On The Robot As A Moral Agent

Shoji Nagataki
School of International Liberal Studies,
Chukyo University
101-2 Yagoto Honmachi, Showa-ku,
Nagoya-shi, Aichi (Japan) 466-8666
shojinagataki@gmail.com

Masayoshi Shibata
Vice President in charge of Education,
Kanazawa University
Kakuma-machi, Kanazawa (Japan) 920-1192
mshibata@staff.kanazawa-u.ac.jp

Takashi Hashimoto
School of Knowledge Science, Japan
Advanced Institute of Science and
Technology
1-1 Asahidai, Nomi, Ishikawa (Japan) 923-1211
hash@jaist.ac.jp

Tatsuya Kashiwabata
Faculty of Letters, Keio University
2-15-45 Mita, Minato-ku, Tokyo (Japan)
108-8345
tatuya@flet.keio.ac.jp

Takeshi Konno
Electrical and Electronic Engineering,
Kanazawa Institute of Technology
7-1 Ohgigaoka, Nonoichi-shi, Ishikawa
(Japan) 921-8501
konno-tks@neptune.kanazawa-it.ac.jp

Hideki Ohira
Graduate School of Informatics,
Nagoya University
Furo-cho, Chikusa-ku, Nagoya-shi,
Aichi (Japan) 464-8601
ohirahideki@gmail.com

Toshihiko Miura
Faculty of Letters, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo (Japan)
113-8654
miurat@jcom.home.ne.jp

Shin'ichi Kubota
Tokoha University (part-time lecturer)
6-1 Yayoicho, Suruga-ku, Shizuoka-shi,
Shizioka (Japan) 422-8581
pxk06600@nifty.com

ABSTRACT

To be a moral agent is to bear its own responsibility which others cannot take for it. We hold that such irreplaceability consists in its having an inner world to which others cannot have direct access. The purpose of this paper is to propose, as a means of gaining support for our thesis, an experiment --- a psychological one in which to assess to what degree we can attribute moral responsibility to a robot. Furthermore, we explore the possibility of a society where humans and robots coexist.

CCS CONCEPTS

• Human-centered computing → Human computer interaction (HCI) → HCI theory, concepts and models

KEYWORDS

Human-robot coexistence, Personality, Moral agent

ACM Reference format:
Shoji Nagataki, Masayoshi Shibata, Takashi Hashimoto, Tatsuya Kashiwabata, Takeshi Konno, Hideki Ohira, Toshihiko Miura, Shin'ichi Kubota. 2018. On The Robot As A Moral Agent. In *Proceedings of A XIX International Conference on Human Computer Interaction (Interacción 2018). ACM, New York, NY, USA, 4 pages.* https://doi.org/10.1145/3233824.3233832.

1 INTRODUCTION

What is necessary for robots to live together with human beings? Arguably one of the most important conditions is the possibility of ascribing morality to them. To be a moral agent is to bear its own responsibility which others cannot take for it. Our thesis is that such irreplaceability consists in its having an inner world --- a private realm or a first-person perspective to which others cannot have direct access. In fact, this kind of otherness is quite familiar. In our daily lives, we almost always find opacity, impenetrability, compulsiveness resistance as well as much affinity between ourselves. Suppose that you and I have lunch together. If you confidently and strongly recommend me to eat something. I have never expected in a restaurant, I might feel as if I have lost my initiative. We usually have a sense of otherness in unexpected transfers of initiative.

Interacción 2018, September 12–14, 2018, Palma, Spain
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6491-1/18/09…$15.00
https://doi.org/10.1145/3233824.3233832

In order to gain support for our thesis, we are planning an experiment of interaction between a robot and a human, and setting human subjects to the task of assessing, in various contexts, how much morality he or she attributes to the robot. We will conduct this experiment with an auxiliary hypothesis in hand: if we feel something inscrutable in an object, we have an inclination to assign some kind of morality to it.

Put roughly, there are two general approaches to the study of humanoid robotics: one focusing on appearance and behavior, putting much weight on mimicking those of humans, the other on explicating and reproducing our "inner" cognitive functions. The latter tries to implement functions similar, in some respects, to those of the human mind, as [1] suggests. In congruence with these approaches, we design the experiment of human-robot interaction in which we employ two types of robots. The type one is just implemented with bodily abilities similar to those of humans, whereas the type two is equipped with those bodily abilities as well as an ability of apparently behaving in consideration of the human mind. After the experiment, we measure how much morality the human subjects attribute to each robot by setting them three ethical problems: Trolley problem, Dictator game and Ultimate game.

