Smart Human, Smarter Robot: How Cheating Affects Perceptions of Social Agency

Daniel Ullman[1], Iolanda Leite[2], Jonathan Phillips[1,3,4], Julia Kim-Cohen[3], and Brian Scassellati[1,2]
[1]Program in Cognitive Science | [2]Department of Computer Science | [3]Department of Psychology | [4]Department of Philosophy
Yale University, New Haven, CT 06520 USA

Abstract

Human-robot interaction studies and human-human interaction studies often obtain similar findings. When manipulating high-level apparent cognitive cues in robots, however, this is not always the case. We investigated to what extent the type of agent (human or robot) and the type of behavior (honest or dishonest) affected perceived features of agency and trustworthiness in the context of a competitive game. We predicted that the human and robot in the dishonest manipulation would receive lower attributions of trustworthiness than the human and robot in the honest manipulation, and that the robot would be perceived as less intelligent and intentional than the human overall. The human and robot in the dishonest manipulation received lower attributions of trustworthiness as predicted, but, surprisingly, the robot was perceived to be more intelligent than the human.

Keywords: social robotics; trustworthiness; intelligence; intentionality; agency; human-robot interaction

Introduction

The importance of recognizing social agentic features is not confined to humans, but extends to other living beings and to nonliving social agents. Inferences about the behavior and cognitive capabilities of an entity greatly influence ascriptions of intelligence (Beer, 1990). Human-like properties related to intelligence can be attributed to animated shapes (Scholl & Tremoulet, 2000), virtual agents (Bickmore & Cassell, 2001), and social robots (Bainbridge, Hart, Kim, & Scassellati, 2011; Short, Hart, Vu, & Scassellati, 2010). While the concept of intelligence has been studied extensively with respect to humans, the properties that contribute to perceptions of other animated beings as intelligent, in particular social robots, are still unclear. A better understanding of how people make social attributions to robots will not only allow roboticists to design robots with better social interactive capabilities, but also will add to the knowledge base on features of social agency.

Previous research by Short et al. (2010) showed that manipulating high-level behavioral cues, specifically cheating versus not cheating, causes attributions of different mental states to a robot. The researchers investigated attributions of mental state and intentionality to a cheating robot in a game of rock-paper-scissors, a high-level examination that explored how variations in robotic behavior affected perceptions of a robot's agency. Participants in the two cheat conditions rated the interaction as less fair and honest than those in the third condition, the no cheat condition. Furthermore, the results pointed toward greater attributions of mental state to the robot in the cheat conditions than in the no cheat condition. The work by Short et al. (2010) directly motivates the present research. We seek to further this line of research by benchmarking an analysis of agentic cues in a human-robot interaction against a comparable analysis of agentic cues in a human-human interaction. Ultimately, we aim to examine perceptions of intelligence and intentionality in a context of cheating behavior.

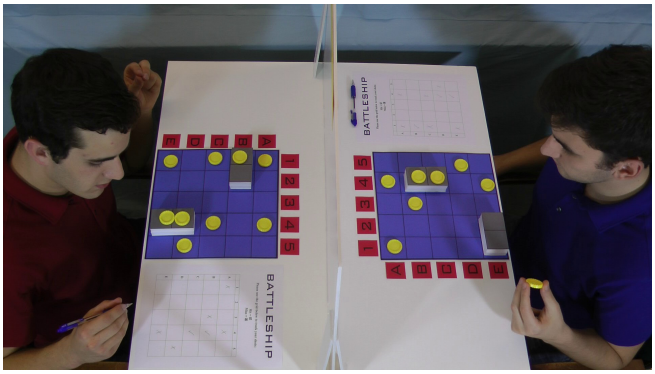

Figure 1. Snapshot of the human manipulation.

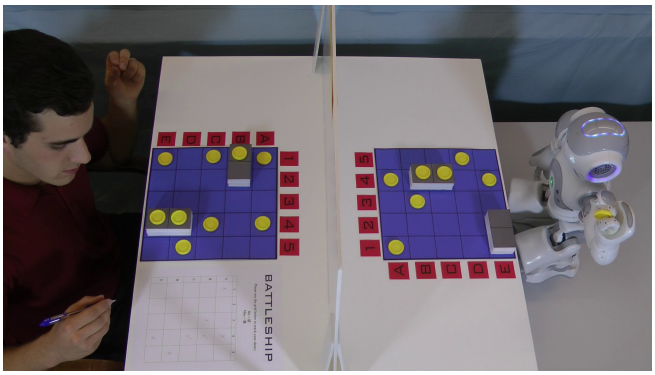

Figure 2. Snapshot of the robot manipulation.

