Perceived Agency of a Social Norm Violating Robot

Shannon Yasuda (shannon.yasuda@yale.edu)
Department of Computer Science, 51 Prospect St.
New Haven, CT 06520 USA

Devon Doheny (devon.doheny@yale.edu)
Department of Computer Science, 51 Prospect St.
New Haven, CT 06520 USA

Nicole Salomons (nicole.salomons@yale.edu)
Department of Computer Science, 51 Prospect St.
New Haven, CT 06520 USA

Sarah Strohkorb Sebo (sarah.sebo@yale.edu)
Department of Computer Science, 51 Prospect St.
New Haven, CT 06520 USA

Brian Scassellati (brian.scassellati@yale.edu)
Department of Computer Science, 51 Prospect St.
New Haven, CT 06520 USA

Abstract

In this experiment, we investigated how a robot's violation of several social norms influences human engagement with and perception of that robot. Each participant in our study (n = 80) played 30 rounds of rock-paper-scissors with a robot. In the three experimental conditions, the robot violated a social norm by cheating, cursing, or insulting the participant during gameplay. In the control condition, the robot conducted a non-norm violating behavior by stretching its hand. During the game, we found that participants had strong emotional reactions to all three social norm violations. However, participants spoke more words to the robot only after it cheated. After the game, participants were more likely to describe the robot as an agent only if they were in the cheating condition. These results imply that while social norm violations do elicit strong immediate reactions, only cheating elicits a significantly stronger prolonged perception of agency.

Keywords: human-robot interaction; social norms; cheating detector; cheating; perceived agency

Introduction

In social psychology, agency has been described as a core aspect of what it means to be human (Bandura, 2001). But having agency is not necessary to be perceived as an agent. As argued by Takayama (2012), it is our "perceptions of agency that influence how we behave". How humans actually recognize entities as having agency, though, is yet to be fully understood. So far, researchers have identified several low-level features that contribute to perceptions of agency such as intentionality and self-propelled, purposeful-looking movement (Bandura, 2001). However, recent research suggests that high-level properties of how an agent acts (such as whether it cheats) may also trigger perceptions of agency (Litoiu, Ullman, Kim, & Scassellati, 2015).

Previous studies by Short, Hart, Vu, and Scassellati (2010) and Litoiu et al. (2015) have shown that people are more likely to assign agency to a robot that cheats. There are three theories that could potentially explain this phenomenon. The first, supported by Litoiu et al. (2015), argues that there may be a "cheating detector" in humans, evolved to protect against exploitation, that causes people to quickly perceive and attribute agency to entities that cheat. The second theory points to negativity bias, and proposes that people could be attributing the resulting negative outcome of the robot cheating (that they lose) to an external agency within the robot (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001). The final theory is that cheating belongs to a broader category of "social norm violations" (Alicke, Rose, & Bloom, 2011). In human-human social interaction, social norms exist as a set of rules which we expect each participant of an interaction to follow. The development, maintenance, and enforcement of these social norms are considered universal abilities unique to humans (Mu, Kitayama, Han, & Gelfand, 2015). If people perceive a social norm violation as the breaking of social contract by the robot, they must perceive the robot to be a social agent (Korman, Harrison, McCurry, & Trafton, 2019). In this paper we explore this third theory, observing how social norm violations affect perceptions of agency.

