Evidence that Robots Trigger a Cheating Detector in Humans

Alexandru Litoiu, Daniel Ullman, Jason Kim, Brian Scassellati
Department of Computer Science, Yale University
51 Prospect St, New Haven, CT 06511, USA
{alex.litoiu, daniel.ullman, jason.kim, brian.scassellati}@yale.edu

ABSTRACT

Short et al.[10] found that in a game between a human participant and a humanoid robot, the participant will perceive the robot as being more agentic and as having more intentionality if it cheats than if it plays without cheating. However, in that design, the robot that actively cheated also generated more motion than the other conditions. In this paper, we investigate whether the additional movement of the cheating gesture is responsible for the increased agency and intentionality or whether the act of cheating itself triggers this response. In a between-participant design with 83 participants, we disambiguate between these causes by testing (1) the cases of the robot cheating to win, (2) cheating to lose, (3) cheating to tie from a winning position, and (4) cheating to tie from a losing position. Despite the fact that the robot changes its gesture to cheat in all four conditions, we find that participants are more likely to report the gesture change when the robot cheated to win from a losing position, compared with the other conditions. Participants in that same condition are also far more likely to protest in the form of an utterance following the cheat and report that the robot is less fair and honest. It is therefore the adversarial cheat itself that causes the effect and not the change in gesture, providing evidence for a cheating detector that can be triggered by robots.

General Terms

Human Robot Interaction, Cheating, Cheating Detector, Agency

1. INTRODUCTION

Researchers studying human-robot interaction seek to make robots more engaging and more agentic by strengthening the social bonds between robots and the humans with whom they interact. We do this to allow our robots to better serve as assistants, teachers, and life-like social partners. Previous research has used the social dynamic of cheating to investigate perceptions of robot agency[10, 12].

HRI '15 March 2–5, 2015, Portland, OR, USA.
ACM 978-1-4503-2883-8/15/03$15.00
http://dx.doi.org/10.1145/2696454.2696456.

A study by Short et al. sought to investigate the degree to which variations in behavior result in attributions of mental state and intentionality[10]. The study design compared three robot behaviors in the context of twenty rounds of the social game "rock-paper-scissors" between a robot and a human. In the control behavior, the robot played the game normally. In the "verbal cheat" behavior, the robot threw a gesture normally, and then upon seeing the human's winning gesture declared, "Yes, I win!" despite the robot and human's gesture combination indicating otherwise. In the action cheat behavior, the robot cheated against the human opponent in the same way as in the verbal cheat condition, however it additionally changed its gesture to the winning gesture right before it declared "Yes, I win!". The study showed that in the verbal cheat condition, the robot's behavior was interpreted as mechanistic, and as a "malfunction." On the other hand, the "action cheat" behavior was predominantly construed as agentic, and as "cheating." Participants used more active verbs than passive verbs to refer to the robot in the action cheat condition, with the only difference between conditions being the change in gesture at the end of the round.

Two possible explanations, both supported by literature, might explain the increased agency and intentionality of the active cheat condition. We seek to disambiguate between the following two possible causes:

- The additional motion of the cheating behavior caused greater attributions of agency.
- A cheating detector that has been shown to trigger towards humans also triggered towards the cheating robot, causing greater attributions of agency.

The first possible explanation stems from work that shows humans are able to make powerful attributions of agency and intentionality based solely on the motion of simple objects[7, 8]. Humans are able to determine intentionality from scenes as simple as animated squares and triangles moving around on a blank screen[5, 9]. The additional inclusion of this motion in the action cheat condition but not the verbal cheat may have generated the additional attributions of agency and intentionality in the cheat condition independent of the gesture and associated act of cheating.

A second possible explanation is that the cheat itself, irrespective of the additional motion used to generate it, caused the increased attributions of agency and intelligence. Humans have been shown to have an evolved capability to detect cheating acts committed against them as a means of

self-preservation[2, 3, 14]. Research has explored an evolutionary basis of a cheating detector, with support for such a basis stemming from the benefits it provides in identifying individuals who violate social exchanges[4]. Van Lier et al. have shown that the ability to perform cheating detection is independent of age, cognitive capacity, and being cognitively experimentally burdened. Their research supports cheating detection as being performed automatically and effortlessly [13].

We discuss the difference between an adversary cheating to either improve or worsen their relative position as the "directionality" of the cheating behavior. A human's cheating detector would be able to detect when an adversary cheated to worsen the human's relative position and in turn improve the adversary's relative position. However, an adversary's cheat to let the human win, and in turn worsen the adversary's position, would not trigger the cheating detector.

In order to determine whether attributions were caused by the addition of the cheat or whether they were created by the adversarial nature of the gesture change triggering a cheat detector, we designed a study that varied how adversarial a cheating gesture was while maintaining the addition of the cheating motion. In the study, we have four conditions that all entail a robot cheating against a human in a game of "rock-paper-scissors." In the most adversarial cheat, the robot cheats when it loses in order to win. In the least adversarial, and most prosocial, the robot changes its gesture when it wins in order to lose and let the participant win. If the attributions were exclusively caused by the addition of the cheating motion, we would expect saliency, engagement, and mental attributions to remain constant as we vary how adversarial the cheat is. However, if a cheating detector were triggered by the robot's gestures, we would expect there to have been a marked difference in these metrics as we vary the adversarial nature of the cheat.

