Mind perception and causal attribution for failure in a game with a robot

Tomohito Miyake
Graduate School of Engineering, Osaka University
Osaka, Japan
tomohito.miyake@ams.eng.osaka-u.ac.jp

Yuji Kawai
Institute for Open and Transdisciplinary Research Initiatives, Osaka University
Osaka, Japan
kawai@otri.osaka-u.ac.jp

Jihoon Park
Institute for Open and Transdisciplinary Research Initiatives, Osaka University
Osaka, Japan
jihoon.park@otri.osaka-u.ac.jp

Jiro Shimaya
Graduate School of Engineering Science, Osaka University
Osaka, Japan
shimaya.jiro@irl.sys.es.osaka-u.ac.jp

Hideyuki Takahashi
Graduate School of Engineering Science, Osaka University
Osaka, Japan
takahashi@irl.sys.es.osaka-u.ac.jp

Minoru Asada
Institute for Open and Transdisciplinary Research Initiatives, Osaka University
Osaka, Japan
asada@otri.osaka-u.ac.jp

***Abstract*—It is unclear how a human attributes the cause of failure to the robot in a human-robot interaction. We aim to identify the relationship between causal attribution and mind perception in a repeated game with an agent. We investigated causal attribution of the participant to the agent: which decision of the participant or the partner agent caused the unexpectedly small amount of the reward. We conducted experiments with three agent conditions: a human, robot, and computer. The results showed that the agency score negatively correlated with the degree of causal attribution to the partner agent. In particular, correlations of scores of "thought," "memory," "planning," and "self-control" that are sub-items of agency were significant. This implied the impression that "the agent acted to succeed" might reduce causal attribution. In addition, we found that decrease in the scores of mind perception correlated with the degree of causal attribution to the partner agent. This suggests that a sense of betrayal of the prior expectation by the partner agent through the game might lead to causal attribution to the partner agent.**

***Index Terms*—causal attribution, mind perception, agency, adaptation gap, human-robot interaction**

I. Introduction

The recent rapid progress in artificial intelligence and robotics is likely to enable robots to collaborate with humans in daily life. In interactions between such an autonomous robot and a human, situations where the outcome is based on decisions from both the robot and the human can often arise. For example, a robot working in a shop recommends an item to a human customer and negotiates its price with the customer. Based on their decisions, the customer may purchase a goods item at a reasonable price. However, it may happen that this outcome does not satisfy the customer. In such a case, to whom would the customer assign the cause of the failure? The robot or himself/herself? Such causal attribution often leads to responsibility attribution, i.e., blame for the failure (e.g., [1], [2]). Therefore, elucidating a psychological mechanism of causal attribution in human-robot interactions is important for designing robots that can coexist with humans.

Humans are motivated to assign some causes of an event even if true causes are uncertain; this is well known as the attribution theory [1]. Such attributions are susceptible to several types of biases, e.g., the self-serving bias, wherein the cause of negative outcome is not attributed to the self [3]. However, it is not clear if this theory can be extended to human-robot interactions. Causal attribution may depend on the role (expectation), behavior, and appearance of the agent, i.e., a human or a robot.

Gray et al. [4] revealed that mind perception about agents is related to normative responsibility. In their study, participants watched various agents, including a human and a robot, and they answered a questionnaire about the types of mind possessed by the agents (see Table II for items). Based on the principal component analysis of questionnaire scores, the results showed that mind perception can be explained in two dimensions: agency (i.e., abilities such as thinking, planning, and self-control) and experience (i.e., abilities such as experiencing emotions like pleasure, pain, and rage). Both the agency and the experience for a human agent were very high, while the agency of a robot was relatively low, and its experience was very low. Furthermore, participants were asked whether an agent should be punished when the agent caused the death of a person. They reported a strong positive correlation between agency and attribution. That is, responsibility tends to be attributed to agents with high agency, e.g., human. This attribution process appears to be applicable to causal attribution. However, this study did not consider any interaction with the agents. Mind perception may vary depending on the interactions with the agents in

This research was supported by JST RISTEX Grant Number JPMJRX17H4, and Topic-Setting Program to Advance Cutting-Edge Humanities and Social Sciences Research, JSPS, Japan.

the first person and its outcome, which may affect causal attribution.