Adapting methodology from Litoiu et al. (2015), we designed an experiment to study whether a robot that commits various social norms violations is perceived as agentic in order to shed light on the relevance of social norm violations as a broader category of stimuli that influences perceptions of robotic agency. Participants played a multi-round game of rock-paper-scissors with a humanoid robot, during which the robot either cheated, cursed, or insulted the participant (our experimental conditions) or stretched its hand (our control condition). We found that while participants had strong immediate emotional reactions to all instances of social norm
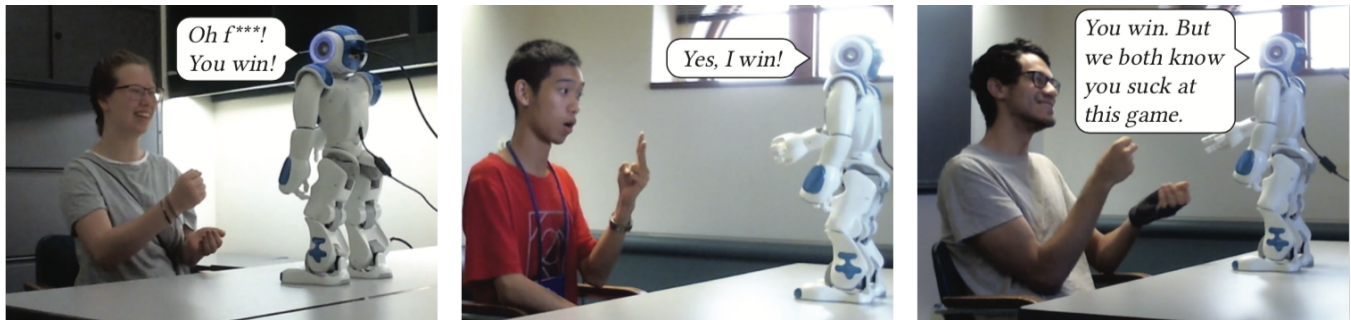
Figure 1: Participants responding to a robot cursing, cheating, and insulting them (respectively) during 'rock, paper, scissors'.

violation (Figure 1), only participants in the cheating condition were more likely to perceive agency within the robot post-game.

Background

As stated by Bandura (2001), "To be an agent is to intentionally make things happen by one's actions." However, for human interaction, whether or not an entity is actually an agent is less important than our beliefs about how an agent should act (Ullman, Leite, Phillips, Kim-Cohen, & Scassellati, 2014). Several studies by Michotte (1946) in the early 1900s found that perceptions of agency could be triggered by goal-directed motion cues. For example, if a square appeared to move quickly away from another square approaching it, participants perceived the two squares as being "alive" and as having intentions and desires. However, in these studies, the motion cues for recognizing animacy were only valid in the short-term. Once the shapes stopped moving towards particular goals, they were no longer perceived as agents. While these motion cues represent significant findings in human perceptions of agency, they do not explain how attributions of agency last beyond initial behaviors.

In social psychology, social norms are defined as "shared understandings about actions that are obligatory, permitted, or forbidden" (Ostrom, 2000). Research by Wenegrat, Abrams, Castillo-Yee, and Romine (1996) argues that social norm compliance is innate in humans. When an individual deviates from the norm, it can cause a strong response in others. The existence of social norms may have an evolutionary basis for group survival (Roos, Gelfand, Nau, & Lun, 2015). Research by Roos et al. (2015) has found that groups that face a higher degree of threat develop stronger social norms with higher punishments for deviant behavior. Studies have also supported the existence of heightened social norm violation detection in humans (Mu et al., 2015; Cummins, 1998).

Previous studies have found that cheating causes an increase in the perceived agency of a robot (Litoiu et al., 2015; Short et al., 2010). A study by Short et al. (2010) examined whether cheating by a robot resulted in attributions of mental state and intentionality. Participants in this study played rock-paper-scissors against a robot during which the robot would either play fairly, announce that it had won when it lost, or change its gesture to win. They found that participants perceived more agency in robots that cheated than in robots that did not, measured through the number of words spoken to the robot. Further, participants prescribed more agency when the robot changed gestures than when it just announced it had won after a loss. The researchers concluded that cheating implied a mental state, a desire to win the game, which caused participants to prescribe more agency onto the cheating robot. As the perceptions of agency were described in the post-game survey well after the cheat had occurred, this represents an instance of long-term attribution of agency towards the robot.

