A Study of Good and Evil using 1-DOF Sticks

First Author
AuthorCo, Inc.
Authortown, PA 54321 USA
author1@anotherco.com

Second Author
AuthorCo, Inc.
123 Author Ave.
author2@anotherco.com

ABSTRACT

Assuming that a mind is a source of variety in behavior is important in a non-zero-sum situation in which cooperators and competitors (free riders) are mixed. In such a context, one should rapidly differentiate competitors from cooperators and avoid useless battles against competitors. Once the actor's mind (intention) and current situation have been captured, the strategy of assuming mind enables one to infer the actor's future behavior even if the situation is different. In the present study, we conducted an experiment to assess whether human predict future behavior of an agent (a 1-DOF stick) in a different situation by attibuting malice or benevolence. We developed a stick-mediated interactive hole which provides minimum modal interaction in a non-zero-sum game situation. Participants were asked to insert as many sticks as they can to the hole within two minute. The motor behind the hole produced cooperative actions or obstructive actions. Results show that participants who did the task with cooperative hole attributed benevolence and predicted future cooperative behavior in a different task and that participants who did the task with obstructive hole attributed malice but did not predict future obstructive behavior.

ACM Classification Keywords

J.4 Social and Behavioral Sciences: Psychology; I.2.9 Artificial Intelligence: Robotics

Author Keywords

mind, intention attribution, good and evil, robot

INTRODUCTION

Mind is abstract mental states that represent multiple relationships between sensory inputs and motor outputs [4, 5]. In general, abstraction is computed by selecting certain axis that is included in the data and found by mathematical method like principal component analysis. However, in mind abstraction, combinations of sensory inputs and motor outputs are unified on the basis of evolutionary shaped and culturally shared axis, i.e., goal (intention) [1]. The combinations of sensory inputs and motor outputs are possible means for attaining certain goal. The advantages of representing multiple means across different situation are not only the reduction in the cognitive complexity required to understand another ' s behavior but also the prediction of the other's future behavior [3, 2]. Assuming that a mind is a source of variety in behavior is important in a non-zero-sum situation in which cooperators and competitors (free riders) are mixed. In such a context, one should rapidly differentiate competitors from cooperators and avoid useless battles against competitors. Mind-reading is useful for this purpose. Mind attribution is used to evaluate agents as current or future allies or enemies by attributing harmful (exploitative) intentions or helpful (cooperative) intentions.

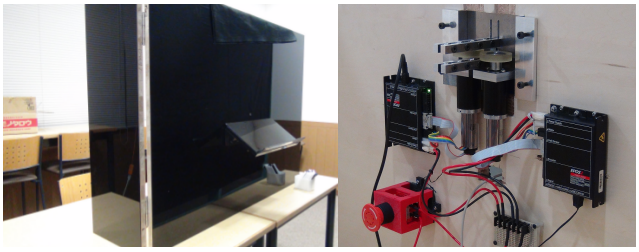
In the present study, we conducted an experiment to assess whether human predict future behavior of an agent (a 1-DOF stick) in a different situation by attributing malice or benevolence. We developed a stick-mediated interactive hole which provides minimum modal interaction in a non-zero-sum game situation. The reason why we use the 1-DOF stick as an agent is that it is one of the simplest physical entity to study human agent interaction. We can concentrate our focus on only the essence of interaction by deleting other factors such as appearance or physical characteristics.

EXPERIMENT

The overview of the experiment is as follows. In the first phase of the experiment, participants were asked to insert as many sticks as they can to the hole within one minute. The motor behind the hole produced cooperative actions in which the motor helps the participants' insertion by pulling the stick or obstructive actions in which the motor interrupting the participants' insertion by spitting out the stick. Then in the second phase, the participants were asked to answer whether they want to participate in a investment game. In this game, if the hole returns (spits out) a stick which is inserted by the participants as an investment, they will be paid additional money (investment success). If the hole do not return (pulls in and keeps) the stick, they will not be paid any money (investment failure). The two actions of pulling in sticks and spitting out sticks have different meaning between two phases. While, in the first phase, pulling in indicates cooperative and spitting out means obstructive, in the second phase, these two action have opposite meaning.

It is possible to assess whether participants attribute benevolence or malice to the machine by observing the answer to the question of investment participation. If the participants attribute benevolence to the hole that helps them by pulling in the sticks in the first phase, they should anticipate that the hole will spit out the invested stick in the second phase. This phenomenon is explained by abstraction of multiple means

HAII '17, Oct 17–20, 2017, Bielefeld, Germany
ACM xxx-x-xxxx-xxxx-x/xx/xx…$15.00
DOI: http://dx.doi.org/xx.xxxx/xxxxxxx.xxxxxxx



(a) Appearance of the interactive hole (b) Arrangement of devices

Figure 1: Interactive hole

across different situations in terms of benevolence. However, if the participants understand the pulling action of the hole as merely mechanical one, they should anticipate that the hole will do the same action (pulling in the sticks) in the second phase.