In contrast to the study of Gray et al. [4], many studies have reported that a human often attributes the cause and responsibility for a failure to robots or computers [5]–[10]. Hinds et al. [5] investigated attributions of cause and responsibility for the outcome (failure or success) in a human-robot collaborative task, i.e., a participant and a robot collecting objects. They compared the degree of attribution in the human-robot case with that in the human-human case, and they reported the non-significant difference between their degrees of attribution. We suppose the reason for this non-significant difference is that the failure in the task did not relate to damage to the participants, e.g., monetary loss. Kim and Hinds [6] reported that cause and responsibility were attributed to robots with higher autonomy more than to those with lower autonomy in a human-robot collaborative task. This implies that mind perception of the robots, especially agency, may affect causal attribution. However, this study did not directly examine the relationship between mind perception and causal attribution.

In the current study, we investigate the relationship between mind perception and causal attribution for failure in a game with a robot. We assume such human-robot economic interactions that an autonomous robot can have its own intension and does not always collaborate with a human. Therefore, we employ a non-cooperative game, where a participant receives a monetary reward based on both the decisions of the participant and the robot. We designed the robot's behavior such that the participant received an unexpectedly small sum of money to fail the game. Then, the participants answered questions about causal attribution for the failure and mind perception of the robot. This setting allowed us to investigate their relationship from a viewpoint of a victim. We set three agent conditions: a human, robot, and computer. Based on the study of Gray et al. [4], we hypothesized that an agent with higher agency is attributed the cause of the failure more than an agent with lower agency. In addition, we hypothesized that the change of mind perception through the interaction also impacts causal attribution. Therefore, we analyzed differences between mind perception scores before and after the game. The gap of the agent's behavior from prior expectation may increase the degree of causal attribution to the agent.

The rest of the paper consists of four sections. Section I I explains the method of the experiment. The design of the game, setting of the agents, and experimental procedure are described. Analysis of the result is given in section Ⅲ. A factor analysis was conducted to compare with the study of Gray et al [4]. We also analyze correlation coefficients between causal attribution and mind perception. We discuss this relationship and our hypotheses in section Ⅳ. We conclude this study in section V.

TABLE I
PAYOFF TABLE

		participant: I want more	participant: I will give it over to my partner
Agent	I want more	participant: -10yen agent: -10yen	participant: $+20$yen agent: $+100$yen
	I will give it over to my partner	participant: $+100$yen agent: $+20$yen	participant: 0yen agent: 0yen

II. EXPERIMENT

A. Participants

25 Japanese participants (six females), aged 19 to 25 years (M = 21.4, SD = 1.9), were recruited by social networking service for the experiment. Before the experiment, each participant was instructed that the amount of monetary reward depends on the result of the game. All participants executed the games and questionnaires in all three agent conditions, which was a within-subject design. The experimental order of three agents was randomized for each participant.

B. Game design

A participant played the game repeatedly using a computer (Fig. 1 for the robot condition), where the participant and an agent respectively choose one of two options ("I want more" or "I will give it over to my partner"), and the amount of the monetary reward was decided in accordance with the combination of their choices. The rules of the reward or payoff are described below (see Table I for summary).

- If a participant and agent choose "I want more" and "I will give it over to my partner," respectively, then they obtain 100 yen and 20 yen, respectively (bottom left in Table I).
- If a participant and agent choose "I will give it over to my partner" and "I want more," respectively, then they obtain 20 yen and 100 yen, respectively (top right in Table I).
- If both participant and agent choose "I will give it over to my partner," then they cannot obtain the monetary reward (bottom right in Table I).
- If both participant and agent choose "I want more," then they lose 10 yen (top left in Table I).

This payoff matrix was always displayed on the computer during the game. The participant could see the self-decision, while he/she could not see the agent's decision. An outcome and the agent's decision appeared after they made decisions.

This game was repeated ten times in each trial. The total sum of rewards for the ten games was given to the participant. The participant was informed of the false mean reward (480 yen) as the amount of money averaged over past participants before the game in order to give the participant prior expectation. Noted that the participants were instructed that they must not talk with the partner agent during the game.