There are a number of factors that contribute to perceptions of entities as agentic. As stated by Bandura (2001), "To be an agent is to intentionally make things happen by one's actions." Researchers over the years have identified features important to ascriptions of agency, including intentionality (Bandura, 2001) and self-propelled, purposeful-looking movement (Premack, 1990; Scholl & Tremoulet, 2000). The concept of agency extends beyond humans; as argued by Takayama (2011), "Regardless of the

absolute status of an entity's agency, it is our *perceptions* of agency that influence how we behave." Several studies have shown that nonhuman entities that act in ways that appear to be goal-directed are likely to be perceived as agentic. For example, a study by Scholl and Tremoulet (2000) showed that goal-directed motion exhibited by small shapes moving around a visual field caused humans to attribute features of animacy and causality to the shapes. Research into the role of goal-directed action in robots has found that infants positively attribute goals to humanoid-robot motion (Kamewari, Kato, Kanda, Ishiguro, & Hiraki, 2005). In a study that manipulated the physical presence of a robot to understand the effect of the robot's embodiment on attributions of goal-directed behavior, Bainbridge et al. (2011) found that participants were more willing to comply with a physically present robot than a robot displayed via a live video feed. Overall, intentionality is a feature that can be ascribed to agents that display sufficient cues of agency (Mutlu, Yamaoka, Kanda, Ishiguro, & Hagita, 2009). Our beliefs about what an agent can do, and how an agent should act, greatly affect our perceptions of an agent.

Cheating is a classification of intentional behavior that can be attributed to social agents. Perceptions of trustworthiness are linked to perceptions of intelligence (Goleman, 1995) and affect how willing one agent is to interact with another, as well as how one agent actually interacts with another. Research has shown that cheating affects trustworthiness, an important feature of social relationships (Rotter, 1980). Robots are becoming increasingly present in contexts that demand relationships built on trust, from robots that deliver medicine in hospitals and robots that provide company for the elderly (Broadbent, Stafford, & MacDonald, 2009; Zhang et al., 2010), to robots that team up with workers on factory lines (Desai et al., 2012). With the potential to be social agents, robots thus also have the potential to be perceived as trustworthy, or even untrustworthy (Vazquez, May, Steinfeld, & Chen, 2011). In fact, robots appear to be held accountable for moral harm that they cause (Kahn et al., 2012).

As robots become further integrated into social situations, it is important to understand how robotic behavior can affect the perception of socially relevant traits. In this paper, we investigate whether manipulating the behavior of an agent in situations involving trust affects perceptions of agentic features of a robot similarly to how it affects perceptions of a human. Specifically, we evaluate how manipulating the type of behavior displayed by a human in a competitive game affects attributions of trustworthiness, intelligence, and intentionality to the human (Figure 1), as compared to identical manipulations of the behavior displayed by a robot (Figure 2). We posited the following hypotheses:

H1: The human and robot in the dishonest manipulation will receive lower attributions of trustworthiness than the human and robot in the honest manipulation.

H2: The robot will be perceived as less intelligent and intentional than the human in both the honest and dishonest manipulations.

Hypothesis H1 stems from the expectation that humans who cheat are perceived to be less trustworthy than humans who do not cheat, and is motivated by findings from Short et al. (2010) that indicate that robots that cheat are perceived as less fair and honest than robots that do not cheat. Hypothesis H2 stems from the expectation that robots display fewer features associated with agency than do humans, resulting in lower attributions of associated features of intelligence and intentionality to robots than humans.

Method

Each participant watched two videos in which two agents, a human and a human or a human and a robot, were playing a modified version of the board game Battleship (Figure 1 and Figure 2). In Battleship, two players sit facing each other with a visual divider in-between the players' ocean grids. The divider hides the opponent's ship locations. The objective of the game is to be the first player to sink the opponent's ships by calling out shot locations.