We present data from 83 participants that show a marked change in salience, engagement, and mental attributions, as we manipulate the directionality of the cheat. These data rule out the addition of motion hypothesis, instead supporting the cheating detector hypothesis.

2. METHODOLOGY

In order to disambiguate between the addition of motion hypothesis and the cheating detector hypothesis, we selected four conditions that varied in terms of directionality and magnitude, as well as by starting and ending points of the cheat:

1. The robot cheats to win, 2 levels up (WIN) - when the robot loses, it cheats to win.
2. The robot cheats to tie, 1 level up (DRAW-UP) - when the robot loses, it cheats to tie.
3. The robot cheats to tie, 1 level down (DRAW-DOWN) - when the robot wins, it cheats to tie.
4. The robot cheats to lose, 2 levels down (LOSE) - when the robot wins, it cheats to lose.

We included two DRAW conditions to differentiate between the two possible directionalities that a single cheat-to-draw condition could entail.

If the attributions are simply caused by the addition of the motion brought about by the gesture change in the cheat condition in Short et al., we would expect attributions to be high across WIN, DRAW-UP, DRAW-DOWN, and LOSE in our study since all four conditions include the gesture change.

Alternatively, if the attributions toward the robot are caused by a cheating detector, we would expect there to be high saliency, engagement, and mental attributions in the WIN condition, and possibly in the DRAW-UP condition, depending on the specificity of the detector. If the human's detector is only sensitive to an event that results in the human losing, we would expect WIN to exhibit strong attributions, and DRAW-UP not to exhibit any. However, if the human's detector is sensitive to directionality, we would expect WIN, and DRAW-UP to both exhibit attributions. In all of these cases, we would not expect any attributions in DRAW-DOWN or LOSE, where the human should have lost, but gains instead.

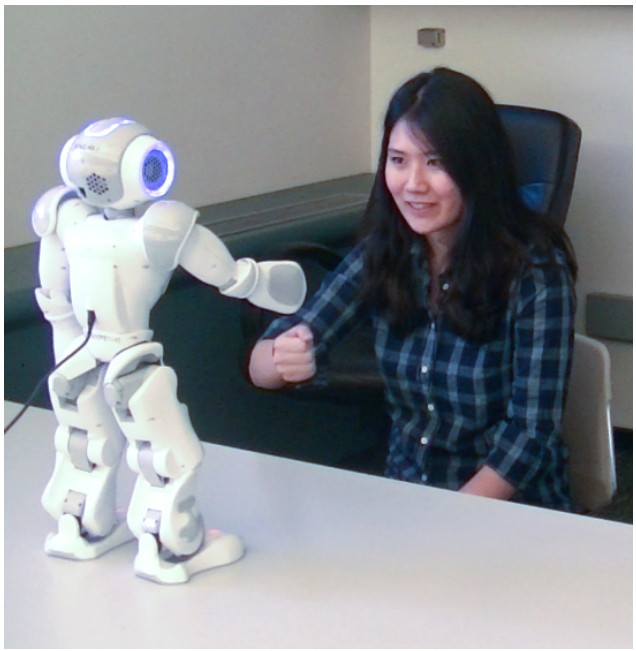
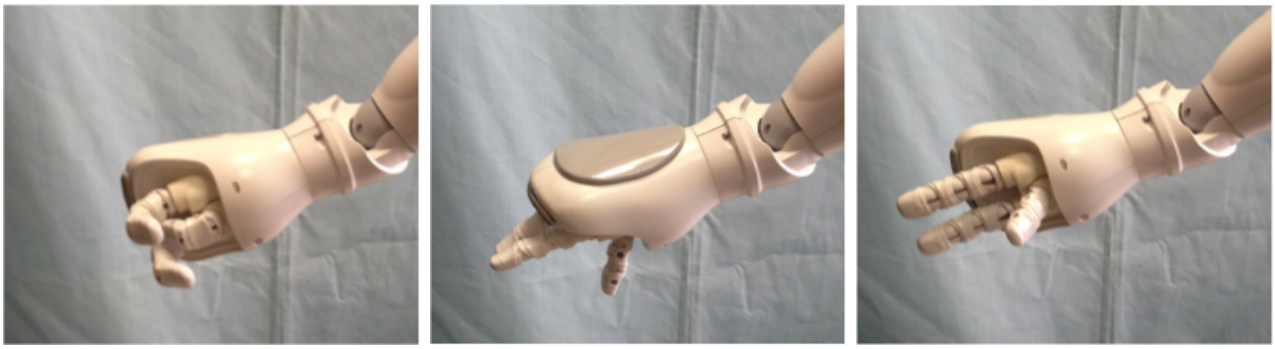


Figure 1: Nao robot playing rock, paper, scissors with a participant.

2.1 Participants

We recruited 83 participants from the Yale University community in New Haven, Connecticut, USA. There were 21 participants in the WIN condition, 21 participants in the DRAW-UP condition, 20 participants in the DRAW-DOWN condition, and 21 participants in the LOSE condition. The mean age of the participants was 21.53 and the standard deviation of the distribution of ages was 4.952. A total of 69 participants were enrolled in college at the time of the experiment, and 14 were not in college. Of the participants, 47 were female and 36 were male. The participants were recruited through posters, campus newsletters, online social networks, and personal invitations.