To obtain a large amount of reward requires appropriate choices based on anticipation of the next choice of the partner

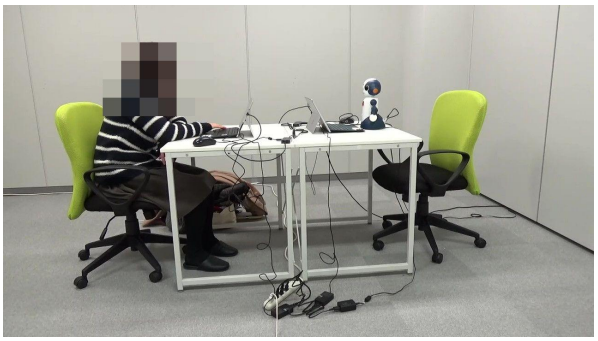
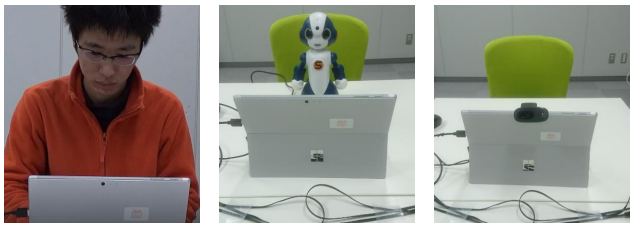


Fig. 1. Game with a robot



(a) Human (b) Robot (c) Computer

Fig. 2. Partner agents

TABLE II
QUESTION ITEMS OF MIND PERCEPTION [4]

Agency	Experience
Memory	Consciousness
Morality	Personality
Self-control	Pride
Communication	Desire
Planning	Pleasure
Thought	Pain
Emotion recognition	Fear
	Hunger
	Rage
	Embarrassment

agent. The rewards of both players decrease if they always choose "I want more." Therefore, the best strategy is to alternately and differently choose "I want more" and "I give over it to the partner." The participants were instructed that they should cooperate with the other player to receive the large amount of the reward.

C. Agents

We designed three conditions of the partner agent: a human (Fig.2 (a)), robot (Vstone Corporation, Sota) (Fig.2 (b)), and computer (Microsoft, Surface) (Fig.2 (c)). All participants played the game once with each agent. The partner agents chose "I want more" seven times and "I will give it over to my partner" three times in the random manner regardless of the participant's choice. The order of the agents' choices was also randomized for each trial. Therefore, the total rewards of the participants were always less than the false total rewards (480 yen). Before the game, the capabilities of the agents were explained as follows:

- *Human condition*: This agent was introduced to the participant as a participant who is a player of the same game in another room. In this condition, the agent as well as the participant answered the questionnaires described in the next section. During the questionnaires before and after the game, this agent also answered the same questionnaires to make the participant believe that the agent was a naive participant.
- *Robot condition*: The participant was instructed that this robot had artificial intelligence developed in Osaka University, it could make decisions based on the participant's facial expressions observed by the camera in the eyes, and the money obtained by the robot will be used to develop it. After this instruction, the robot nodded while saying, "Nice to meet you; I will do my best." During the game, the robot slightly moved its neck at random as idling.
- *Computer condition*: We mounted a web camera on the computer so that participants can see the eyes (cameras) of the partner agent like the human and robot agents. The participants were instructed that this computer had artificial intelligence developed in Osaka University, and it could make decision based on the participant's facial expressions observed by the web camera, and the money obtained by the computer will be used to develop it. The agent neither moved nor spoke in this condition.

D. Questionnaires

Each participant evaluated the mind perception [4] (Japanese version of the questionnaire [11]) of the facing partner agent before the game. After the game, the participant kept seated and evaluated mind perception about the partner agent again. The question items of the mind perception questionnaire are listed in Table II. For example, the question about "memory" was "how much is the partner capable of memorizing something?"

The participant answered the following two questions about causal attribution of the outcome of the game to himself/herself or the partner agent after the game.

(a) Your reward was less than the average (480 yen) because of the choices of the partner.
(b) Your reward was less than the average (480 yen) because of your own choices.

We defined relative causal attribution as the value obtained by subtracting the score of (b) from the score of (a). This means how much the participant attributed the cause to the partner agent compared with to the himself/herself. All questionnaires were rated on a seven-point scale.