2 EXPERIMENTAL DESIGN

In order to make explicit such an aspect of our daily experience, we developed an experiment using Bodily Coordinated Motion task (BCM task). A bodily coordination is a social art and one of the key elements which enables us to have a social relationship with others [2,3,4]. When coordinating ourselves well and getting along with each other, we would feel an affinity between us, while when failing in it, a sense of otherness or impenetrability would be imposed upon us.

Participants engaged in BCM task are evaluated how much morality she/he attributes to the paired partner (a human and robots with poor or rich inner mechanism) in the setting of BCM test. Our prediction is that the participant will attribute more morality to the “rich” robot than to the “poor” one.

2.1 Participants

The participants are 40 right-handed people (all male). They interact with a partner in the BCM task. Based on a between-subjects design, they are allocated into one of 4 conditions.

2.2 Experimental Design

A participant engages in BCM task with a human or a robot. In interaction with a human, there are two conditions: face to face and back to back. We have two types of mechanism of robot's motion: rich and poor. The rich mechanism robot follows human's rotation, and the poor mechanism changes the direction of rotation with random intervals. The interaction with a robot is always face to face. We use a humanoid type robot, Pepper (SoftBank Robotics Corp.) . The conditions are summarized in Table 1.

Table 1: Frequency of Special Characters

Partner	Types of interaction	
Human	Face to Face	Back to Back
Robot	Rich mechanism	Poor mechanism

2.3 Bodily Coordinated Motion Task

Each participant conducts a bodily coordinated motion task with a human partner (Fig. 1) or a robot partner (Fig. 2). She/He is told that the experiment is to evaluate the smoothness and flexibility of motion of humans and robots, though its actual purpose will be revealed afterward. She/He is asked to continue rotating the handle, at a moderate and constant speed, during the task, changing the direction of rotation randomly. The task consits of 5 sessions and one session continues 90 seconds. A typical example of motion of a participant and a human partner is shown in Fig. 3.

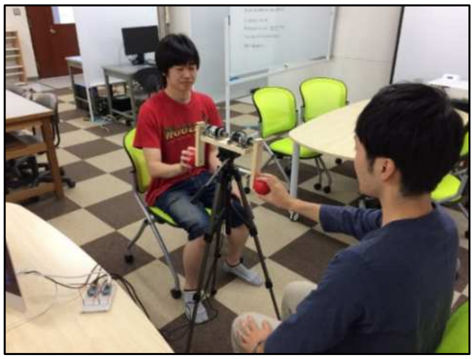

Figure 1: BCM task in human-human and in face to face condition

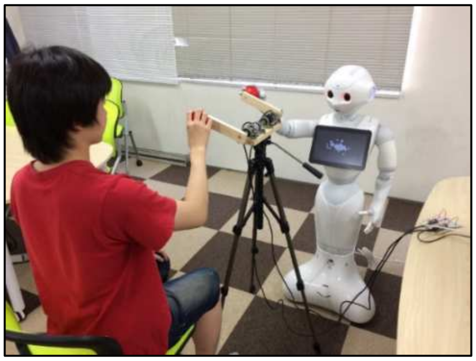

Figure 2: BCM task in human-robot condition

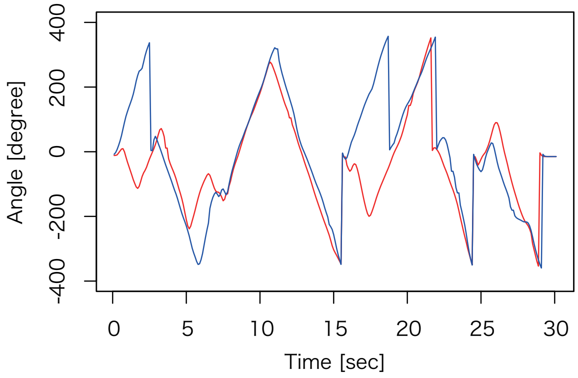

Figure 3: Example of time series of rotation motion in the human-human condition. The X axis is time and the Y axis is the angle from the initial position.

2.4. Measures

The participants' impression related to moral judgement about the partners are evaluated using the trolley problem, the dictator game, and the ultimate game. The participants are instructed that these games are the different experiment from the BCM task, which will be debriefed at the end of experiment.

2.4.1 Trolley Problem. In Fig. 4, there is a lever which can change direction of the trolley [5,6,7].

1. Do nothing, and the trolley kills the five people.
2. Pull the lever, the trolley will kill one person.

The human partner or the robot partner in the BCM task makes a utilitarian decision (pull the lever to kill one person to save five people). Participants rate how moral responsibility the human or robot partner has. We predict the result of this test as indicated in Fig. 5.

