No Fair!!
An Interaction with a Cheating Robot

Elaine Short, Justin Hart, Michelle Vu, Brian Scassellati
Department of Computer Science
Yale University
New Haven, CT 06511
elaine.short@yale.edu, justin.hart@yale.edu, michelle.vu@yale.edu, scaz@cs.yale.edu

***Abstract*—Using a humanoid robot and a simple children's game, we examine the degree to which variations in behavior result in attributions of mental state and intentionality. Participants play the well-known children's game "rock-paper-scissors" against a robot that either plays fairly, or that cheats in one of two ways. In the "verbal cheat" condition, the robot announces the wrong outcome on several rounds which it loses, declaring itself the winner. In the "action cheat" condition, the robot changes its gesture after seeing its opponent's play. We find that participants display a greater level of social engagement and make greater attributions of mental state when playing against the robot in the conditions in which it cheats.**

***Index Terms*—Affective & emotional responses, Beliefs about robots, Mental models of robot behavior**

I. Introduction

When people play games with each other, they do not merely go through the motions of game play. The intrigue of poker is in bluffing, fooling your opponent so that they fall into your trap. This dynamic can be seen in the actions taken in the game, and it is reflective of the thought processes behind game play. Bandura states, "to be an agent is to intentionally make things happen by one's actions" [1]. In this experiment, we model this very human aspect of games in a humanoid robot by having it cheat during a simple children's game. By setting up the task in this way, we are seeking to make participants treat the robot not as a device mechanically running through the steps of a program, but as an entity with underlying desires and motivations. That is, we hope to have them attribute mental state to the robot and treat it like an agent performing a task rather than a device passively processing input. To the participant, it is obvious that the robot has failed to play by the rules when it cheats. With this behavior we hope to create the perception that it wants to win. Is this simple manipulation sufficient to increase attributions of mental state? Will participants find the interaction more human-like and engaging? What if its behavior might easily be confused with a malfunction? Is the unexpected behavior sufficient to produce these effects?

Many social robots have been designed to foster attributions of intentionality as a mechanism to promote more engaging human-robot interactions (i.e. [2]). A simple social cue related to intentionality is used in [3], in which the authors set up a guessing game where participants ask a robot "yes/no" questions in order to find a chosen object from a set of objects laid out on a table. During play, the robot briefly glances at the object that it has selected. This "nonverbal leakage" is a subtle expression of mental state on the part of the robot. Proto-social responses have also been used to increase the human participants' feeling of engagement, as well as the length of time they fix their attention to a robot [4]. Breazeal [5] uses affective communication to create engagement between naïve participants and a humanoid robot. Mental state attribution may also play a role in studies that observe other phenomena. In their study of humans teaching robots, Kim, Leyzberg, Tsui, and Scassellati found that participants speak more to a robot that has trouble performing a learning task [6].

Simple experiments involving the motion of shapes across a computer screen or stage have also been used to probe questions of intentionality. Early work demonstrated that simple shapes moving across a screen will be treated as animate, provided they move in a way which allows such an interpretation [7]–[9]. Premack suggests that infants perceive the motion of blocks as intentional when they appear to move independently of external stimuli. The infants show surprise when an "intentionally moving" shape does not demonstrate preference for shapes of its own kind [10]. Infants even read morality into the actions of blocks that either "help" or "hinder" another block and preferentially look at a "helping" block over a "hindering" block [11]. These attributions are based on the actions that the blocks take, and are related to perceptions of mental state on the part of the infant. The infant perceives that one block is trying to impede the other and prefers the helpful one.

Games are a popular medium for human robot interaction [3], [12], [13]. Kahn et al. [14] have proposed a number of design patterns for social human robot interaction, including "reciprocal turn-taking in game context". In this work, we use a similar pattern, except that players' actions during rounds of this game occur simultaneously. Participants play the children's game *rock-paper-scissors* with a humanoid robot. Although the participant assumes the robot will play normally, we break from this expected paradigm of behavior. In some cases, the robot makes an ambiguous error which could be perceived as either cheating or a malfunction, in others it clearly cheats. We observe the effects of this manipulation by observing the video-recorded behavior of the participants and administering a post-interaction survey. Our research is guided by the following questions: Will participants find the clearly cheating robot to be more engaging? Will participants make

978-1-4244-4893-7/10/$25.00 © 2010 IEEE

greater attributions of mental state to it? Is any departure from the expected paradigm sufficient to produce these results?

II. Methodology

We devised a simple interaction based on the children's game rock-paper-scissors in order to answer our questions regarding how engagement and attributions of mental state to a robot vary in cases in which the robot either is likely to be perceived as malfunctioning or is clearly cheating. We used a between-participants design, with each participant seeing the robot play in one of three conditions: the *control conditon*, in which the robot plays according to the rules and announces all outcomes correctly; the *verbal cheat condition*, in which the robot sometimes declares itself the winner despite having lost or tied the round; or the *action cheat condition*, in which the robot sometimes changes the symbol that it has thrown after having seen that it has lost or tied the round. After a brief introduction to the task, which excludes any mention of possible cheating, participants are left to play twenty rounds of rock-paper-scissors with the robot. A short questionnaire is then administered.

Given these conditions, we propose the following hypotheses:

H1 The verbal cheat will be characterized as a malfunction while the action cheat will be interpreted as intentional cheating.