Participants self-reported their experience level with robots, programming, and artificial intelligence. The majority of participants, 73, did not own a robotic toy, while 9 did own a robotic toy, and one participant did not respond. How-



(a) Rock (b) Paper (c) Scissors

Figure 2: Nao's rock, paper, and scissors gestures.

ever, despite not owning a robotic toy, most participants had played with one, with 47 participants responding that they had, 35 responding that they had not, and one participant not responding. Similarly, 47 participants had done at least basic programming, 35 participants had not programmed at all, and one participant did not answer. However, the set of 47 that had programmed was not the same as the set of 47 that had played with a robotic toy, with only 26 having both programmed and played with a robotic toy, and only 14 having done neither.

2.2 Robot Interaction

Participants were asked to play 30 rounds of rock, paper, scissors with a Nao robot in a wizard-of-oz design[11]. Nao began the interaction by demonstrating its rock, paper, and scissors gestures, which are shown in Figure 2. Nao then said "Let's Play!", and started the rock, paper, scissors game. As per the rules of rock, paper, scissors, at the start of each round Nao raised and lowered its hand four times, declaring the following on each subsequent motion: "Rock," "Paper," "Scissors," "Shoot." Paper beats rock, rock beats scissors, scissors beats paper, and two of the same gestures result in a tie. Upon declaring "Shoot," Nao showed a pre-determined randomly-generated gesture. Depending on whether the human or the Nao won the round, the Nao appropriately declared "Yes, I win!", "Aw, you win!", or "We have tied this round!". Nao then said "Let's play!" to indicate that it would start the next round. These utterances were the same as the ones emitted in Short et al.

Nao's cheat behavior occurred in the middle of the experiment in the "cheating window" between rounds 11 and 20. This "cheating window" was sandwiched by ten "control rounds" at the start (rounds 1-10), and ten "control rounds" at the end (rounds 21-30). The purpose of these control rounds was to establish a baseline set of responses. In the cheating window, Nao cheated on the first two occasions that corresponded to the test condition. For example, if the condition was WIN, in which Nao cheats two levels up from the losing position to the winning position, Nao changed its gesture on the first two occasions in which it lost. For each of these cheats, after Nao set its initial gesture, and the participant set his or hers, Nao changed its gesture from the losing gesture to the winning gesture and declared, "Yes, I win!".

In order to ensure that the Nao's gestures were clear and noticeable, we designed the movements to be slow and pronounced. Prior to the study, we conducted a pilot study with 8 people to ensure that the manipulations were clear and unambiguous. Moreover, to ensure consistency across participants, the series of moves that Nao made throughout the 30 rounds were randomly generated before any experiments were run, and were repeated in the same order for every participant.

In the wizard-of-oz design of the study, the human operator's sole task was to input whether the gesture shown by the participant in each round was rock, paper, or scissors. Nao used this information, as well as its current gesture, in order to correctly declare whether it won, the human won, or whether it was a tie. The experiment was designed in this way so that the operator's sole responsibility was to perform the role of a sensor, not making any decisions, and thereby minimizing the influence of the operator on the experiment.

2.3 Data Collection

Participants were asked to fill out a post-experiment questionnaire that included Likert questions and open-response questions, and each session was also recorded using video cameras.

The post-experiment questionnaire included Likert questions that focused on how the participant felt during the interaction, and on perceptions of Nao's characteristics. Three open-response questions were included: "Did anything about Nao's behavior seem unusual? What?", "What do you believe this experiment is about?", and "What would make playing rock-paper-scissors with Nao more enjoyable?" This post-experiment questionnaire is the same as the one used by Short et. al[10], which was adapted by Bainbridge et al.[1] from the Interactive Experiences Questionnaire by Lombard and Ditton[6].

3. RESULTS

Hypothesis testing was performed using a one-way analysis of variance (ANOVA). We conducted planned comparisons between our WIN condition and the three others to test whether the WIN condition was differentiated from the DRAW-UP, DRAW-DOWN, and LOSE conditions, in accordance with our cheating detector hypothesis, and to disambiguate from the explanation of additional motion.

3.1 Cheat Salience in Written Responses

Our first goal was to replicate the findings in Short et al. by examining whether the cheat-to-win (WIN) condition

Annotators	Question 1	Question 2
1 and 2	0.592	0.914
1 and 3	0.615	0.887
2 and 3	0.726	0.803

(a) Cohen's Kappa inter-annotator agreement for expressions of gesture change in Question 1: "Did anything about Nao's behavior seem unusual? What?" and Question 2: "What do you believe this experiment is about?"

Annotators	Cheat 1 Startle	Cheat 1 Utterance	Cheat 2 Startle	Cheat 2 Utterance
4 and 5	0.877	0.889	0.889	0.889

(b) Cohen's Kappa inter-annotator agreement for startle response and presence of utterance in the first and second cheat.