Materials

The experimental setup replicated the game of Battleship, with the game and experimental setup modified to accommodate the physical capabilities of the robot. We used the robot Nao V3.2, a humanoid robot from Aldebaran Robotics (pictured on the right in Figure 2). We recorded four video clips of in-progress games of Battleship, one for each condition. The humans in the videos tracked game progress using a sheet with a grid, while participants were told that the robot would track shots using its memory.

Procedure

The study employed a mixed design, with participants randomly assigned to one of four conditions. The first independent variable, presented between subjects, was player type: *human* or *robot*. The second independent variable, presented within subjects, was behavior type: *honest*, the absence of cheating, or *dishonest*, the presence of cheating.

Participants were presented with the rules of the game they were going to observe, which explicitly stated: "Do not change the position of any ship once the game has begun." The cheat behavior consisted of the human or robot cheater moving a ship out of the line of fire and categorizing a shot as a miss instead of a hit, directly violating the stated game rules. Each participant was presented with both the honest and dishonest videos of the player type they were randomly assigned. The order of the videos was counterbalanced so that half of the participants viewed the honest video first and half viewed the dishonest video first. David was the name given to the human opponent on the left, while Kevin was the name given to the human or robot opponent on the right.

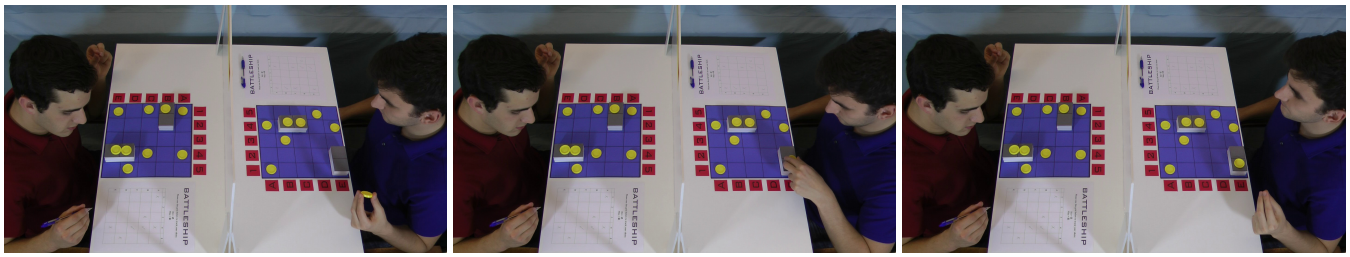

Figure 3. Series of snapshots of the honest human condition.
Human-human interaction with human on the right placing piece, following game rules.

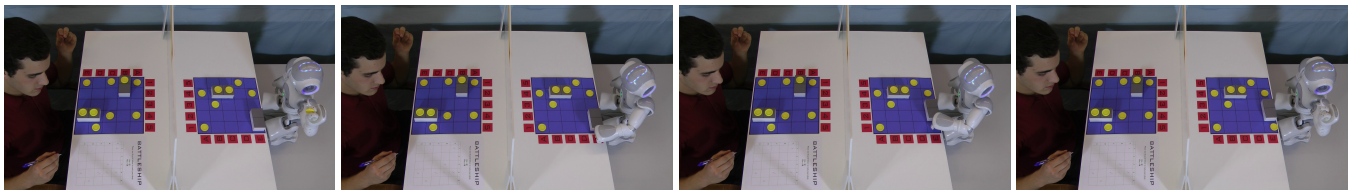

Figure 4. Series of snapshots of the dishonest robot condition.
Human-robot interaction with robot moving ship and then placing piece, violating game rules.

For each video, participants were told that David, on the left, is playing against Kevin, on the right, and that it is Kevin's turn. The game proceeded as follows: Kevin took a turn, David took a turn and Kevin either did not cheat (honest manipulation) or did cheat (dishonest manipulation), and then Kevin took a final turn and won the game. For each condition, participants were presented with the first video, answered survey questions for the first video, were presented with the second video, and answered the same set of survey questions for the second video. Participants were then asked to answer optional demographic questions.

Figure 3 shows a series of snapshots of the honest human condition, portraying the no cheating behavior. Figure 4 shows a series of snapshots of the dishonest robot condition, portraying the cheating behavior.