H2 Participants in the action cheat condition will become more socially engaged with the robot than in the other conditions.

H3 Participants in the action cheat condition will interpret such cheats as intentional attempts to modify the outcome of the game, and will thus make greater attributions of mental state to the robot.

A. Experimental Setup

Our experiment uses the "Wizard of Oz" arrangement [15], in which a human operator controls the actions of the robot from behind the scenes. The participant is unaware of the remote operator, and interacts with the robot as if the robot were autonomous. This setup allows us to quickly design rich interactions and to avoid software errors that could compromise our results. The latter consideration is especially important in this experiment, as one of the manipulations is intended to be easily mistaken for a malfunction.

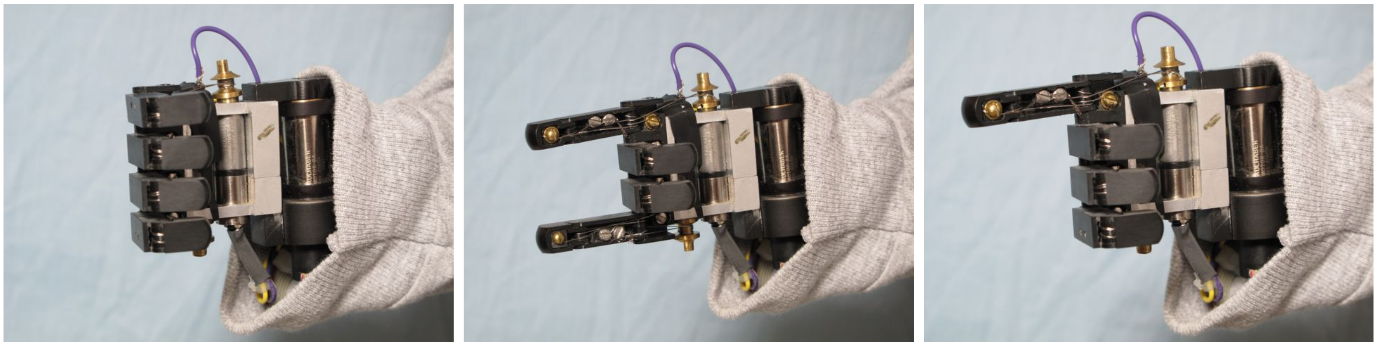
The robot sits on a desk facing the participant. A pair of standard computer speakers serve to play the robot's voice, while a microphone under the desk and a video camera behind the robot provide video and audio recording. The remote operator, or "wizard", controls the robot remotely from another room. Cameras in the robot's eyes allow the wizard to see the participant and interact appropriately.

During the task, the robot plays the game rock-paper-scissors with participants. Rock-paper-scissors is a game familiar to most Americans. It is played by children and adults alike to settle minor disputes and make simple decisions. To play, two participants face each other and each make a fist, raising their forearms parallel to their chests. The participants then lower their fists parallel to the ground. They repeat this raising-lowering motion three more times. These gestures may or may not be accompanied by the phrase, "rock, paper, scissors, shoot". The last time they lower their fist, they change their fist into one of three shapes. They may choose to retain the fist shape, symbolizing "rock", to flatten their hand out, symbolizing "paper", or to curl their ring and little fingers in towards their palm, spreading the index and middle fingers apart, and curling their thumb over the ring and little fingers, symbolizing "scissors". The symbol for "scissors" is identical to the well-known "peace" or "victory" hand gestures, but is held with the fingers pointing horizontally away from the body. The arm motions are performed in unison by the two participants, ensuring that the final hand gestures are made at the same time. The winner is determined by identifying which symbol "beats" the other. "Rock" smashes "scissors", "scissors" cut "paper", "paper" wraps around "rock". The cyclic nature of these relationships assures that no symbol has an advantage over the others.

We have chosen this game for its simple rules and familiarity, so that cheating behaviors will be more obvious to participants. The rounds are short, so many rounds can be played without exhausting the participant. The gestures used are easy to program into our robot. Finally, the game does not involve any physical contact, ensuring the safety of both human and robot.

To play, the robot raises and lowers its arm four times, and makes a gesture immediately after lowering its arm for the fourth time: "Rock, paper, scissors, shoot!". After each round, the robot announces the result: "Yes, I win", "Aw, you win" or "We have tied this round." The robot's choice of symbol on all rounds for all participants were randomly generated prior to the experiment, so all participants see the same sequence of throws. This also makes it easy for the wizard to keep a consistent rhythm while playing, and to determine the outcome of each round quickly.

In all conditions, the participant and the robot play 20 rounds of rock-paper-scissors, unless additional rounds are needed in the cheating conditions, as explained below. In the control condition, exactly 20 rounds are played, and the robot announces the outcome of each round correctly. In the verbal cheat condition, the robot incorrectly declares itself the victor on the 4th, 8th, and 15th rounds. In the action cheat condition, the robot cheats on the 4th, 8th, and 15th rounds by changing its hand gesture to the winning gesture (without repeating the full arm movement) and then declaring itself the victor. If the robot in fact wins fairly on a round designated for cheating, the cheat is attempted on the next round. The subsequent cheat rounds are pushed back, such that the spacing between rounds of cheating is preserved. For instance, if the robot wins the 4th round outright, but cheats on the 5th round, the next attempt to cheat is made on the 9th round, instead of the 8th. The entire interaction is then lengthened accordingly, such that there are always five rounds where the robot does not cheat after the final round of cheating. The randomly generated sequence



(a) Rock

(b) Paper

(c) Scissors

Fig. 1. The three modified gestures Nico uses while playing rock-paper-scissors. Mechanical limitations prevent use of the traditional gestures.

of throws is longer than 20, assuring that participants see a consistent sequence even during an extended interaction.