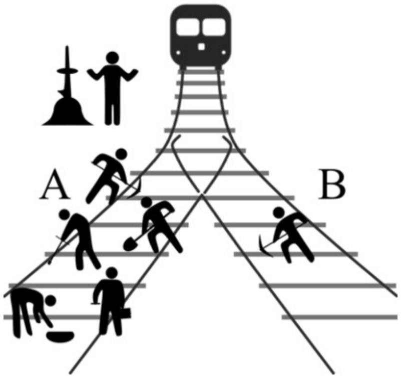

Figure 4: Trolley problem: A trolley is barreling down the railway tracks.

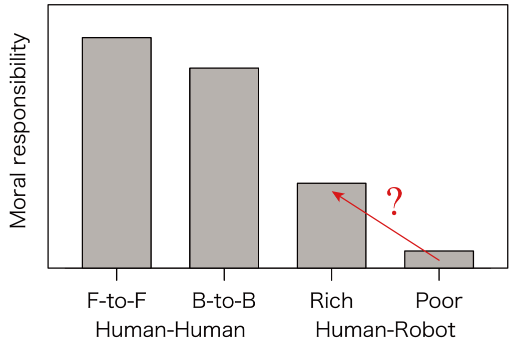

Figure 5: Prediction of the results in the trolley problem. "F-to-F" and "B-to-B" mean the "face to face" condition and the "back to back" condition with a human partner.

2.4.2 Dictator Game. Two players splits an amount of money (e.g. 10 Euros)[8].

1. A dictator can decide how to split the money.
2. A recipient simply receives the allocated money.
   e.g. A dictator takes 7 euros and a recipient is given 3 euros.

The dictator has no risk to make any policy to split the money. Thus, the endowment from the dictator to the recipient is thought to reflect selfishness or preference of equality. Each participant plays the role of dictator. We predict the result of this test as indicated in Fig. 6.

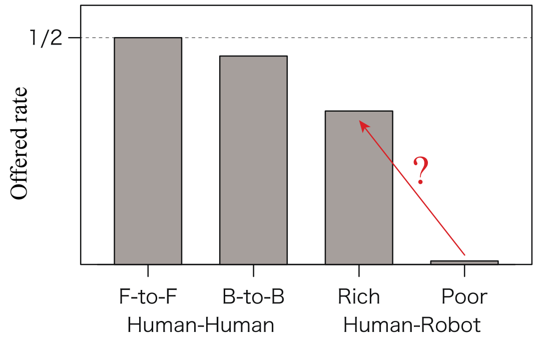

Figure 6: Prediction of the results in the dictator game. The Y axis is the rate of offering in the game.

2.4.3 Ultimate Game. Two players negotiate to divide an amount of money (e.g.10 Euros).

1. A proposer makes a proposal how to split the money.

   e.g. A proposer takes 7 euros and give 3 euros to a responder.

A responder makes a decision whether he or she accepts or rejects the proposal. If the responder rejects the proposal, both players can get no money. Thus, rejection in this game is thought as costly punishment to unfair others. Each participants play the role of the proposer and then the responder. Our predictions of this game are shown in Fig. 7 when the participant is the proposer and in Fig. 8 when she/he is the responder.

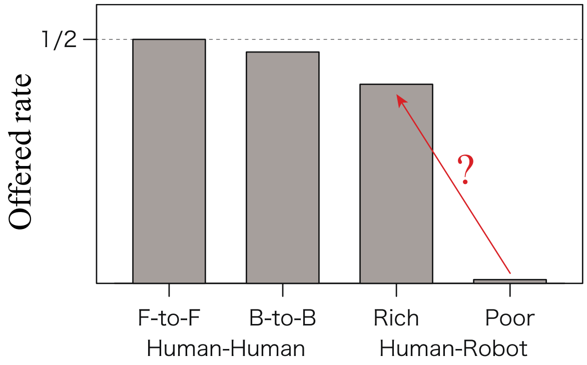

Figure 7: Prediction of the results in the ultimate game when a participant plays the role of proposer. The Y axis is the rate of offering in the game.

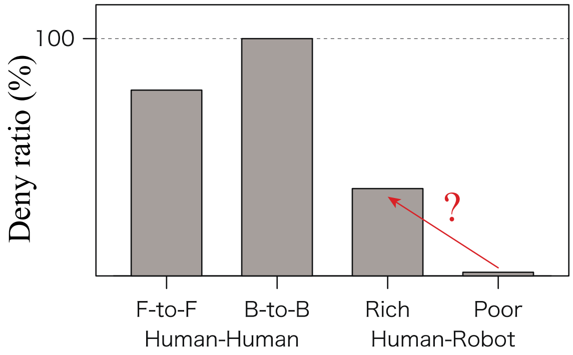

Figure 8: Prediction of the results in the ultimate game when a participant plays the role of responder. The Y axis is the rate of denying in the game.