Participants

A total of 200 adults (137 male, 63 female) participated in the study. The mean age of participants was 31.24 years (*SD* = 10.60). Participants reported race/ethnicity as follows: 153 "Non-Hispanic White or Euro-American"; 14 "Black, Afro-Caribbean, or African American"; 4 "Latino or Hispanic American"; 16 "Asian or Asian American"; 2 "Native American or Alaskan Native"; 6 "Multi-racial"; 0 "Other"; 5 "I would prefer not to answer."

We recruited participants via the web-based resource Amazon Mechanical Turk. Participants were directed to a survey designed using the web-based resource Qualtrics. To ensure that workers completed the survey, participants received an end of survey completion code in order to receive payment through Mechanical Turk.

Participants were recruited through Amazon Mechanical Turk with the following restrictions in place: having an overall approval rating > 95% and being geographically located in America (determined by IP addresses). Participants were excluded if they self-reported experiencing technical trouble watching or hearing either video, if they reported being a non-native English speaker, or if they reported a 4 or below on a 7-point Likert item concerning how well they understood the rules of the game they were observing. Participants were also excluded if they failed to correctly answer a control question on the number of humans present in each video. Participants' IP addresses were checked to ensure that there were no repeat participants; none were found. After exclusions, a total of 179 participants remained: 87 in the human manipulation and 92 in the robot manipulation. Participants were paid $0.50 each.

Measures

Participants were presented with the prompt "How would you rate Kevin in terms of the following:" and these 7-point Likert items: Intelligence, Cleverness, Intentionality, Fairness, Honesty, Trustworthiness.

Trustworthiness Data for the dependent variable of trustworthiness were computed by combining three ratings on fairness, honesty, and trustworthiness, with internal consistency for the honest conditions (Cronbach's $\alpha = .93$, $n = 3$) and the dishonest conditions (Cronbach's $\alpha = .96$, $n = 3$).

Intelligence Data for the dependent variable of intelligence were computed by combining two ratings on intelligence and cleverness, with internal consistency for the honest conditions (Cronbach's $\alpha = .76$, $n = 2$) and the dishonest conditions (Cronbach's $\alpha = .82$, $n = 2$).

Intentionality Data for the dependent variable of intentionality were taken from the one rating on intentionality.

Results

Three 2 x 2 ANOVAs were conducted. All ANOVAs compared the two independent variables of player type (human or robot) and behavior type (honest or dishonest).

Trustworthiness

A 2 x 2 ANOVA was conducted to investigate the impact of player type and behavior type on perceived trustworthiness (Figure 5). There was a significant main effect of player type, with the human conditions (M = 3.63, SE = 0.12) rated lower than the robot conditions (M = 3.99, SE = 0.12), $F(1, 177) = 4.70$, $p = .03$, $\eta_p^2 = .03$. There was also a significant main effect of behavior type, with the honest conditions (M = 5.50, SE = 0.11) rated higher than the dishonest conditions (M = 2.12, SE = 0.13), $F(1, 177) = 433.90$, $p < .001$, $\eta_p^2 = .71$. There was no significant interaction effect between player type and behavior type, $F(1, 177) = 0.52$, $p = .47$, $\eta_p^2 = .00$. Participants' perception of trustworthiness was greater for the human in the honest condition (M = 5.38, SE = 0.15) than in the dishonest condition (M = 1.88, SE = 0.18), and participants' perception of trustworthiness was greater for the robot in the honest condition (M = 5.63, SE = 0.15) than in the dishonest condition (M = 2.36, SE = 0.18).

Intelligence

A 2 x 2 ANOVA was conducted to investigate the impact of player type and behavior type on perceived intelligence (Figure 6). There was a significant main effect of player type, with the human conditions (M = 4.58, SE = 0.12) rated lower than the robot conditions (M = 5.09, SE = 0.12), $F(1, 177) = 8.89$, $p < .01$, $\eta_p^2 = .05$. There was no significant main effect of behavior type, with no statistically significant difference between the honest conditions (M = 4.82, SE = 0.09) and the dishonest conditions (M = 4.85, SE = 0.11), $F(1, 177) = 0.12$, $p = .73$, $\eta_p^2 = .00$. There was a significant interaction effect between player type and behavior type, $F(1, 177) = 6.63$, $p = .01$, $\eta_p^2 = .04$. Participants' perception of intelligence was greater for the human in the honest condition (M = 4.68, SE = 0.12) than in the dishonest condition (M = 4.48, SE = 0.15), whereas participants' perception of intelligence was greater for the robot in the dishonest condition (M = 5.22, SE = 0.15) than in the honest condition (M = 4.95, SE = 0.12).