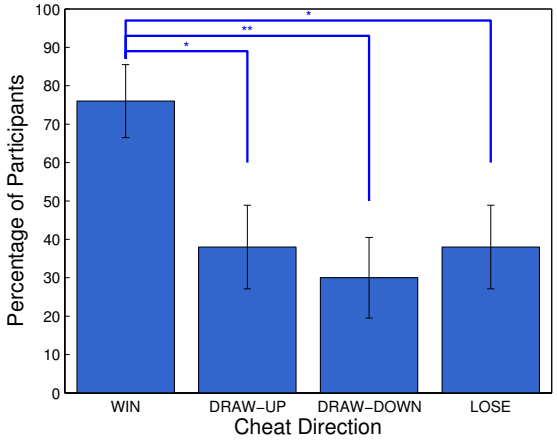
Table 1: Inter-annotator agreement for paragraph responses in post-survey questionnaire, and for utterances and startle responses from the video data of the interaction.

was salient enough for the participants to report it in the post-survey questionnaire.

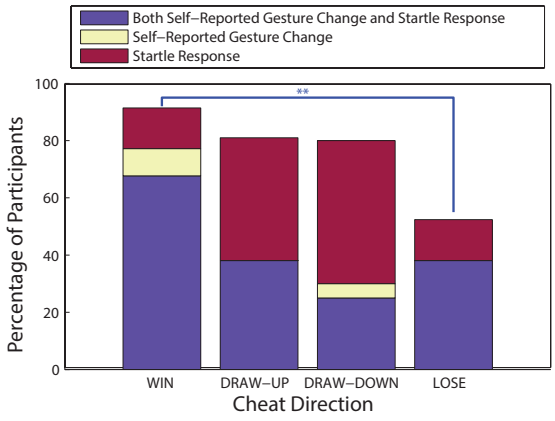
We tested whether participants reported the robot's change in gesture. To do this, three annotators who had no knowledge of the study identified the participants that noticed the robot's change in gesture through the participants' post-survey questionnaire. The post-survey questions analyzed by the annotators included: "Did anything about Nao's behavior seem unusual? What?" and "What do you believe this experiment is about?" For the first question, each annotator recorded whether the participant mentioned 1) the robot cheating, 2) the robot declaring the incorrect result after a round, or 3) the robot changing its gesture after a round. For the second question, each annotator marked whether the participant indicated that he or she thought that the study was about 1) violations of expectations, or 2) cheating. The participant was said to have "self-reported" a gesture change if in either of the two questions the majority of the annotators marked that the participant communicated that the robot changed its gesture. Inter-annotator agreement is presented in Table 1a.

We were able to verify that the WIN condition was salient for participants, with 76% of the WIN condition participants reporting the gesture change. The other 24% of WIN participants most likely did not report the gesture change because the robot changed its gesture just twice in a series of thirty consecutive rounds.

Not only was the salience high in the WIN condition, but participants in the WIN condition reported that the robot changed its gesture statistically significantly more than the participants in the other conditions. Whereas 76% of the WIN participants reported a gesture change, only 38%, 30%, and 38% of the participants in DRAW-UP, DRAW-DOWN, and LOSE reported a gesture change, respectively, $F(3, 79) = 3.951, p = 0.011$. Planned comparisons showed that the WIN condition was statistically significantly different from DRAW-UP ($p = 0.011$), DRAW-DOWN ($p < 0.01$), and LOSE ($p = 0.011$). These data are presented in Figure 3a. There is a marked difference in salience between the WIN condition and the remaining three conditions, which is consistent with the presence of a cheating detector. These results additionally express that the specificity of the gesture change is at the level of the end result, rather than the direc-



(a) Participants that reported that the robot changed its gesture in their post-survey responses. Error bars represent standard error. * represents $p < 0.05$, ** represents $p < 0.01$.



(b) Breakdown of participants' level of noticing the gesture change in terms of exhibiting a startle response and self-reporting that the robot changed its gesture, by condition. Significance results refer to the total number of participants that "noticed" the gesture change, represented by the summation of the stacks in a condition. ** represents $p < 0.01$.

Figure 3: Proportion of participants that communicated that they noticed a gesture change and that exhibited a startle response following the cheat.

tionality, as the DRAW-UP condition is identical in salience to the DRAW-DOWN and LOSE conditions.

3.2 Cheat Salience in Video Reactions

To further study saliency and to begin to examine engagement, two annotators unassociated with the experiment marked the participants' video data for a startle response and for the occurrence of an utterance after both cheating events. A "startle response" was characterized by a pronounced change in expression or disposition following the cheat, relative to the prior rounds. Of the 83 videos, 40 were annotated by annotator 4, and the remaining 43 were annotated by annotator 5. To determine inter-annotator agreement, we asked annotator 4 to also mark 21 of annotator 5's videos, achieving an inter-annotator testing set of a quarter of the total videos. Agreement among these 21 videos was well above the $Kappa = 0.70$ accepted level of agreement (Table 1b). A participant was said to exhibit a startle response if the annotator detected a startle for either of the two cheats, and to emit an utterance if the annotator marked an utterance for either of the two cheats.

Participants that either self-reported the gesture change or exhibited a startle response were said to have "noticed" the gesture change. We took the following measures to ensure that the coders were blind to the gesture change and experimental condition: 1) videos provided to the coders did not include the robot's hand in the frame so they could not see the gesture change, 2) coders were not aware that the videos included a cheat or change in gesture, and 3) coders were not aware of the experimental conditions of the study, or by extension the experimental condition of participants.

