Robot As Moral Agent: A Philosophical and Empirical Approach

Name : Shoji Nagataki[†1], Masayoshi Shibata[†2], Tatsuya Kashiwabata[†3], Takashi Hashimoto[†4], Takeshi Konno[†5], Hideki Ohira[†6], Toshihiko Miura[†7], Shinichi Kubota

E-Mail : shojinagataki@gmail.com †1 Chukyo University †2 Kanazawa University †3 Keio University †4 Japan Advanced Institute of Science and Technology(JAIST) †5 Kanazawa Institute of Technology †6 Nagoya University †7 University of Tokyo

I. Introduction

What is necessary for robots to coexist with human beings? In order to do so, robots must be moral agents. To be a moral agent is to bear its own responsibility which others cannot take for it. We argue that such an irreplaceability consists in its having an inner world. The personality of a moral agent is firmly rooted in such an inner world.

II. Purpose of This Presentation

It is necessary to implement similar bodily and psychological abilities in someone, or something to be accepted as a moral agent, or another person in a human society. Then the irreplaceability, which we mentioned, can be viewed along another dimension; it is related to the problem of whether a first-person perspective can be attributed to *the other* in question. This kind of perspective involves a private realm to which other people cannot have direct access. This is where our personality and irreplaceability, including that of moral responsibility, lie in.

In fact, such an otherness is familiar. It is a common experience that we find similarities as well as differences between us. Suppose that you and I agree to have lunch together, but you force me to eat something I have never expected in a restaurant. In that situation, I would feel I've lost my initiative. This happens in our everyday life. We have a sense of otherness in unexpected transfers of initiative.

In order to make explicit such an aspect of our daily experience, we design an experiment of 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 (cf., William H. McNeill (1997) *Keeping Together in Time: Dance and Drill in Human History*, Harvard University Press). When coordinating ourselves well and getting along with each other, we feel an affinity between us, while when failing in it, a sense of otherness or impenetrability is imposed upon us.

The purpose of this presentation is to explicate a condition in which humans attribute the status of moral agency to a robot. For that purpose, we are planning to set up an experiment of interaction between a robot and a human. Before that, we develop some hypotheses about the anticipated results of the experiment.

IV. Anticipated Results and Two Hypotheses

One hypothesis is that a subject will attribute a certain moral agency to the other (even to a robot) with whom the subject can bodily coordinate in a better manner. This is because the coordination involves the process of mutual understanding in some respects. 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. 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, subjects may succeed in bodily coordination or fail. There are also conflicts as to which subject 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 opponent'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 subject will find a richness within his opponent through a complicated process of co-ordinations, divergences and transitions of the initiative. This process occurs when, for instance, the subject feels its opponent'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.

III. Experimental Design

Participants

Students of university in Japan : 40 right-handed people (all male).

Based on a between-subjects design, they were allocated into one of 4 conditions.

Four conditions

Human-Human		Human-Robot	
Face-to-face	Back-to-back	Rich mechanism	Poor mechanism
1. HH.FF	2. HH.BB	3. HR.Rich	4. HR.Poor

- Rich : Turn-taking mechanism
- Poor : Random mechanism

Tasks

1. Bodily Coordinated Motion Task

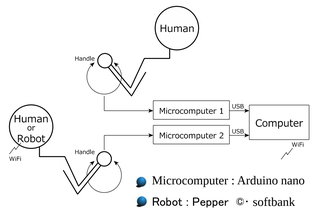


Fig.1 System diagram

Fig.2 Human-human condition

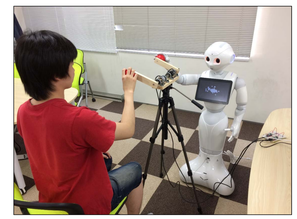


Fig.3 Human-robot condition

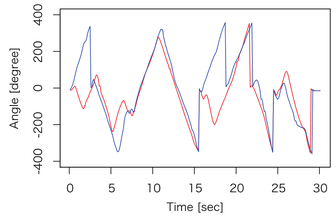
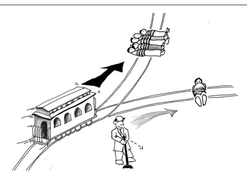


Fig.4 Time series of two angles in the human-human condition.

Each participant conducted a bodily coordinated motion task with a human partner or a robot partner. She/He was asked to continue to rotate a handle during the task and to try to match location of a red ball with the handle in another handle which a partner rotates (coordinated motion). The speed and direction of rotation were spontaneously decided by the players, without any verbal communication. With a human partner, a participant conducted this task in the face-to-face condition and in the back-to-back condition (control condition). In the rich mechanism condition, motion of the robot were complicatedly structured to simulate real-humans' motion. In the poor mechanism condition, the robot simply tracked the participant's motion. A typical example of motion of a participant and a human partner is shown in Fig.4.

2. Moral Judgment Task

Trolley problem



A trolley is barreling down the railway tracks. There is a lever which can change direction of the trolley.
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 Bodily Coordinated Motion 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.

Dictator game

Two players splits an amount of money (e.g. ¥1000).
1. A dictator can decide how to split the money.

Dictator ¥700 or ¥300 Recipient

2. A recipient simply receives the allocated money.

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.

Ultimatum game

Two players negotiate to divide an amount of money (e.g. ¥1000).
1. A proposer makes a proposal how to split the money.

Proposer ¥700 or ¥300 Responder

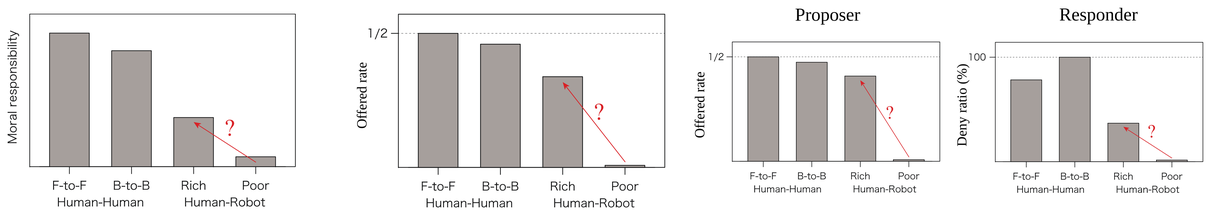
2. A responder makes a decision whether he or she accepts or rejects the proposal.

¥700 Accept ¥300

¥0 Rejection ¥0

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.

Hypothesis of results



A usual robot ("poor robot" 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 Bodily Coordinated Motion Task ("rich robot" 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.