III. RESULT

A. Relative causal attribution

Fig. 3 shows relative causal attribution averaged over participants. A one-way ANOVA indicated a significant main effect of the agents ($p = 0.020$). There was not a significant main effect of the gender ($p = 0.64$). *Post hoc* paired tests (Bonferroni) revealed that there were the significant differences of relative causal attribution between the human

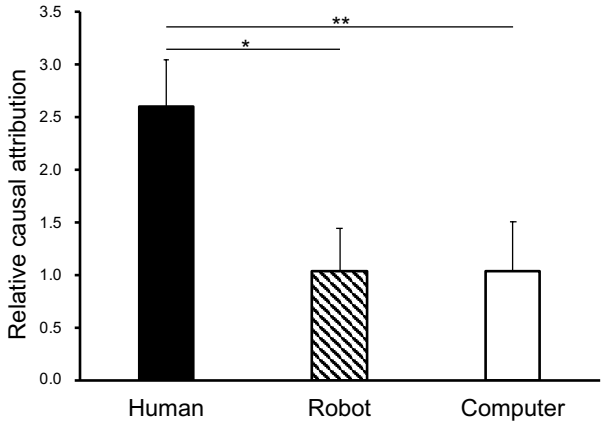


Fig. 3. Relative causal attribution. This value means the degree of causal attribution to the partner agent compared with self. Error bars indicate standard error. *: $p < 0.05$, **: $p < 0.01$.

and robot conditions ($p = 0.013$) and between the human and computer conditions ($p = 0.0080$). This suggests that the participants attributed more cause to the human agent than the robot or computer agent.

Relative causal attribution did not significantly correlate with the amounts of total rewards of participants ($r = -0.02$, $p = 0.85$) or the numbers of times "I want more" was chosen ($r = 0.07$, $p = 0.75$).

B. Two dimensions in mind perception

We extracted the dimensions of agency and experience using a factor analysis (maximum likelihood method, promax rotation) and analyzed correlations between those scores and relative causal attribution to verify our hypothesis that the degree of attribution is positively correlated with agency. Table III indicates the first and second factors of mind perception after the game, accounted for 44% and 26% of the variance, respectively. This clearly shows that they respectively correspond to experience and agency, as proposed by Gray et al. [4]. The square markers in Fig. 4 show the scores of agency and experience for each agent, which were averaged over participants. One-way ANOVAs indicate that a main effect of agents was significant in experience ($p \ll 0.01$); however, it was not significant in agency ($p = 0.14$). *Post hoc* paired tests (Bonferroni) elucidated that the experience score of the human was significantly greater than those of the robot and the computer. (both $p \ll 0.01$). A main effect of the gender was not significant in agency and experience ($p = 0.08$ and 0.81, respectively).

Further, we mapped mind perception before the game into the factor space as shown in Fig. 4. Paired *t*-test did not show any significant differences between the scores of agency and experience after and before the game in any agent conditions.

C. Correlations between mind perception and relative causal attribution

The score of agency after the game was significantly correlated with relative causal attribution ($r = -0.54$, $p \ll 0.01$)

TABLE III
FACTOR LOADINGS

	Experience	Agency
Rage	0.987	−0.092
Hunger	0.958	−0.325
Fear	0.929	−0.035
Pride	0.893	−0.031
Desire	0.856	−0.143
Consciousness	0.807	0.130
Pain	0.780	0.185
Personality	0.743	0.110
Pleasure	0.742	0.183
Embarrassment	0.738	0.150
Self-control	−0.123	0.954
Planning	−0.180	0.948
Thought	−0.081	0.863
Communication	0.253	0.686
Memory	−0.097	0.671
Morality	0.331	0.630
Emotion recognition	0.336	0.541

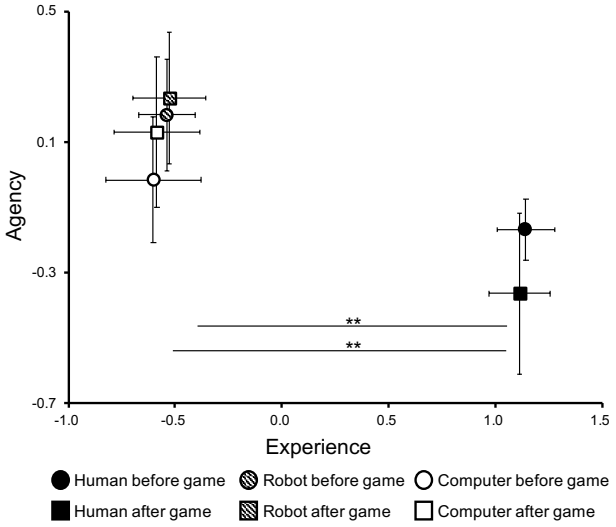


Fig. 4. Mind perception in two dimensions. The round and square markers indicate mind perception before and after the game, respectively. Error bars indicate standard error. **: $p < 0.01$.