During the experiment, the wizard sits in an office separate from the lab where the robot is. In a conference room, another experimenter briefs the study participant, then brings them into the lab. The experimenter demonstrates three rounds of rock-paper-scissors so that the participant can see each of the robot's hand gestures and hear each of the robot's statements, before leaving the lab. The participant plays the requisite number of rounds of rock-paper-scissors with the robot, after which the robot thanks the participant for playing by saying, "Thank you for playing with me. Someone will come get you shortly." The experimenter comes back into the lab and brings the participant back to the conference room. The participant fills out the questionnaire and is then debriefed.

B. The Robot Platform

Nico is an upper torso humanoid robot with 22 mechanical degrees of freedom: 6 in each arm, 6 in the head, 2 in the right hand, a shoulder motion and the ability to twist its torso (see Figure 2). It is designed to mimic the size, proportions, and kinematic structure of a 12-month-old human male infant at the fiftieth percentile [16]. Nico has a foveated active vision system with two NTSC color cameras mounted in each eye. The robot is clothed in a child's t-shirt and cap, and has stylistic accents that give it a non-threatening appearance.

For this experiment, Nico's software is run on a local area network of 3 different computers. The programs to play its voice and to stream video from its cameras over the network run on a Linux machine located in the same room as the experimental area. A QNX machine connected to the network of microcontrollers driving Nico's actuators for the hand, arm, and head is also located in this room. The hand is driven by a custom motor board attached to an Atmega16 microcontroller running software that was developed in-house. Finally, the wizard uses a Linux laptop set up in another room connected over the network.

Nico's hand was designed with two mechanical degrees of freedom. The hand cannot exactly mimic the typical rock-paper-scissors gestures used by a human, so we developed a simple set of similar gestures for the robot to use (Figure 1). The robot's modified gestures are demonstrated to participants before the game; participants are allowed to use the normal gestures.

C. Survey

This study uses a survey that is adapted from the one used by Bainbridge et al. [17], which is modified from the version of the Interactive Experiences Questionnaire, originally developed by Lombard and Ditton [18], used in Kidd and Breazeal [19]. Kidd and Breazeal [19] used this questionnaire as a test of the perceived social presence of a set of characters in an interaction. We use it for its questions concerning the social characteristics of an agent. Our modified version. In our modified version, we added two Likert scale questions - *How well did Nico play the game?*, *Would you like to play rock, paper, scissors with Nico again?* and one open-ended question - *Did anything about Nico's behavior seem unusual? What?*

D. Coding

Five coders with no other involvement with the experiment examined responses to the question, "Did anything about Nico's behavior seem unusual? What?" These coders were broken up into two groups. Three coders were responsible for coding whether responses referred to Nico as "cheating" or "malfunctioning or making a mistake", and whether or not the participant appeared to anthropomorphize the robot in their response. Using the third coder as a tie-breaker, the majority response was taken for each of these measures. Inter-annotator agreement is listed in Table I.

We are also interested in the extent to which participants attributed mental state to the robot. Did they think of the robot as making choices and taking actions, or as a passive processor of the task at hand? The remaining two coders were given lists of all of the verb phrases related to Nico appearing in response to the question, "Did anything about Nico's behavior seem unusual? What?" They were tasked with classifying each verb phrase as being either in active or passive

Coders	Metric	"Cheating"	"Malfunction or Mistake"	"Anthropomorphize"
1&2	Cohen's Kappa	0.860	0.655	0.552
	Chi-squared	$\chi^2(1, N = 58) = 43.066, p < 0.001$	$\chi^2(1, N = 58) = 24.865, p < 0.001$	$\chi^2(1, N = 58) = 19.376, p < 0.001$
1&3	Cohen's Kappa	0.678	0.506	0.345
	Chi-squared	$\chi^2(1, N = 58) = 28.458, p < 0.001$	$\chi^2(1, N = 58) = 15.975, p < 0.001$	$\chi^2(1, N = 58) = 7.108, p = 0.008$
2&3	Cohen's Kappa	0.817	0.582	0.552
	Chi-squared	$\chi^2(1, N = 58) = 40.030, p < 0.001$	$\chi^2(1, N = 58) = 21.136, p < 0.001$	$\chi^2(1, N = 58) = 22.095, p < 0.001$

TABLE I
INTER-ANNOTATOR AGREEMENT RATING RESPONSES TO THE QUESTION: "DID ANYTHING ABOUT NICO'S BEHAVIOR SEEM UNUSUAL? WHAT?"