Intentionality

A 2 x 2 ANOVA was conducted to investigate the impact of player type and behavior type on perceived intentionality (Figure 7). There was no significant main effect of player type, with no statistically significant difference between the human conditions (M = 5.41, SE = 0.15) and the robot conditions (M = 5.10, SE = 0.15), $F(1, 177) = 2.17$, $p = .14$, $\eta_p^2 = .01$. There was a main effect of behavior type, with the honest conditions (M = 4.97, SE = 0.11) rated lower than the dishonest conditions (M = 5.55, SE = 0.13), $F(1, 177) = 23.93$, $p < .001$, $\eta_p^2 = .12$. There was a marginally significant interaction effect between player type and behavior type, $F(1, 177) = 2.85$, $p = .09$, $\eta_p^2 = .02$. Participants' perception of intentionality was much greater for the human in the dishonest condition (M = 5.81, SE =

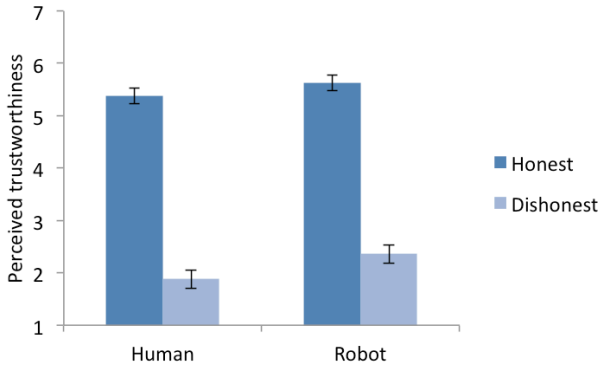

Figure 5. Mean perceptions of trustworthiness. Error bars show SE mean.

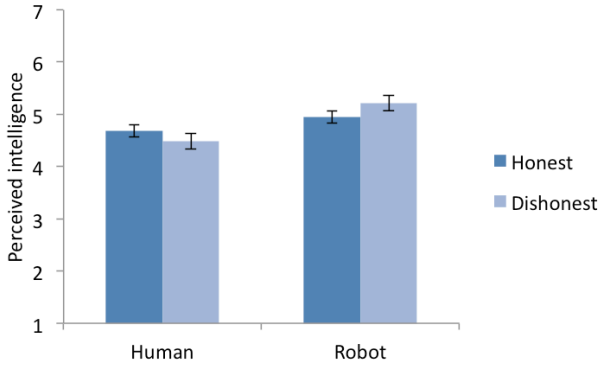

Figure 6. Mean perceptions of intelligence. Error bars show SE mean.

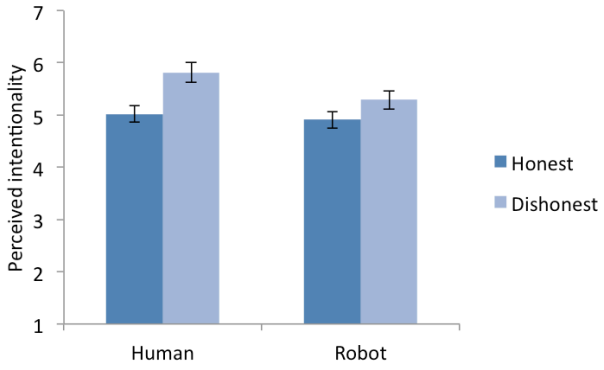

Figure 7. Mean perceptions of intentionality. Error bars show SE mean.

0.19) than in the honest condition (M = 5.02, SE = 0.16), whereas participants' perception of intentionality was only slightly greater for the robot in the dishonest condition (M = 5.29, SE = 0.18) than in the honest condition (M = 4.91, SE = 0.16).