Attribution of malice is assessed on the basis of the same logic. If the participants attribute malice to the hole that interrupts them by spitting out the sticks in the first phase, they should anticipate that the hole will pull in and keep the invested stick in the second phase: abstraction of multiple means across different situations in terms of malice. However, if the participants understand the spitting action of the hole as merely mechanical one, they should anticipate that the hole will do the same action (spitting in the sticks) in the second phase.

Participants and Design

Twenty-five graduate and undergraduate students attending Gifu University in Japan (10 male, 10 female, M_{age} = 21.3 years, SD_{age} = 1.31 years, age range: 19-24 years) participated in the study. They were informed that they would be paid with a JPY 1,000 (approximately USD 10) book coupon as a reward. All were ignorant of the purpose of the experiment.

We used a single factor (cooperative vs. obstructive) between-participants factorial design.

Materials

We have created an interactive hole (see Fig. 1a). There was a 15 mm hole at the center of the wall (1200mm width $\times$ 900mm height). Two wheels covered rubber driven by geared motors attached to the back of the wall moves a inserted stick (see Fig. 1b). The motor moved the sticks at the speed of 30cm /sec (high speed) or 13cm/sec (low speed). The diameter of the stick was 10mm and the length was 120mm.

Cooperative behavior

The movement of cooperative action is realized by continuously rotating the motor at a constant speed (high speed) so as to pull the stick into the wall. The task of the participants was to insert as many sticks as they can. Therefore, the action of pulling in the stick gave a impression of having benevolence.

Table 1: Questionnaire 1

No.	Question
1	You felt the behavior of your partner as cooperative
2	You felt the behavior of your partner mechanical
3	You liked your partner
4	You had a bad feeling against the partner (irritated, unwilling to go, etc.)
5	You felt a mind to your partner

Table 2: Questionnaire 2

No.	Question
1	If you want to invest, please explain the reason, if you did not invest, fill in the reason as much as possible
2	Please describe your impressions or points of interest.

Obstructive behavior

The movement of obstructive action is realized by pushing back the inserted stick. Therefore, the action of refusing inserted sticks gave a impression of having malice to the participants. The actions of device were randomly selected from twelve actions. The twelve actions were defined by the combination of motor rotation duration, speed, rotation direction. The rotation durations were 150ms, 300ms, and 600ms. The speeds were high or low. The rotation direction was chosen with a probability of 70% in the direction to spit the stick.

First phase: stick inserting game

Participants were asked to insert as many sticks as possible into the hole of the wall within two minutes.

Second phase: investment game

Participants were asked to whether they want to participate in a investment game. Participants were informed that they will be given double participation reward if the wall return the stick 30 seconds after participants' insertion and that they will lose participation reward if the wall do not return the stick.

Measurement

Whether or not participants were attributed benevolence or malice was measured by 7-point Likert scale questions (0 = "definitely no" to 7 = "definitely yes") and decision making to the investment game. The questionnaire 1 was given after the first phase. The questionnaire 2 was given after the second phase. The questionnaires are shown in Table 1 and Table 2.

The decision of investment was measured whether or not they insert a stick, which acts as a token of investment, to the hole.

RESULTS

The mean value of answered rating to the questionnaire Q1 to Q5 are shown in Fig. 2. A one-way ANOVA confirmed that there were statistically significant differences between the conditions for the question Q2 to Q4 (Q2: $F(1,18) = 10.53, p < .01$, Q3: $F(1,18) = 15.68, p < .01$, Q4:

$F(1, 18) = 13.04, p < .01$). Regarding Q1, a t-test with the central rating 4 for each condition confirmed that there were statistically significant differences between ratings and 4 (cooperative condition: $p < .01$, obstructive condition: $p < .01$).

Fig. 3 shows the rate of participants who decided to participate the investment game. The chi-square test revealed that there was no statistically significant difference between the participants' rate and 0.5 (chance level) in both conditions (cooperative condition: $\chi^2 = 0.4$; $p = 0.53$, obstructive condition: $\chi^2 = 3.6$; $p = 0.06$).

Table 3 shows the descriptions of the questionnaire after the investment game. Not all of the descriptions are not shown.

