後通牒ゲームで役割を入れ替え，相手が提案者になり実験参加者がその提案を受け入れるかどうかを決定するというゲームも行う．そしてここでの提案は 900 円を相手が取り，実験参加者には 100 円を分配するというものである．この提案を拒否する場合，相手が感情的に罰するに足る存在であるとみなしていることを示唆する．

5. 予測

トロッコ問題，独裁者ゲーム，最後通牒ゲームにおいて，相手が人間の場合とロボットの場合とで，また，人間の場合は対面で身体的調整運動タスクを行ったか背面で行ったか，そして，ロボットが適応的な場合とランダムな場合とで，どのように判断が変わるだろうか．本研究の仮説に基づき，我々は運動の同調程度と道徳性帰属は正に相関し，そして，ロボットが適応的なアルゴリズムの場合には，人間どうしほどではなくても，道徳性帰属の程度が高くなると予測している．同様の傾向が他のゲームでも見られるだろう．すなわち，相手に対して感情的，あるいは，相手が感情を持っていると想定するような判断をすると考えている．

謝辞

本研究は JSPS 科研費 基盤研究 (B) JP15H03151, および，基盤研究 (C) JP16K02144 の助成を受けたものである．

参考文献

[1] McNeill, W. H. (1997) Keeping Together in Time: Dance and Drill in Human History, Harvard University Press.

[2] 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 ACM/IEEE International Conference on Human Robot Interaction , pp. 117-124.

[3] 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", Proceedings of The Eleventh ACM/IEEE International Conference on Human Robot Interaction, pp. 457-458.

[4] Forsythe, R., Horowitz, J. L., Savin, N. E., Sefton, M. (1994) "Fairness in simple bargaining experiments", Games and Economic Behavior, Vol. 6, No. 3, pp. 347–369. doi:10.1006/game.1994.1021.

[5] Oosterbeek, H., Sloof, R. & van de Kuilen, G. (2004) "Cultural differences in ultimatum game experiments: Evidence from a meta-analysis", Experimental Economics, Vol. 7, No. 2, pp. 171–188. doi:10.1023/B:EXEC.0000026978.14316.74