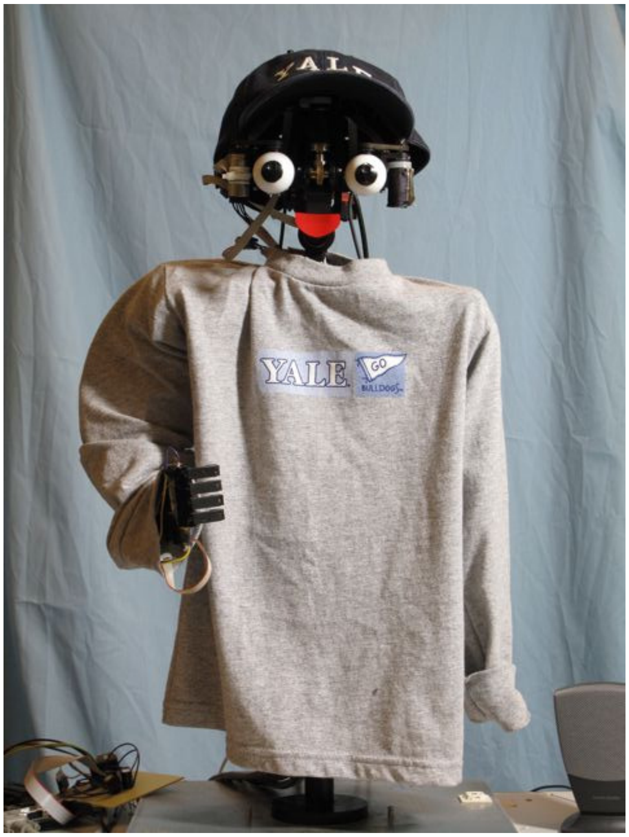


Fig. 2. The humanoid robot, Nico.

voice. Inter-annotator agreement was Cohen's Kappa = 0.394, $\chi^2(1, N = 144) = 35.332, p < 0.001$. The mean number of verbs classified as active was taken for each response. These sum was divided by the number of verb phrases appearing in each response, normalizing it to one. That is, we measured the proportion of verbs in each response classified as "active" or "passive".

Finally, from the video taken of the participants playing with Nico, we transcribed the participants' utterances and counted the number of words spoken by each participant during the interaction with Nico. We discarded those words spoken to the experimenter while the experimenter is in the room with the participant demonstrating the task. Because the phrase does not contain communicative intent, and is mostly used to synchronize timing and keep track of the number of movements before the throw, we did not count "Rock, paper, scissors, shoot" in the number of words uttered by the participant. We use the adjusted word count as a metric to quantify social engagement between conditions.

III. RESULTS

We recruited 73 participants from the Yale community through posters, the Facebook social networking site, and personal invitations. We discarded seven because of operator error, three because of technical malfunctions, and three because of failure to properly participate in the task. Operator errors consisted mainly of cheating in the wrong way during trials, or failure to start recording equipment. Mechanical problems included a stripped set-screw in the hand which needed replacing, and a cable in the hand that broke during one trial. Discounting discarded participants, 21 participants were placed in our control group, 20 into the verbal cheat group, and 19 into the action cheat group. Of these, we had 23 male participants, 32 female participants, and five who did not report demographic information.

One of our first concerns is establishing whether or not the paradigm of the experiment works – that is, is it ambiguous to the verbal cheat group whether Nico was cheating or malfunctioning, and do participants in the action cheat group perceive Nico as cheating more often than in the other groups? Hypothesis testing via a one-way analysis of variance shows that Nico's behavior affected participants perceptions both of cheating, $F(2, 56) = 33.407, p < .001$ and malfunctioning or making a mistake, $F(2, 56) = 14.000, p < .001$. Unless otherwise specified, all post hoc analyses presented in this paper use Fischer's LSD post hoc criterion. Post hoc analyses indicate that all mean differences are significant for cheating, and that all differences but that between the control group and action cheat group are significant for malfunctioning or making a mistake at $p < 0.05$. Malfunctioning or making a mistake is borderline significant at $p = 0.089$. Results from the classification of the written responses appear in Figure 3. Confirming our first hypothesis, the participants who saw the action cheat mentioned cheating, while the participants who saw the verbal cheat frequently described it as a mistake or malfunction, while only sometimes calling it cheating. The participants in the control group, unsurprisingly, were not sure at all about what the unusual behavior might be,

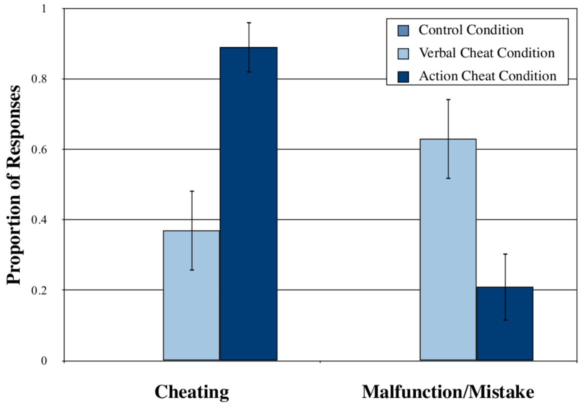


Fig. 3. Proportion of participants who indicated that Nico cheated or made a malfunction or mistake. Bars represent standard error.