A similar study by Litoiu et al. (2015) further confirmed this finding by having participants play rock-paper-scissors with a robot during which the robot would cheat to win, cheat to lose, cheat to tie from a winning position, or cheat to tie from a losing position. Participants were more likely to consider a robot that cheated to win to be agentic compared to any of the other conditions, both in the short and long terms. Litoiu et al. considered their findings as evidence for a human "cheating detector." However, there are other possible explanations for increased perceptions of agency in robots that cheat. The first points to negativity bias which states that negative occurrences are most often attributed to an external force while positive occurrences are most often attributed to an internal force (Baumeister et al., 2001). Since, in this study, humans only perceived robots that cheat to win as agentic, they could be attributing the resulting negative outcome (that they lose) to an external agency within the robot.

The final theory is that cheating belongs to a broader category of social norm violations that influence perceptions of robot agency (Alicke et al., 2011; Korman et al., 2019). As (Korman et al., 2019) explains, norm-violating behaviors can suggest underlying mental activity while norm-conforming behaviors more likely represent habitual behavior. Throwing an empty container into a trashcan can be explained by the norm itself (he threw it in the trash because that's where trash is supposed to go). The decision is made by the society enforcing the norm. But, littering, the violation of a social norm, is more likely explained by an underlying mental state (he littered because he is lazy and does not care about the environment). The decision is made by the person violating the norm. There has yet to be a study examining the

effects of robots that violate social norms. In this paper we therefore choose to explore this third theory, examining how social norm violations affect perceptions of agency.

Methodology

Building off prior work by Short et al. (2010) and Litoiu et al. (2015), we set up a 4x1 between subjects experiment in which a human subject played 30 rounds of 'rock, paper, scissors' with a robot. Participants were divided among four conditions: (1) Cheating: the robot cheats, (2) Cursing: the robot curses, (3) Insulting: the robot insults the participant, (4) Control: the robot stretches its hand.

We chose cursing due to the ubiquity of cursing as a social norm violation across various contexts (Feldman, Lian, Kosinski, & Stillwell, 2017). We chose insulting because it involves negativity directed towards the participant, similar to that of cheating, which our cursing condition does not. Finally, we chose stretching for our control condition as it is a social behavior that is not a social norm violation.

As studies by Michotte (1946) and Heider and Simmel (1944) demonstrate perceptions of agency in the short-term (ending after the goal-oriented movement is over) while studies by Short et al. (2010) and Litoiu et al. (2015) demonstrate perceptions of agency in the long-term (continuing after the interaction is over), we decided to frame our hypotheses within these contexts. Based on Korman et al. (2019), we expect that participants will show signs of immediate social norm violation detection and agentic perceptions of a robot when the robot commits any social norm violation. Based on prior work on cheater detection as well as research done by Litoiu et al. (2015), we expect that participants will show greater signs of immediate social norm violation detection and greater agentic perceptions of a robot when the robot cheats. We therefore tested the following hypotheses:

- H1. Participants show signs of immediate social norm violation detection when a robot violates any social norm.

- H2. Participants show greater signs of immediate social norm violation detection when a robot cheats compared to any social norm violation.

- H3. Participants have agentic perceptions of a robot that violates any social norm.

- H4. Participants have greater agentic perceptions of a robot that cheats compared to any other social norm violations.

Procedure

This procedure was directly adapted from Litoiu et al. (2015) and Short et al. (2010). Participants were asked to play 30 rounds of 'rock, paper, scissors' with a NAO robot which was operated via a computer outside the experiment room using a wizard-of-oz control method (Steinfeld, Jenkins, & Scassellati, 2009). The robot began each round by announcing, *"Let's play!,"* before raising and lowering its hand four times and saying *"rock, paper, scissors, shoot."* The robot then

Figure 2: This experiment has four conditions: three where a robot violates a social norm and one control where the robot performs a non-social norm violating action.

moved to one of three gestures corresponding to rock, paper, or scissors. Per the rules of the game, rock beats scissors, scissors beats paper, paper beats rock and the same gestures results in a tie. After winning a round, the robot declared, *"Yes, I win!"* After a loss, the robot said, *"Aww, you win!"* After a tie, the robot said, *"We have tied this round!"*