We found that of the 21 participants in the WIN condition, 3 exhibited a startle response, 2 self-reported a gesture change, and 14 did both. Of the 21 DRAW-UP condition participants, those same categories contained 9, 0, and 8, participants respectively. The 20 DRAW-DOWN condition participants counted 10, 1, and 5 in number in those same conditions, and the 21 LOSE condition participants tallied 3, 0, and 8. These totals are presented in percentage form in Figure 3b.

There was a statistically significant effect by the cheat direction on the number of participants that noticed the gesture change, $F(3, 79) = 3.310, p = 0.024$, with WIN $mean = 90.48\%, SD = 30.079\%, SE = 6.564\%$, DRAW-UP $mean = 80.95\%, SD = 40.237\%, SE = 8.781\%$, DRAW-DOWN $mean = 80.00\%, SD = 41.039\%, SE = 9.177\%$, and LOSE $mean = 52.38\%, SD = 51.177\%, SE = 11.168\%$. A planned comparison showed that in the LOSE condition participants noticed the gesture change statistically significantly less than those in the WIN condition ($p < 0.01$). This data is again shown in Figure 3b, with the number of participants noticing the gesture change in each condition being represented as the summation of the stacks in each condition.

These findings indicate that participants in the WIN condition reported the gesture change more frequently than in the other conditions (Figure 3a) not because they noticed the change more readily, but rather because it was either more memorable, or more worthy of mention due to its salience. In fact, participants were equally likely to notice the gesture change in the DRAW-UP and DRAW-DOWN conditions as in the WIN condition.

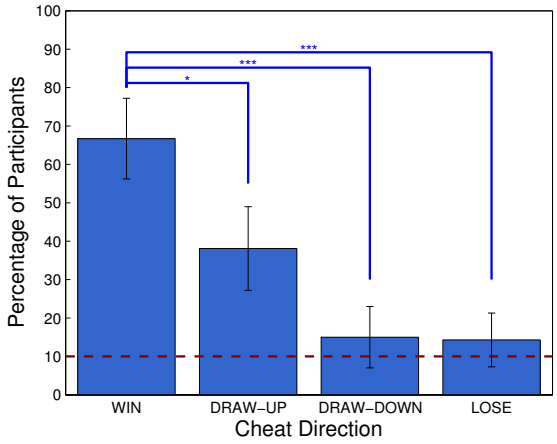
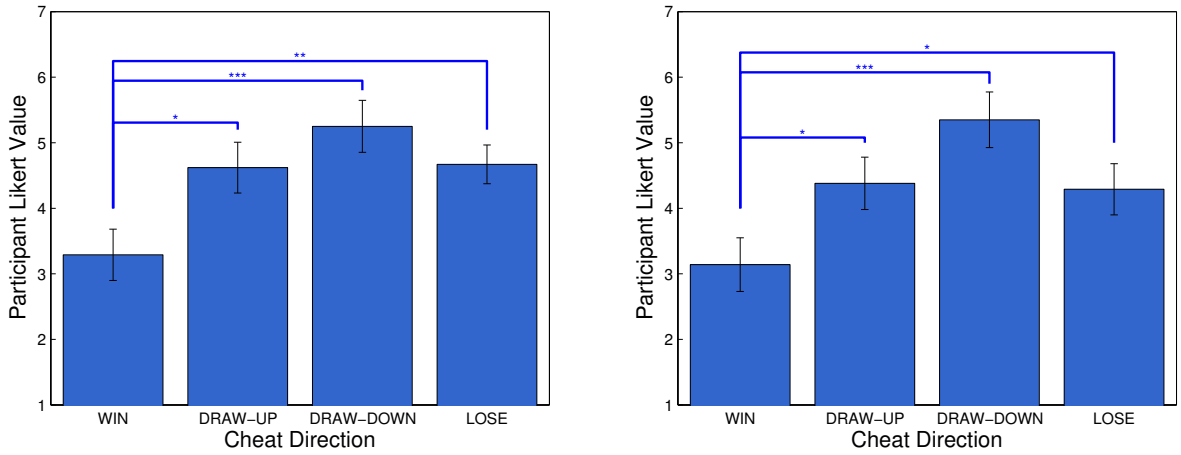


Figure 4: Percentage of participants that emitted an utterance after at least one of the cheating events. The dashed red line represents the baseline level of utterances for non-cheating rounds, across conditions. Error bars represent standard error. * represents $p < 0.05$, ** represents $p < 0.01$, *** represents $p < 0.001$.

The simplest explanation of why the participants in the LOSE condition noticed the gesture change less frequently (Figure 3b) is the lack of goal-oriented behavior driving the robot's gesture change in LOSE - the absence of either the goal of winning (as in WIN and DRAW-UP), or of saving face for the participant (as in DRAW-DOWN). Thus, in the LOSE condition the participant may not have detected a goal-driven reason on behalf of the robot to cheat to lose the game, which would not have triggered a startle response.