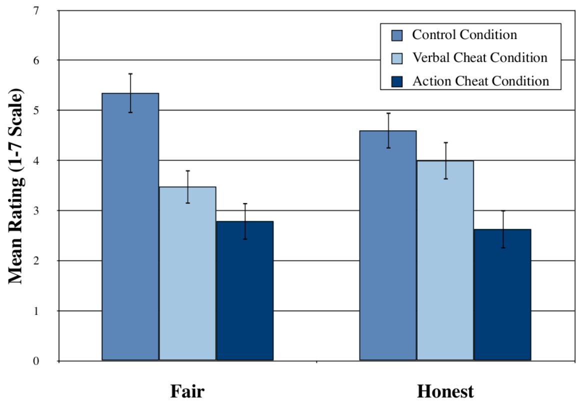


Fig. 4. Seven point Likert scale ratings evaluating Nico as "fair" and "honest" by participants. Bars represent standard error.

although many noticed that the computer-generated voice lacked emotion:

- "Robotic voice? Some fingers didn't work? I'm not really sure."
- "His expressions when he won seemed a bit flat."

The participants in the verbal cheat group might be aware that Nico always won when he cheated, but suspected that the cheat was just an error:

- "Sometimes I won but he said he did... did he just misrecognize my symbol?"
- "Sometimes he would pronounce the results of the match wrong [...] but he would never make a mistake in my favor."

Participants in the action cheat group were often quite certain of what the robot was doing:

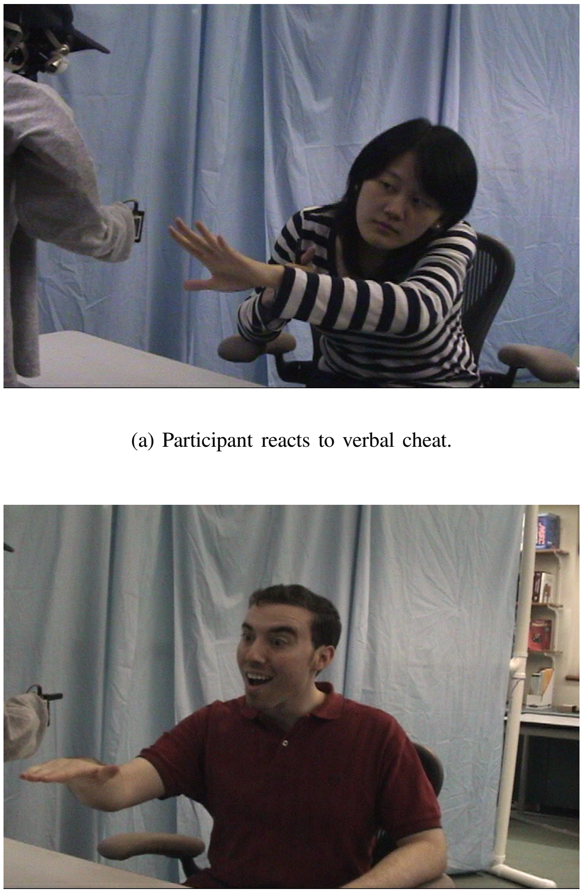
- "He cheats! Would change his sign after he saw mine!"
- "It was unusual that it KNEW how to cheat. I did not expect that at all."

Participants often had an emotional reaction to the robot's behavior and made unfavorable character attributions to the cheating robot. In Figure 4 we see that participants in the action cheat condition regarded the robot as less honest than did participants in the other two conditions. An analysis of variance shows that Nico's actions affected perceptions of the robot as "fair", $F(2, 59) = 12.573, p < .001$, and "honest", $F(2, 59) = 6.985, p = 0.002$. Post hoc analyses indicate that differences for both the action and verbal cheat conditions are significant when compared to control, but not when compared to each other for perceptions of the robot as "fair" $p < 0.05$. For perceptions of the robot as "honest", all mean differences are significant except for that between control and verbal at $p < 0.05$. As one might expect, in both the action cheat condition and the verbal cheat condition the game is rated less "fair" than in the control condition. Even if the robot does not mean to falsely report the outcome of the game, the game has not been played according to the rules, and this is reflected in their responses.

Qualitatively, we see that the participants' reactions in the verbal cheat case (Figure 5(a)) tend more towards confusion, while their reactions to the action cheat (Figure 5(b)) are more exaggerated, showing surprise, amusement, and occasionally anger. People speak more when they are socially engaged. Looking at the responses when participants were asked to rate "How often did you want to or did you speak to Nico?" on a Likert scale of one to seven, we see that they report wanting to speak with the verbal cheat robot more than the action-cheat robot (Figure 6(a)). An analysis of variance shows that there is a significant interaction between the experimental condition and this score $F(2, 51) = 4.683, p = 0.014$. Post hoc analyses using the Scheffé post hoc criterion indicate that the only significant interaction is between the verbal cheat and control groups $p = 0.004$. The interaction between the action and verbal cheat groups is nearly significant at $p = 0.062$. The interaction between the control and action groups is not significant at $p > 0.308$. When we count the actual number of words spoken by each participant during the interaction (Figure 6(b)) however, the result is suggestive that they might actually speak more to the action-cheat robot, although analysis of variance indicates that there is no significant interaction between the independent variable and the number of words participants spoke, $F(2, 51) = 1.248, p = 0.296$. It is possible that in their survey evaluation of their experience, they were simply trying to punish the "bad" cheating robot, but were more sympathetic to the robot which makes mistakes. In terms of our second hypothesis, our results, though inconclusive, are suggestive that any deviation from expected operation is sufficient to create a greater degree of engagement in the interaction. Though the action cheat may perform slightly better, this comes at the cost of ill-will towards the robot.