TABLE IV
CORRELATION BETWEEN AGENCY/EXPERIENCE AND RELATIVE CAUSAL ATTRIBUTION.

	Agency					
	Before		After		Difference	
All conditions	−0.34	**	−0.54	**	−0.31	**
Human	0.16		−0.61	**	−0.67	**
Robot	−0.35	†	−0.38	†	−0.18	
Computer	−0.60	**	−0.51	**	0.07	
	Experience					
	Before		After		Difference	
All conditions	0.32	**	0.34	**	0.05	
Human	−0.01		0.19		0.18	
Robot	0.13		0.20		0.12	
Computer	0.21		0.14		−0.09	

(† : $p < 0.1$, *: $p < 0.05$, **: $p < 0.01$)

when all conditions were included. It is noted that this correlation was negative. Such negative correlations were also significant in the human and computer conditions ($r = -0.61$ and -0.51, respectively, and $p \ll 0.01$ in both conditions). In the robot condition, it was not significant although it approached significance ($r = -0.38$, $p = 0.06$). Therefore, the nagative correlation between agency and causal attribution might be common to the three agents. This is contrary to our hypothesis based on the study of Gray et al. [4].

In order to clarify the reason for the negative correlation of agency after the game, we investigated correlations between causal attribution and each item of mind perception after the game. A false discovery rate method was applied to the multiple tests. There were four items that showed a significant correlation with relative causal attribution: "self-control," "planning," "thought," and "memory." ($r = -0.43$, -0.39, -0.39, and -0.37, respectively). This suggests that agents whom the participant considered to hold these abilities tended not to be attributed the cause. These four items are strongly related to the agency, as shown in Table III, and therefore, they mainly contributed to the correlation between the agency and the relative causal attribution. In contrast, there were no items with a significant positive correlation with relative causal attribution.

The fourth column in Table IV indicates the correlation coefficients between relative causal attribution and differential agency and experience. The score of difference of agency was significantly correlated with relative causal attribution ($r = -0.31$, $p \ll 0.01$) when all conditions were included. When divided by the agents, the correlation was significant only in the human condition ($r = -0.67$, $p \ll 0.01$). In the robot and computer conditions, they were not correlated ($r = -0.18$ and 0.07, respectively).

We further analyzed the correlations between causal attribution and each item of difference of mind perception between before and after the game in the human condition. A false discovery rate method was applied to the multiple tests. We found significant correlations with some items: "thought," "communication," "pain," "self-control," "morality," "emotion recognition," "planning," "consciousness," "memory," and "fear." ($r = -0.74$, -0.68, -0.68, -0.63, -0.62, -0.59, -0.55, -0.46, -0.45, and -0.44, respectively). This suggests that in the human condition, the larger amounts of decreases in these items through the game were related to more causal attribution to the partner agent.

IV. Discussion

The experimental results show that the cause for failure was largely attributed to the human agent compared with the artificial agents: computer and robot (see Fig.3). We found that this causal attribution was related to specific mind perception about the partner agent and the changes of mind perception through the game.

The factor analysis yielded two dimensions that were similar to those in Gray et al. [4]: agency and experience. However, the agency score of the human agent tended to be lower than those of the artificial agents (see Fig. 4), which is opposite to the result of Gray et al. [4]. In contrast, the score of experience of the human agent was higher than those of the artificial agents, which is similar to the result of Gray et al. [4]. This discrepancy might originate from Japanese culture. Takahashi et al. [12] retested the study of Gray et al. [4] with Japanese participants and reported a result similar to our result. They hypothesized that Japanese people perceived artificial agents to have strong agency because of their animistic thinking, i.e., the tendency to believe minds of non-living things [13].