Discussion

The examination of trustworthiness provides the most straightforward results, supporting our first hypothesis. Hypothesis H1 posited that participants would make lower attributions of trustworthiness to the human and robot in the dishonest manipulation than in the honest manipulation. There was indeed a large, significant main effect of behavior type on trustworthiness, such that the human and robot received lower attributions of trustworthiness in the dishonest manipulation than in the honest manipulation. This is both an expected and logical finding, especially given the comparable results on fairness and honesty for a robot obtained by Short et al. (2010). There was an additional small, significant main effect of player type on trustworthiness, such that the robot was perceived to be more trustworthy than the human. A possible explanation for this effect may be that participants rated the human and robot relative to their experiences of a typical human and robot, with the findings suggesting that participants perceived the robot to be slightly more trustworthy compared to their experience of a typical robot. As most people do not often interact with robots, especially not in contexts involving honest or dishonest behaviors, this might have contributed to slightly inflated attributions in the robot conditions. An additional possible explanation is that while people may readily infer that a person is, in general, untrustworthy, they may be less willing to infer that a robot is untrustworthy from a single interaction.

Hypothesis H2 posited that participants would perceive the robot as less intelligent and intentional than the human. This hypothesis was not supported. Rather, the results suggested a different interpretation, and warranted separately considering intelligence and intentionality. We first analyzed the results from intelligence. There was indeed a significant main effect of player type, however it was in the opposite direction of the prediction; participants rated the robot higher on intelligence than the human. At first glance, this finding seems to contradict expectations suggested by Short et al. (2010). However, upon closer examination, this result in fact seems to parallel the effect of player type on trustworthiness explained earlier; that is, the agent rated as more intelligent, the robot, was also rated as more trustworthy.

While there was no significant main effect of behavior type on intelligence, there was a significant interaction effect between player type and behavior type. The interaction effect was a crossover interaction, such that participants' perception of intelligence was lower for the human in the dishonest condition than in the honest condition, whereas participants' perception of intelligence was greater for the robot in the dishonest condition than in the honest condition. In light of the possibility that participants rated the human and robot relative to their experiences of a typical human and robot, the interaction effect appears logical. People perceived the robot as more intelligent when it was dishonest, while the human was rated as less intelligent when dishonest; these ratings may partially stem from the actual cheat behavior itself, which participants might have considered an intelligent behavior for a robot, but an unintelligent behavior for a human. Participants likely considered the cheat behavior of the robot a novel, surprising behavior for a robot, adding to the perceived intelligence of the robot. It might be possible to tease out such a novelty effect in a future repeated interactions study.

As for intentionality, there was no significant main effect of player type on perceived intentionality, but there was a significant main effect of behavior type. The significant main effect of behavior type seems to align with findings associated with the side-effect effect (Knobe, 2003), such that participants attributed greater intentionality when the agent performed an immoral action than when the agent performed a morally neutral action. It is interesting to note that the dishonest robot was rated as less intentional than the dishonest human, and was not rated as low as the human on trustworthiness; that is, it appears participants held the robot less accountable for its cheating behavior than the human. There was a marginal interaction effect, p = .09, such that participants' perception of intentionality was much greater for the human in the dishonest condition than in the honest condition, whereas participants' perception of intentionality was only slightly greater for the robot in the dishonest condition than in the honest condition. It is possible, however, that the item on intentionality was confusing, as the results do not align with previous research that implicates perceptions of intentionality as correlated with other perceptions of agentic behavior (Premack, 1990; Scholl & Tremoulet, 2000). Alternatively, it is also possible that this is a novel finding and will incite future research into a potential separation of perceived intelligence and intentionality in social robotics. Further research is ultimately required to tease out such an interaction effect and to better understand the results indicated by the item on intentionality.

Conclusion

As robots become increasingly present in everyday settings, and especially as they take on roles that necessitate greater social interaction with humans, the field of social robotics requires a more thorough understanding of the features that influence people's perceptions of robots. This trend necessitates research into how humans perceive robots' social traits, as well as to what extent robotic behavior affects attributions of agency.