Finally, we investigate attributions of mental state to Nico across the three conditions. Do participants perceive the cheating behavior of the action cheat robot as more intentional and attribute more mental state to Nico in this condition, while perceiving the malfunctioning robot as simply running a flawed program?

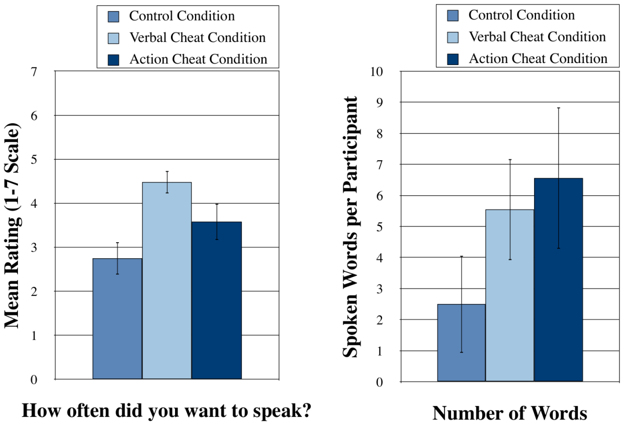
It is not easy to measure the attribution of mental state. How do we identify whether participants think of the robot as an



(a) Participant reacts to verbal cheat.

(b) Participant reacts to action cheat.

Fig. 5. Participant Reactions



(a) Seven point Likert scale responses to the question "How often did you want to or did you speak to Nico?" Bars represent standard error.

(b) Number of words spoken by participants while interacting with Nico. Bars represent standard error.

Fig. 6. Participants' self-reported and actual speech with the robot.

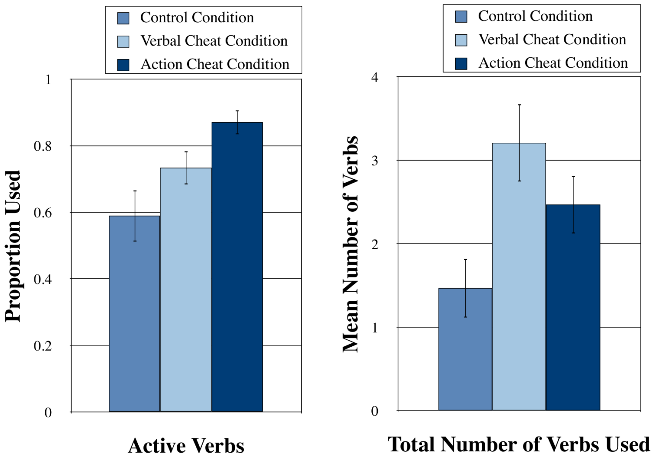
agent reasoning about a game, rather than a machine stepping through a task? Asking "How much does the robot think?" is not sufficient. Instead we have to rely on more subtle cues in the participants' behavior and written responses. The nature of participants' responses to the open-response question, "Did you notice anything unusual about Nico's behavior? What?" gives us insight into their attributions of mental state to Nico across the various conditions. For instance, members of the action cheat group said:

- "He seemed to change from rock to scissors or paper sometimes to win, but it seemed like a cute behavior rather than an annoying one - I liked it that he wanted to win!"
- "I noticed Nico cheated a few times. I didn't mind, it made him more interesting to play with."

While members of the verbal cheat group said:

- "He either misread my choice or his hands didn't move as they should or he cheats. Either way, at least 3 times he won when he should have lost. I think scissors is difficult. I would have corrected him if I thought he was programmed to listen."
- "More often than not, Nico either misinterpreted my hand position or was dishonest about the outcome of the game. The position he mostly had trouble with or wasn't honest about was scissors."

As can be seen in Figure 7(a), participants in the action cheat group use a higher ratio of active voice to passive voice verbs than those in the other two groups. An analysis of variance indicates that the interaction between study group and this ratio is significant, $F(2, 49) = 6.686, p = 0.003$. Post hoc analyses indicate that all pairwise interactions are borderline-significant $p <= 0.071$, and that the interaction between the action cheat and control groups is significant at $p < 0.001$. This use of active voice implies action and choice, and attributions of mental state to the robot, whereas passive voice represents things that happen to the robot. Participants perceive Nico as a more active participant in the scene when it cheats, but malfunctioning is something that happens to a device due to a flaw in its design. Participants in the verbal cheat group seemed to use more verbs to describe what happened, as if, in their uncertainty, they wanted to cover all possible explanations for the behavior (Figure 7(b)). However, an analysis of variance indicates a significant interaction between the study group and this statistic, $F(2, 56) = 4.978, p = 0.010$. Post hoc analyses indicate that the only pairwise interaction that is significant is between verbal and control $p = 0.003$, with action and control borderline significant at $p = 0.076$ and action and verbal not significant at $p = 0.188$. Additionally, participants anthropomorphize the active condition the most ($M = 0.89$, $SD = 0.072$), followed by the verbal ($M = 0.74$, $SD = 0.104$), then control ($M = 0.42$, $SD = 0.116$). Hypothesis testing by analysis of variance shows that the interaction between experimental condition and this variable is significant $F(2, 56) = 5.906, p = 0.005$. Post hoc analyses indicate that all pairwise interactions with the control group



(a) Proportion of active voice verbs used to describe Nico's actions in each condition. Bars represent standard error.

(b) Number of verbs used to describe Nico's actions in each condition. Bars represent standard error.