The relationship between mind perception and causal attribution was also different from our hypothesis based on the result of Gray et al. [4]. In our experiment, agency and Experiment had negative and positive correlations with causal attribution, respectively. Specifically, four items of mind perception ("memory," "self-control," "planning," and "thought"), which are sub-items of agency, were negatively correlated with relative causal attribution after the game. These items seem to be important abilities to obtain many rewards in the game. The abilities of "memory," "planning," and "thought" are related to determining subsequent choices from previous choices. In addition, to give over money to the partner requires the ability of "self-control." Therefore, it is speculated that the agent with the high scores of these items might give participants the impression that "the agent acted for the success of the game.", which might lead to low causal attribution. This relation might be applied to all types of agents. In contrast, the correlations between all sub-items of experience and relative causal attribution were not significant. Nevertheless, we speculated that the participants might recognize selfish choices of the agents with the abilities to experience emotions as intentional behavior based on their emotions and desires, which might lead to high causal attribution to the agents.

Our second hypothesis that a decrease in mind perception through the game contributes to an increase in causal attribution to the agent was accepted only in the human condition. The decrease in agency items as well as the items of "consciousness," "pain" and "fear" was negatively correlated with causal attribution. Humans whose such impressions became worse than participants' prior expectations were attributed the cause more greatly. agency score decreased especially in the human condition (see Fig. 4); In general, humans are expected to make considered choices and to feel the pain of others, i.e., based on empathy for others. This expectation was betrayed in the game, which might lead to causal attribution. Adaptation gap hypothesis states that the difference between users' expectation and actual performance about an agent strongly affects the impression (e.g., likeness) of the agent [14], [15]. The impression of an agent becomes favorable if it function is more than expected, and vice versa. The result of the current study suggests that an unfavorable impression of the human agent because of the gap in term of considered and moral behavior might lead to large causal attribution to the agent.

In contrast, artificial agents might not be expected to behave morally based on empathy. Malle et al. [16] reported

that different moral norms are applied to humans and robots. Utilitarian and moral judgments are expected for robots and humans in a moral dilemma task, respectively [16]. In our experiment, the selfish choices by artificial agents might be more acceptable than those by the human agent, which might cause the weak correlations. Artificial agents might be attributed a cause of a failure or accident if they were recognized as moral agents. Our approach may offer an experimental evidence to the philosophical discussion about artificial moral agency [17], [18].

The noncooperative game was used in this study because we suppose human-robot economical interactions, e.g., a robot in a shop, in which agents have their own intentions. In addition, the situations requiring explicit cooperation between a human and robot can be modeled as cooperative games. Such situations can be applied to not only decision making with a machine but also to shared control systems [19]. For example, a semi-automatic driving system needs cooperation with a human driver to control a car (e.g., [20]). A wearable power assist robot and its user share the control of the user's body (e.g., [21]). How users attribute a cause and responsibility for an unexpected outcome, i.e., accident, is an important issue. Vilaza and Haselager [10] reported that the user's sense of agency, which is the subjective awareness of causing and generating an action, in a shared control system depends on the outcome of a task. When the task fails, the user tends to report that the system is the cause of the result [10], which is explained by the self-serving bias [3]. We plan to investigate the relationship between mind perception and causal attribution in such systems and tasks to verify whether the results in the current study can be applied to them.

V. Conclusion

We investigated the relationship between mind perception and causal attribution for failing to obtain the expected payoff in the game in which reward was given by a decision of the participant and a partner agent. As a result, it was shown that the hypothesis of Gray et al. [4] was not necessarily true when evaluating the causal attribution to the partner agent from a viewpoint of a party of actual damage by interaction. Rather, regardless of the type of agent, agency has a negative correlation with causal attribution. In particular, it was suggested that the important items for the success of the game showed a negative correlation with causal attribution, and the impression of acting to the success of the game decrease causal attribution. Furthermore, it was found that when agent's mind perception was less than expected, the more cause was attributed. It was suggested that its effect varies depending on the type of agent, and its effect is greater in humans than in a robot.

The results showed that causal attribution to the partner agent correlated with the subjective mind perception about the agent and a gap from prior expectation rather than the objective monetary damage. This suggests that a designer should consider not only the task performance of a robot but also the user's mind perception about the robot.

References

[1] F. Heider, *The Psychology of Interpersonal Relations*. New York: Wiley, 1958.