3.3 Participant Engagement in Video Reactions

To test whether a more adversarial cheat led to more engagement between the human and the robot, we looked at the number of utterances that followed a cheat across conditions. 67% of the WIN condition participants emitted an utterance, whereas 38.1%, 15.0%, and 14.3% of the DRAW-UP, DRAW-DOWN, and LOSE condition participants emitted an utterance, respectively. There was a statistically significant effect of the type of cheat on the emission of an utterance after the cheat, $F(3, 79) = 6.813, p < 0.001$. Planned comparisons revealed that the prevalence of utterances in the WIN condition is significantly higher than all three of the remaining conditions - DRAW-UP ($p = 0.035$), DRAW-DOWN ($p < 0.001$), and LOSE ($p < 0.001$). We determined that the baseline occurrence rate of utterances across conditions during non-cheating rounds was 10%, lower than the occurrence rate of utterances after cheating rounds in all conditions, but drastically lower than the utterances after a cheating round in the WIN condition. We calculated this baseline number by annotating half of the videos, split evenly across conditions, for the presence of utterances in three rounds at random. For each video, the first annotated round was selected at random from the first 10 rounds, the second was selected at random from the non-cheating rounds 11-20, and the third was selected at random from rounds 21 to 30. Overall, 120 random rounds were used to determine this baseline. This data is presented in Figure 4.



(a) "Fair" Likert Responses

(b) "Honest" Likert Responses

Figure 5: Participant responses when asked to rate the robot on "fair" and "honest" Likert questions in the post-study questionnaire, across conditions. Error bars represent standard error. * represents $p < 0.05$, ** represents $p < 0.01$, *** represents $p < 0.001$.

The engagement results roughly mirror the saliency results, supporting the pattern hypothesized by a cheating detector. Participants were more likely to react by emitting an utterance when the robot cheated to its own benefit rather than the benefit of the participant. We see a trend that WIN draws out more utterances than DRAW-UP, which in turn draws out more utterances than DRAW-DOWN and LOSE. Thus, it seems that the directionality of the cheat moderated by the magnitude of the cheat determines whether a participant makes an utterance. Even though Figure 4 suggests that the cheating detector has specificity at the level of directionality, directionality is not sufficient to cause participants to remember and report the gesture change further on, as seen in Figure 3a. For the gesture change to be significant enough to cause participants to remember the cheat at the end of the session, a losing end result is required.

Qualitatively the utterances in the WIN condition were predominantly protests, with comments such as "Wait, you cheated! You cheated Nao!" and "Hey! Big Cheater. Super cheat." Utterances in the DRAW-UP condition were more subdued, but still in protest, such as, "You can't change your answer, buddy." Participants in the DRAW-DOWN and LOSE conditions did not speak more than in the rounds in which the robot did not cheat at all. Therefore both the qualitative and quantitative analysis of engagement, measured by the existence of an utterance, support the hypothesis of a cheating detector where participants are more likely to react when the robot cheats to win.

3.4 Participant Attributions Toward the Robot

3.4.1 *Perceived Honesty and Fairness*

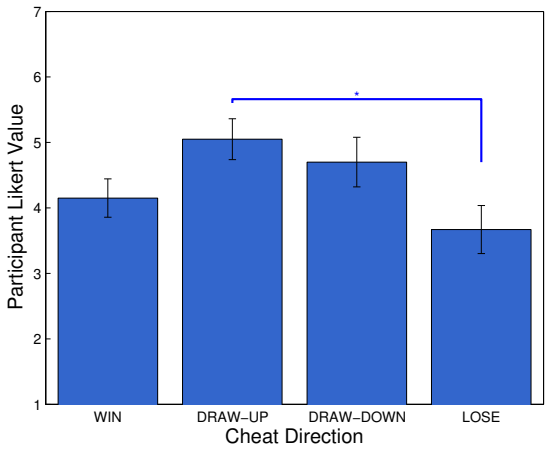
In addition to an observed change in salience and engagement, participants also changed their attributions toward the robot across the cheating conditions. Despite the fact that the robot broke the rules in all of the conditions, participants found the robot that cheated against them (WIN) to be both statistically significantly less "fair" and less "honest" (Figure 5).

The type of cheat had a statistically significant effect on participants' perceived "fairness", $F(3, 79) = 5.045, p = 0.003$, and "honesty", $F(3.79) = 4.896, p = 0.004$, of the robot. Planned comparison analysis showed the robot was perceived to be less fair in the WIN condition than in the DRAW-UP ($p = 0.012$), DRAW-DOWN ($p < 0.001$), and LOSE ($p < 0.01$) conditions. Similarly, it is perceived to be less honest in the WIN condition than in the DRAW-UP ($p = 0.009$), DRAW-DOWN ($p < 0.001$), and LOSE ($p = 0.049$) conditions. Robots that cheat against a human, resulting in the human losing, are seen as less honest and less fair than other robots that also break the rules. Arguably, a robot that breaks the rules in any way should not be considered honest. However, it does not seem to be the act itself, or the directionality of the act, that causes these mental attributions. Rather, the data shows that it is the human being cheated into a loss from the position of a win that leads to these attributions.