Fig. 7. Participants' self-reported and actual speech with the robot.

are significant $p <= 0.029$, but that the mean difference between the action and verbal cheat groups is not significant $p = 0.266$ These data agree with our third hypothesis that participants would attribute greater mental state to the robot that is perceived as actively trying to change the outcome of the match in the action cheat condition.

IV. Discussion

We have crafted a task around the children's game rock-paper-scissors in order to study attributions of mental state and social engagement with a robot. The different conditions in this experiment are distinguished by small, easy-to-program behaviors that can be attributed to either cheating or malfunctioning. This simple interaction lends itself to modification. Further investigation into the attribution of animacy, intentionality and mental state might focus on the modification of other parameters. One might vary the robot's movement between groups: one group seeing a smooth-moving, lifelike robot, while the other sees a robot that moves in a more mechanical-looking way.

Participants find both the verbal cheat and action cheat conditions to be more engaging than the control condition. Qualitatively, the participants in both cheating groups are more entertained than the control participants, most of whom appear to be bored halfway through the interaction. We have tested this by analyzing their vocalizations for word count and note that between the three groups, action cheat receives the most words, then verbal cheat, then control. This result is not statistically significant, however, and bears further investigation.

Even though it would appear that participants find both experimental conditions to be more engaging than the control, they are far more likely to make negative personality attributions to the robot with a clear "cheating" behavior. They rate the interaction with both cheating robots as less fair than the control, but they rate the action cheat robot as less honest, and they say that they do not want to speak to it as much, even though the result of counting the number of words spoken to each robot indicates that they may actually speak to it more. That is, their negative feelings towards the cheating robot lead them to say that they did not want to speak to it as much, independent of their actions. They sympathize with the robot in the verbal cheat condition, behave punitively towards the robot in the action cheat condition.

The attribution of mental state to a robotic partner has dramatic consequences for the relationship between robot and human. A friend has mental state, a vacuum cleaner does not. In our experiment, mental state is implied in the idea of "cheating" – a cheating opponent is acting out of a desire to win the game. A malfunction, on the other hand, is entirely accidental, and could be the fault of physical design or faulty programming, and can even occur entirely without involvement on the part of the robot. A malfunction is something that happens to a machine, while a social entity that can cheat has motivations and desires.

We find support for the idea of the "cheater" as an agent with mental state in our classification of the verbs in the written responses. Participants in the action cheat group use more active verbs to describe the robot's behavior, while the participants in the verbal cheat group use more passive verbs. Active verbs imply will on the robot's part, as it is deliberately acting in order to win the game. On the other hand, passive verbs are associated with objects and entities which exist in the world without acting upon it. The verbal cheat group perceives the robot's cheating as a flaw in the design, a mistake, a classification error, or a shortcoming in the techniques used to program it. The action cheat group, on the other hand, perceives the robot's cheating as a deliberately unscrupulous act, in which it not only wants to win, but planned out the appropriate steps to accomplish that goal.

When studying how humans perceive and interact with robots, we often wish to encourage them to think of the robot as a social entity. While other research has moved towards creating sophisticated emotional models which cause complex behavior [2], we demonstrate that there are much simpler ways to foster the attribution of mental state to a robot, as well as to enhance participants' engagement in the interaction. The action cheat behavior appears to create slightly more engagement than the verbal cheat, but at the cost of causing negative feelings towards a robot that is perceived as dishonest. We attribute this engagement partly to the deviation from expected behavior, and partly to attributions of mental state towards the robot. We believe that many interactions can be improved by the introduction of such simple behaviors, and that this should be exploited by designers in HRI. Bringing human and robot together to perform a simple, repetitive, familiar task and then having the robot behave unexpectedly can increase engagement and mental state attribution without complex behavioral or mechanical additions.

This study was limited in part by the characteristics of our participants. Participants were mostly Yale undergraduate and graduate students, and while none of them had seen the robot before, they were mostly younger, and therefore grew up with computers and similar technology. We would like to have been able to recruit a group of participants which more widely represented different age ranges and degrees of familiarity with technology. Additionally, we would have liked to have added more survey questions which could contribute to our conclusions about the participants' perception of the robot's mental state. Finally, we would like to have reduced the delay in the system, because although it was the same across all conditions and would not affect our conclusions, many of the participants found the interaction to be stilted because of it.

V. Conclusion

We have designed a robot that plays rock-paper-scissors with human participants. In one of our experimental conditions, the robot cheats by announcing the incorrect result (always in the robot's favor). In another, the robot cheats by changing its throw to the winning throw.

We find that, as expected, that participants are more likely to consider the verbal cheat as a malfunction and the action cheat as a cheating behavior. Participants in both cheating groups rate the interaction as less fair than those in the control group. Furthermore, the participants in those groups are more engaged in the interaction than the participants in the control group. They both report wanting to speak more, and actually speak more with the robot. Although we do not find a statistically significant difference in how much they speak to the action-cheat robot versus the verbal cheat robot, the results suggest that they speak more to the robot that cheats by changing its throw than the one whose actions could be interpreted as a mere malfunction.

Our results point towards a greater attribution of mental state to a cheating robot than one that behaves entirely as expected, providing a simple way for robotics researchers to increase the engagement of a human participant with the robot, although at the expense of positive emotions towards the robot.