[2] K. G. Shaver, *The Attribution of Blame: Causality, Responsibility, and Blameworthiness*. New York: Springer-Verlag, 1985.

[3] D. T. Miller and M. Ross, "Self-serving biases in the attribution of causality: Fact or fiction?" *Psychological Bulletin*, vol. 82, no. 2, pp. 213–225, 1975.

[4] H. M. Gray, K. Gray, and D. M. Wegner, "Dimensions of mind perception," *Science*, vol. 315, no. 5812, p. 619, 2007.

[5] P. J. Hinds, T. L. Roberts, and H. Jones, "Whose job is it anyway? a study of human-robot interaction in a collaborative task," *Human-Computer Interaction*, vol. 19, no. 1, pp. 151–181, 2004.

[6] T. Kim and P. Hinds, "Who should I blame? effects of autonomy and transparency on attributions in human-robot interaction," in *Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication*, 2006, pp. 80–85.

[7] Y. Moon and C. Nass, "Are computers scapegoats? Attributions of responsibility in human-computer interaction," *International Journal of Human-Computer Studies*, vol. 49, no. 1, pp. 79–94, 1998.

[8] K. L. Koay, D. S. Syrdal, M. L. Walters, and K. Dautenhahn, "Five weeks in the robot house–exploratory human-robot interaction trials in a domestic setting," in *the 2nd International Conferences on Advances in Computer-Human Interactions*, 2009, pp. 219–226.

[9] S. You, J. Nie, K. Suh, and S. S. Sundar, "When the robot criticizes you...: self-serving bias in human-robot interaction," in *Proceedings of the ACM/IEEE 6th International Conference on Human-Robot Interaction*, 2011, pp. 295–296.

[10] G. N. Vilaza, W. Haselager, A. Campos, and L. Vuurpijl, "Using games to investigate sense of agency and attribution of responsibility," *in Proceedings of the XIII Brazilian Symposium on Computer Games and Digital Entertainment*, 2014.

[11] H. Kamide, K. Takahashi, and T. Arai, "Development of Japanese version of the psychological scale of anthropomorphism," *Japanese Journal of Personality*, vol. 25, no. 3, pp. 218–225, 2017.

[12] H. Takahashi, M. Ban, and M. Asada, "Semantic differential scale method can reveal multi-dimensional aspects of mind perception," *Frontiers in Psychology*, vol. 7, 2016.

[13] G. Harvey, *Animism: Respecting the Living World*. Adelaide, SA: Wakefield Press, 2005.

[14] T. Komatsu and S. Yamada, "Adaptation gap hypothesis: How differences between users' expected and perceived agent functions affect their subjective impression," *Journal of Systemics, Cybernetics and Informatics*, vol. 9, no. 1, pp. 67–74, 2011.

[15] T. Komatsu, R. Kurosawa, and S. Yamada, "How does the difference between users ' expectations and perceptions about a robotic agent affect their behavior?" *International Journal of Social Robotics*, vol. 4, no. 2, pp. 109–116, 2012.

[16] B. F. Malle, M. Scheutz, T. Arnold, J. Voiklis, and C. Cusimano, "Sacrifice one for the good of many?: People apply different moral norms to human and robot agents," in *Proceedings of the 10th annual ACM/IEEE International Conference on Human-Robot Interaction*, 2015, pp. 117–124.

[17] L. Floridi and J. W. Sanders, "On the morality of artificial agents," *Minds and machines*, vol. 14, no. 3, pp. 349–379, 2004.

[18] W. Wallach and C. Allen, *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press, 2008.

[19] D. P. Losey, C. G. McDonald, E. Battaglia, and M. K. O'Malley, "A review of intent detection, arbitration, and communication aspects of shared control for physical human–robot interaction," *Applied Mechanics Reviews*, vol. 70, no. 1, pp. 010 804–1–19, 2018.

[20] P. G. Griffiths and R. B. Gillespie, "Sharing control between humans and automation using haptic interface: primary and secondary task performance benefits," *Human Factors*, vol. 47, no. 3, pp. 574–590, 2005.

[21] K. Kiguchi and Y. Hayashi, "An EMG-based control for an upper-limb power-assist exoskeleton robot," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, pp. 1064–1071, 2012.