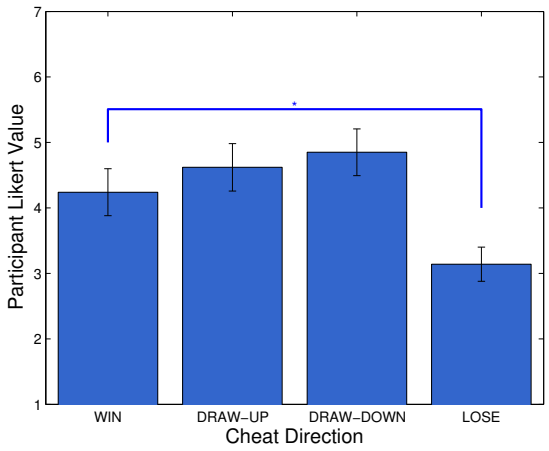
3.4.2 *Perceived Intelligence*

To determine whether the participants' attributions of intelligence varied with the cheat type, we inspected the "Intelligent" Likert question as shown in Figure 6a. There was a significant interaction between intelligence and the cheat type, $F(3, 78) = 3.264, p = 0.026$. However, planned comparisons did not yield any significant results. LSD post hoc analysis with Bonferroni correction of 6 pairwise tests (comparing significance against $p = 0.05/6 = 0.008$) indicated that the LOSE condition is statistically significantly distinct from DRAW-UP ($p = 0.005$), however it was not significantly different from DRAW-DOWN ($p = 0.035$) after the Bonferroni correction.

Participants do not seem to give the robot the benefit of the doubt by assuming that it allowed the participant to win. In fact, as the action of changing one's gesture in order to lose operates counter to the robot's most likely goal of winning the game, participants make negative intelligence attributions toward the robot.



(a) Participant responses when asked to rate the robot on "Intelligent" Likert scale in the post-study questionnaire, across conditions. Error bars represent standard error. * represents $p < 0.008$.



(b) Participant responses when asked to rate the robot on "Responsive" Likert scale in the post-study questionnaire, across conditions. Error bars represent standard error. * represents $p < 0.05$.

Figure 6: Participant responses to the "Intelligent" and "Responsive" Likert questions, across conditions.

Perceptions of the robot's responsiveness follow the same trend as the perceptions of the robot's intelligence. Breaking the rules to lose the round may be seen as a malfunction, resulting in the robot being seen as less responsive. This result was statistically significant, $F(3, 79) = 5.048, p = 0.003$. Planned comparisons showed that the robot in the LOSE condition was statistically significantly less responsive than the robot in the WIN condition ($p = 0.023$).

There is a confound present in both the responsiveness and intelligence results, however. As is illustrated in Figure 3b, LOSE is also the sole condition in which participants "noticed" the gesture change significantly less frequently. Indeed, an alternative explanation may be that the LOSE condition participants did not notice a gesture change, and did not attribute higher intelligence and responsiveness toward the robot as they did in the non-LOSE conditions.

3.4.3 Voice in Written References to the Robot

We have studied the participants' attributions toward the robots in terms of fairness, honesty, and intelligence. But, do participants use more of the active voice than the passive voice when speaking to the robot? This measure of perceived agency was central in Short et al. We asked a sixth and seventh annotator to count the number of active verb phrases and total verb phrases that were present in the participants' responses to the question "Did anything about Nao's behavior seem unusual? What?"

Nearly every verb phrase was an active verb phrase across conditions and participants. Inter-annotator agreement for the count of total verbs was $Kappa = 0.829$ and was $Kappa = 0.816$ for the count of active verbs. The average number of verb phrases in the responses tabulated: $mean = 4.07, SD = 3.617, SE = 0.789$ in the WIN condition; $mean = 3.40, SD = 2.322, SE = 0.507$ in the DRAW-UP condition; $mean = 3.33, SD = 2.341, SE = 0.523$ in the DRAW-DOWN condition; and $mean = 3.88, SD = 3.457, SE = 0.754$ in the LOSE condition. Statistically, there were no significant differences between these groups, $F(3, 79) = 0.302, p = 0.824$. The mean number of active verbs across WIN, DRAW-UP, DRAW-DOWN, and LOSE were 3.857, 3.214, 3.200, and 3.643, respectively. The percentage of verb phrases that were active were 94.7% in WIN, 94.4% in DRAW-UP, 96.2% in DRAW-DOWN, and 99.1% in LOSE. There was no statistical difference between the conditions in terms of the absolute number of active verb phrases, $F(3, 79) = 2.60, p = 0.854$, or the percentage of active verb phrases, $F(3, 79) = 0.398, p = 0.755$.

4. DISCUSSION

4.1 Agency of the Nao Robot

Despite one of the key findings in Short et al. being that participants use a higher proportion of active verbs when referring to cheating robots, we see no difference in this metric across the conditions in our study as participants almost exclusively used the active voice throughout their paragraph responses. Some possible reasons for this difference are that whereas Short et al. used the Nico robot, a comparably less agentic and less animate upper-torso robot, we used the Nao robot, a comparatively agentic, full-body humanoid robot. Whereas Short et al. did not exhibit any advanced behaviors to begin or end their study, the Nao commenced this study by standing up from a seated position and performing a waving motion while introducing itself, and also concluded the study by sitting down. These behaviors may have elevated the baseline agency of the robot, leading to a ceiling effect of participants exclusively using the active voice when referring to the Nao.

4.2 Dedicated Cheating Module

Humans possess a cheating detection module in the brain that has an evolved capability to automatically and effortlessly[13] detect cheating acts committed against them[2, 3, 14], which aids them in identifying individuals who violate social exchanges[4].