In a game such as rock-paper-scissors, people expect consistent behavior. When the robot does something unexpected, especially when the behavior has a clear intention (as opposed to a malfunction) they are surprised into interacting socially. In the words of one participant, "You totally cheated! You're not allowed to cheat. You're the robot."

Acknowledgements

The authors would like to thank Ahmet Aktay, Emily Bernier, Christian Csar, Julie Hersh, Michael Mester, and Drew Westphal for the time that they spent coding. Support for this work was provided by National Science Foundation awards #0534610 (Quantitative Measures of Social Response in Autism), #0835767 (Understanding Regulation of Visual Attention in Autism through Computational and Robotic Modeling) and CAREER award #0238334 (Social Robots and Human Social Development). The authors also acknowledge the generous support of the DARPA Computer Science Futures II program, Microsoft, and the Sloan Foundation. This work was made possible through a software grant from QNX Software Systems Ltd and hardware grants by Ugobe Inc.

References

[1] A. Bandura, "Social cognitive theory: An agentic perspective," *Annual Review of Psychology*, vol. 52, pp. 1–26, 2001.

[2] C. Breazeal and B. Scassellati, "How to build robots that make friends and influence people," in *IEEE International Conference on Intelligent Robots and Systems (IROS-99)*, vol. 2. Kyongju, Korea: IEEE, August 1999, pp. 858–863.

[3] B. Mutlu, F. Yamaoka, T. Kanda, H. Ishiguro, and N. Hagita, "Nonverbal leakage in robots: communication of intentions through seemingly unintentional behavior," in *HRI '09: Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. La Jolla, California: ACM, March 2009, pp. 69–76.

[4] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artificial Intelligence*, vol. 166, no. 1–2, pp. 140–164, 2005.

[5] C. Breazeal, "Affective interaction between humans and robots," in *Proceedings of the 6th European Conference on Advances in Artificial Life (ECAL '01)*. Bremen, Germany: Springer-Verlag, September 2001, pp. 582–591.

[6] E. S. Kim, D. Leyzberg, K. M. Tsui, and B. Scassellati, "How people talk when teaching a robot," in *HRI '09: Proceedings of the 4th ACM/IEEE international conference on human-robot interaction*. La Jolla, California: ACM, March 2009, pp. 23–30.

[7] C. Crick and B. Scassellati, "Inferring narrative and intention from playground games," in *Proceedings of the 7th IEEE International Conference on Development and Learning (ICDL 2008)*. Monterrey, California: IEEE, August 2008.

[8] F. Heider and M. Simmel, "An experimental study of apparent behavior," *The American Journal of Psychology*, vol. 57, no. 2, pp. 243–259, 1944.

[9] B. J. Scholl and P. D. Tremoulet, "Perceptual causality and animacy," *Trends in cognitive sciences*, vol. 4, no. 8, pp. 299–309, 2000.

[10] D. Premack, "The infant's theory of self-propelled objects," *Cognition*, vol. 36, no. 1, pp. 1–16, 1990.

[11] J. K. Hamlin, K. Wynn, and P. Bloom, "Social evaluation by preverbal infants," *Nature*, vol. 450, no. 7169, pp. 557–559, 2007.

[12] B. Mutlu, S. Osman, J. Forlizzi, J. Hodgins, and S. Kiesler, "Perceptions of asimo: An exploration on co-operation and competition with humans and humanoid robots," in *HRI '06: Proceedings of the 1st ACM/IEEE Conference on Human-Robot Interaction*, vol. 2006. Salt Lake City, Utah: ACM, March 2006, pp. 351–352.

[13] C. Nass, Y. Moon, B. J. Fogg, B. Reeves, and D. C. Dryer, "Can computer personalities be human personalities?" *International Journal of Human - Computer Studies*, vol. 43, no. 2, pp. 223–239, 1995.

[14] P. H. Kahn, N. G. Freier, T. Kanda, H. Ishiguro, J. H. Ruckert, R. L. Severson, and S. K. Kane, "Design patterns for sociality in human-robot interaction," in *HRI '08: Proceedings of the 3rd ACM/IEEE international conference on human-robot interaction*. Amsterdam, The Netherlands: ACM, March 2008, pp. 97–104.

[15] A. Steinfeld, O. C. Jenkins, and B. Scassellati, "The oz of wizard: Simulating the human for interaction research," in *HRI '09: Proceedings of the 4th ACM/IEEE international conference on human-robot interaction*. La Jolla, California, USA: ACM, March 2008, pp. 101–107.

[16] G. Sun and B. Scassellati, "A fast and efficient model for learning to reach." *International Journal of Humanoid Robotics*, vol. 2, no. 4, pp. 391–413, 2005.

[17] W. A. Bainbridge, J. W. Hart, E. S. Kim, and B. Scassellati, "The effect of presence on human-robot interaction," in *17th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. Munich, Germany: IEEE, August 2008.

[18] M. Lombard, T. B. Ditton, D. Crane, B. Davis, G. Gil-Egui, K. Horvath, and J. Rossman, "Measuring presence: A literature-based approach to the development of a standardized paper-and-pencil instrument," in *Presence 2000: The Third International Workshop on Presence, Netherlands*, Delft, The Netherlands, 2000.

[19] C. D. Kidd and C. Breazeal, "Effect of a robot on user perceptions," in *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 4. Sendai, Japan: IEEE, October 2004, pp. 3559–3564.