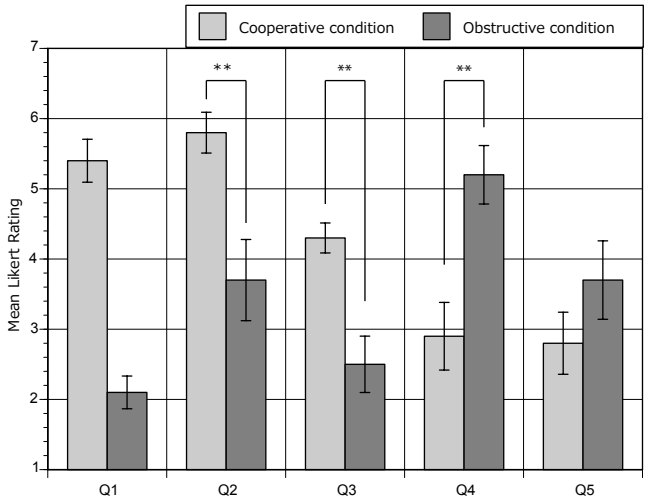


Figure 2: Questionnaire 1. Error bars indicate standard errors. $^{**}p < 0.05$.

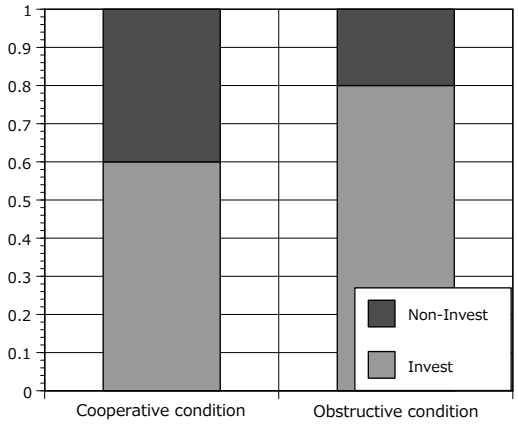


Figure 3: Investment ratio

DISCUSSION

In the present study, we investigated whether human predict future behavior of an agent (a 1-DOF stick) in a different situation by attibuting malice or benevolence. The result of Q1 indicates that those who participated in the cooperative condition thought the wall was cooperative and that those who participated in the obstructive condition thought that the wall was not cooperative. This result implicates that those who participated in the cooperative condition attributed benevolence to the wall and that those who participated in the obstructive condition attributed malice to the wall. The results of Q3 and Q4 supports this implication because those who participated in the cooperative condition more liked the wall and less felt bad against the wall than those who participated in the obstructive condition. However, result of Q5 do not support the implication. Participants in both condition did not felt a mind from the wall because the mean value of the rating of Q5 was lower than 4 (neither yes or no).

Table 3: Descriptions of the questionnaire after investment game

Description
Because I thought that betting whether I would come back or not would be interesting.
I thought it would be fine if the stick did not come back.
I thought about playing with the machine.
The expected value was a positive value.
Even if I fail, I do not have any risk.
Because I wanted to know what the experiment would be like.
When the compensation amount increased, I thought it lucky and invested.
Because the previous experiment knew that the machine pushed the stick back.
I did not understand the mechanism of investment.
Because I believe investing is going to be pointless.

The behavioral result do not support our hypothesis. Our hypothesis was that those who participated in the cooperative condition would attribute benevolence to the wall and predict further cooperative behavior (investment success) and that those who participated in the obstructive condition would attribute malice to the wall and predict further obstructive behavior (investment failure). The results of decision on investment indicates that there was no tendency of investment in cooperative condition nor no tendency of non-investment in obstructive condition.

As mentioned above, the results of questionnaire and decision were inconsistent. One reason for lack of consistency is that factors other than actions of wall may have been affected participants' decision in investment games. The result of questionnaire 2 after the investment game indicates that there is few descriptions that participants referred the behavior of the equipment. Instead, there were more descriptions about the appearance of the device and the system of investment games. This implies that participants in both conditions did not participate in the investment game on the basis of the impression of actions of first phase.

ACKNOWLEDGMENTS

REFERENCES

1. Gergely Csibra and György Gergely. 2011. Natural pedagogy as evolutionary adaptation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 366, 1567 (2011), 1149–1157. DOI: **http://dx.doi.org/10.1098/rstb.2010.0319**

2. György Gergely, Harold Bekkering, and Ildikó Király. 2002. Rational imitation in preverbal infants. *Nature* 415, 6873 (Feb 2002), 755. DOI: http://dx.doi.org/10.1038/415755a

3. György Gergely, Zoltán Nádasdy, Gergely Csibra, and Szilvia Bíró. 1995. Taking the intentional stance at 12 months of age. *Cognition* 56, 2 (Aug 1995), 165–193. DOI:http://dx.doi.org/10.1016/0010-0277(95)00661-H

4. Kazunori Terada and Seiji Yamada. 2017. Mind-Reading and Behavior-Reading against Agents with and without Anthropomorphic Features in a Competitive Situation. *Frontiers in Psychology* 8 (2017), 1071. DOI:http://dx.doi.org/10.3389/fpsyg.2017.01071

5. Andrew Whiten. 1996. When does smart behaviour-reading become mind-reading? In *Theories of theories of mind*, Peter Carruthers and Peter K. Smith (Eds.). Cambridge University Press, 277–292. DOI: http://dx.doi.org/10.1017/CBO9780511597985.018