In this paper, we investigated the extent to which the type of agent (human or robot) and the type of behavior (honest or dishonest) affected perceptions of trustworthiness, intelligence, and intentionality in the context of a

competitive game. The results on trustworthiness were expected, while interpretation of the results on intelligence and intentionality yielded unexpected but intriguing findings. Participants' perceptions of intelligence trended in opposite directions for the human and robot; the robot was perceived as more intelligent in the dishonest manipulation, while the human was perceived as less intelligent in the dishonest manipulation. An interesting implication of the study and a potential for follow-up research concerns the results on intentionality. The marginal interaction effect for intentionality suggests that attributions of intentionality to agents are affected by the type of agent (human or robot) in combination with the type of behavior (honest or dishonest); future work focused on the role of intentionality in this paradigm will hopefully better illuminate the observed effect.

Most human-robot interaction studies obtain results in the same direction as human-human interaction studies. Our findings, with respect to perceived intelligence, did not align with this trend. Rather, our results suggest that the behavior of a robot not only affects to what extent it is perceived as an intelligent agent, similarly to a human, but also demonstrate that individuals perceive a robot differently from a comparably performing human. These results indicate that robot designers cannot simply transfer findings from social human-human interaction to human-robot interaction, but instead must further investigate features that affect human-robot interaction in their own right.

Acknowledgments

The authors gratefully acknowledge Alex Litoiu for help recording the stimuli, Henny Admoni and Elena Corina Grigore for feedback on drafts of this paper, and anonymous reviewers for their insightful comments. This material is based upon work supported by the National Science Foundation award #1117801 (Manipulating Perceptions of Robot Agency) and award #1139078 (Socially Assistive Robots).

References

Bainbridge, W. A., Hart, J. W., Kim, E. S., & Scassellati, B. (2011). The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics*, *3*, 41-52.

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual Review of Psychology*, *52*, 1-26.

Beer, R. D. (1990). *Intelligence as adaptive behavior: An experiment in computational neuroethology*. San Diego, CA: Academic Press.

Bickmore, T., & Cassell, J. (2001). Relational agents: A model and implementation of building user trust. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 396-403.

Broadbent, E., Stafford, R., & MacDonald, B. (2009). Acceptance of healthcare robots for the older population: Review and future directions. *International Journal of Social Robotics*, *1*, 319-330.

Desai, M., Medvedev, M., Vazquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., … & Yanco, H. (2012). Effects of changing reliability on trust of robot systems. *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction*, 73-80.

Goleman, D. (1995). *Emotional intelligence: Why it can matter more than IQ*. New York, NY: Bantam Books.

Kahn, P. H., Jr., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., … & Severson, R. L. (2012). Do people hold a humanoid robot morally accountable for the harm it causes?. *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction*, 33-40.

Kamewari, K., Kato, M., Kanda, T., Ishiguro, H., & Hiraki, K. (2005). Six-and-a-half-month-old children positively attribute goals to human action and to humanoid-robot motion. *Cognitive Development*, *20*, 303-320.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*, 190-193.

Mutlu, B., Yamaoka, F., Kanda, T., Ishiguro, H., & Hagita, N. (2009). Nonverbal leakage in robots: Communication of intentions through seemingly unintentional behavior. *Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction*, 69-76.

Premack, D. (1990). The infant's theory of self-propelled objects. *Cognition*, *36*, 1-16.

Rotter, J. B. (1980). Interpersonal trust, trustworthiness, and gullibility. *American Psychologist*, *35*, 1-7.

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4*, 299-309.

Short, E., Hart, J., Vu, M., & Scassellati, B. (2010). No fair!!: An interaction with a cheating robot. *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction*, 219-226.

Takayama, L. (2011). Perspectives on agency: Interacting with and through personal robots. *In Zacarias, M. & Oliveira, J. V. (Eds.), Human-Computer Interaction: The Agency Perspective*. Berlin: Springer.

Vazquez, M., May, A., Steinfeld, A., & Chen, W. H. (2011). A deceptive robot referee in a multiplayer gaming environment. *International Conference on Collaboration Technologies and Systems*, 204-211.

Zhang, T., Kaber, D. B., Zhu, B., Swangnetr, M., Mosaly, P., & Hodge, L. (2010). Service robot feature design effects on user perceptions and emotional responses. *Intelligent Service Robotics*, *3*, 73-88.