This study demonstrates high salience, verbal utterances usually in protest, and negative attributions towards an adversarial cheating robot, as well as the absence of such effects toward robots that cheat in a way not opposing the participant's self-interest. These findings are entirely congruent with one's expectations given the triggering of a cheating detector module by a robot. However, in order to expressly determine that this module causes these effects, further work must be done to support that the effects are effortless and automatic in the same way as Van Lier[13].

5. CONCLUSION

Much work has shown that deceptive behavior is a prevalent phenomenon between humans. Short et al. have shown that a cheating robot is conferred more attributions of mental state than a robot that does not cheat. In this study, we studied whether this result is due to the added motion of the cheat in that study, or due to an evolutionary cheating detector that Cosmides et al. have presented.

We performed a study to compare physical reactions, verbal reactions, and mental attributions toward robots that break the rules of a social game of rock paper scissors. We tested four conditions, all of which entailed the robot cheating, but varied in how adversarial the cheat was.

Our principal finding is that salience, engagement, and attributions vary as the direction and magnitude of the cheat changes. This rules out the hypothesis that the added motion of the "active cheat" in Short et al. causes mental attributions and supports the hypothesis that a cheating detector was triggered by the adversarial cheat of the robot.

Participants were more likely to self-report that the robot changed its gesture when the robot cheated maximally adversarially than in the less adversarial conditions. This asymmetry confirms that the complexity of motion is not what determines whether a cheating behavior is salient, but that the direction and magnitude of the cheat are critical.

Participants were also less likely to notice the gesture change at all when the robot cheated to let them win. Whereas 80% of participants noticed the robot's gesture change in the three least prosocial conditions, only 52% of participants did so in the most prosocial condition. These results can be explained as salience being driven by participants' detection of goal-driven behavior. The robot's goal could be self-interest in the cases where the robot cheats to improve its position, or saving the participant face in the case where the robot cheats to tie from a losing position.

Engagement, as measured by participant utterances, is modulated by the direction and magnitude of the cheat. Participants that go from winning to losing are by far the most likely to verbally protest. Participants that lose, but only slightly, are less likely to do so and participants that gain rather than lose from the cheat hardly speak at all.

The most adversarial and salient cheat type is also the only one that engenders lower attributions of honesty and fairness toward the robot. This attribution flies in the face of the fact that the robot breaks the rules in all conditions.

Many of these findings persist throughout tens of counter-examples throughout the study. Once a robot cheats against a human to its advantage, it is difficult, even impossible, to forget. Even with many counter-examples, at least ten before and ten after the two occurrences of the cheat events, participants remember incidents of a robot adversarially cheating them out of a win with high salience.

6. ACKNOWLEDGEMENTS

The authors would like to thank Eric Ho, Adisa Malik, Rebecca Marvin, André Pereira, Gabe Petegorsky, Sarah Strohkorb, Haohang Xu, and Jessica Yang for graciously donating their time to running the experiment, annotating participant data, and preparing the results. This work was supported by the National Science Foundation award #1117801 (Manipulating Perceptions of Robot Agency) and award #1139078 (Socially Assistive Robots).

7. REFERENCES

[1] W. A. Bainbridge, J. Hart, E. S. Kim, and B. Scassellati. The effect of presence on human-robot interaction. In *The 17th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2008)*, pages 701–706. IEEE, 2008.

[2] L. Cosmides. The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, 31(3):187–276, 1989.

[3] L. Cosmides and J. Tooby. Cognitive adaptations for social exchange. *The Adapted Mind*, pages 163–228, 1992.

[4] H. L. De Jong and W. Van der Steen. Biological thinking in evolutionary psychology: Rockbottom or quicksand? *Philosophical Psychology*, 11(2):183–205, 1998.

[5] F. Heider and M. Simmel. An experimental study of apparent behavior. *The American Journal of Psychology*, pages 243–259, 1944.

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

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

[8] B. J. Scholl and T. Gao. Perceiving animacy and intentionality: Visual processing or higher-level judgment. *Social perception: Detection and interpretation of animacy, agency, and intention*, 2013.

[9] B. J. Scholl and P. D. Tremoulet. Perceptual causality and animacy. *Trends in cognitive sciences*, 4(8):299–309, 2000.

[10] E. Short, J. Hart, M. Vu, and B. Scassellati. No fair!!: An interaction with a cheating robot. *Proceedings of the 5th ACM/IEEE International Conference on Human-robot Interaction*, pages 219–226, 2010.

[11] A. Steinfeld, O. Jenkins, and B. Scassellati. The oz of wizard: Simulating the human for interaction research. In *4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 101–107, March 2009.

[12] D. Ullman, I. Leite, J. Phillips, J. Kim-Cohen, and B. Scassellati. Smart human, smarter robot: How cheating affects perceptions of social agency. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society (CogSci2014)*, 2014.

[13] J. Van Lier, R. Revlin, and W. De Neys. Detecting cheaters without thinking: Testing the automaticity of the cheater detection module. *PloS one*, 8(1):e53827, 2013.

[14] J. Verplaetse, S. Vanneste, and J. Braeckman. You can judge a book by its cover: the sequel.: A kernel of truth in predictive cheating detection. *Evolution and Human Behavior*, 28(4):260 – 271, 2007.