A usual robot (the "poor mechanism" in this study) is not recognized as a moral agent. Thus, participants will not attribute moral responsibility to the robot's utilitarian decision in the trolley problem, will allocate almost no money to the robot in the dictator game, and will behave in a very selfish ways in the ultimatum game. Contrarily, a robot which showed complicated patterns of synchronized actions in the BCM task (the "rich mechanism" in this study) will be somehow regarded as a moral agent like a human. Thus, participants' evaluation for the utilitarian decision by the robot in the trolley problem, and behaviors to the robot in economical negotiation games will be similar to those to a human.

3 ANTICIPATED RESULT AND TWO HYPOTHESIS

One hypothesis is that a participant will attribute a certain moral agency to the other (even to a robot) with whom the participant can bodily coordinate in a better manner. This is because the coordination involves the process of mutual understanding in some respects [9]. Generally speaking, even with a new acquaintance of others, we naturally develop a concern for them. In parallel with that, we come to think that others should have a similar concern for us in turn. One can recognize a primitive basis for ethics in this situation.

Another hypothesis is that the richer world we recognize within others, the more demand for morality we make (as mentioned above, we set up three kinds of moral judgment task to test this hypothesis). On a simple setting of BCM task, however, what is it like to recognize a richness — or an inscrutable realm which underlies personality — within others?

In our experiment, participants may succeed in bodily coordination or fail. There are also conflicts as to which participant possesses the initiative. In the case of a conflict where the coordination once fails, a transition of the initiative eventually will take place, and a new coordination will hold, we presume. In the bodily coordination process with the other, you may have not only a sense of the partner's being in tune with yourself. You may also feel her/his resistance or the shift of the initiative to the other side. The experience of coordination can be a complicated one full of twists and turns.

We think that a participant will find a richness within his partner through a complicated process of coordinations, divergences and transitions of the initiative. This process occurs when, for instance, the participant feels its partner's purposely making an unexpected move. In such a situation, it seems natural for us to attribute intention, desire, responsibility, and so on to others. This is when we recognize others as moral agents and accept them into the intersubjective world of morality.

Therefore, we would recognize personality and morality within a machine if we could interact with them in our everyday life. In the near future, we might experience a new world coexisting with robots which have a kind of subjectivity and morality.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (B), Grant Number JP15H03151, (C), Grant Number JP16K02144, Innovative Areas Grant Number JP18H04206, and Innovative Areas #4903 (Evolinguistics) Grant Number JP17H06383..

REFERENCES

[1] Shoji Nagataki, M. Shibata, T. Konno, T. Hashimoto, and H. Ohira. 2013. Reciprocal Ascription of Intentions Realized in Robot-human Interaction. In *Proceedings of the 35th annual meeting of the cognitive science society (CogSci2013)*. 4061..

[2] William H. McNeill. 1995. *Keeping Together in Time: Dance and Drill in Human History*. Harvard University Press.

[3] Alexander Mortl, A. T. Lorenz, and S. Hirche. 2014. Rhythm patterns interaction-synchronization behavior for human-robot joint action. PloS one, 9(4):e95195. DOI: 10.1371/journal.pone.0095195.

[4] Kyongsik Yun, K. Watanabe, and S. Shimojo. 2012. Interpersonal body and neural synchronization as a marker of implicit social interaction. Scientific reports, 2(959). DOI: 10.1038/srep00959.

[5] Joshua. D. Greene, L. E. Nystrom, A. D. Engell, J. M. Darley, and J. D. Cohen.2004. The neural bases of cognitive conflict and control in moral judgment. Neuron, 44(2):389–400. DOI: 10.1016/j.neuron.2004.09.027.

[6] Bertram F. Malle, M. Scheutz, T. Arnold, J. Voiklis, and C. Cusimano. 2015. Sacrifice One For the Good of Many?: People Apply Different Moral Norms to Human and Robot Agents In *HRI '15 Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* 117-124. DOI: 10.1145/2696454.2696458.

[7] Takanori Komatsu. 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 Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on, 457–458. IEEE, DOI: 10.1109/HRI.2016.7451804.

[8] Takahiro Osumi and H. Ohira. 2009. Cardiac responses predict decisions: An investigation of the relation between orienting response and decisions in the ultimatum game. International Journal of Psychophysiology, 74(1):74–79. DOI: 10.1016/j.ijpsycho.2009.07.007.

[9] Scott S. Wiltermuth and C. Heath. 2009. Synchrony and Cooperation. In *Psychological Science*, Vol.20, No. 1, 1-5. DOI: 10.1111/j.1467-9280.2008.02253.x.