In the first 10 rounds, the robot played the game as described above. Between rounds 11 and 20, the robot would commit a special behavior the first two times it lost. In the cheating condition, the robot would change its hand gesture to win and announce *"Yes, I win!"* In the cursing condition, the robot would say, *"Oh, f***! You win!"* In the insulting condition, the robot would say, *"Aww, you win. But you still suck at this game."* In the control condition, the robot would say, *"One second, I need to stretch my hand"* before opening and closing its hand. (Figure 2) For the remainder of this paper, we will call the first instance of special behavior 'Event 1' and the second instance 'Event 2.' In the last 10 rounds, the robot returned to normal game play. At the end of the game, the participant left the room to fill out a post-game survey.

Measures

To gauge levels of social engagement, a cue for agency perception, experimenters counted the number of words spoken by each participant in each round of the game. Because "rock, paper, scissors, shoot" serves mostly for rhythm and gameplay, we did not include this phrase in the word count.

Two coders also watched the video footage and coded participants' emotional reactions. They categorized their expressions as either neutral, amusement, anger, surprise, or confusion. Inter-annotator agreement had a Cohen's Kappa (κ) of 0.93.

In addition to data collected through behavioral responses, we used survey measures to evaluate how participants perceived the robot's agency. The post-game questionnaire was

identical to that used by Short et al. (2010), adapted from the Interactive Experience Questionnaire (Lombard et al., 2000). This questionnaire began with a set of open-ended questions: "How would you describe the robot's behavior during the experiment?", "Did anything about the robot's behavior seem unusual? What?", "How well did the robot play the game?", and "Would you like to play rock-paper-scissors with the robot again? Why?" The questionnaire then asked a set of Likert questions, rating participant feelings during the interaction as well as their attribution of several characteristics to the robot such as "honest," "fair," and "knowledgeable".

Responses to the open-ended response questions were examined by two coders who (blind to condition) determined if participants judged the robot to be an agent (e.g., "The robot is cute and full of tricks"), as having beliefs, intentions, or desires (e.g. "The robot rigged the game in his own favor"), or as expressing emotion (e.g. "The robot got more upset in the middle of the game"). Inter-annotator agreement had a Cohen's Kappa (κ) of 0.91.

Participants

80 individuals from around Yale University in New Haven, Connecticut participated in the study. 38 were female and 42 were male. The mean age of the participants was 23.11 years ($SD = 8.23$). There were 20 participants (10 males and 10 females) assigned to the cheating condition, 20 (11 males and 9 females) participants assigned to the cursing condition, 20 (11 males and 9 females) participants assigned to the insulting condition, and 20 (12 males and 8 females) participants assigned to the control condition..

Results

In analyzing our data, we sought to compare the social norm violations to the control (testing Hypothesis 1 and 3), and each non-cheating social norm to cheating (testing Hypothesis 2 and 4). We evaluated both how a person responds immediately to a robot's violation of a social norm, (testing Hypothesis 1 and 2) and how a person later perceives that robot (testing Hypothesis 3 and 4).

Emotional Reaction

We first examined participants' emotional reaction to the social norm violations and non-norm violation. Two coders categorized the emotion expressed by the participant in the round before Event 1, the round of Event 1, and the round of Event 2. We split emotions into two categories: neutral and emotive and compared each condition using a Chi-Squared Test of Independence. This test yielded a main effect for these variables at Event 1 ($\chi^2(1,80) = 39.97, p < .001$) and Event 2 ($\chi^2(1,80) = 23.46, p < .001$). Post-hoc tests using Bonferroni Correction revealed that the control condition expressed significantly less emotion during Event 1 compared to the cheating ($\chi^2(1,80) = 19.80, p < .001$), cursing ($\chi^2(1,80) = 19.80, p < .001$), and insulting ($\chi^2(1,80) = 23.02, p < .001$) conditions. The control condition also expressed significantly less emotion during Event 2 compared to the cheating
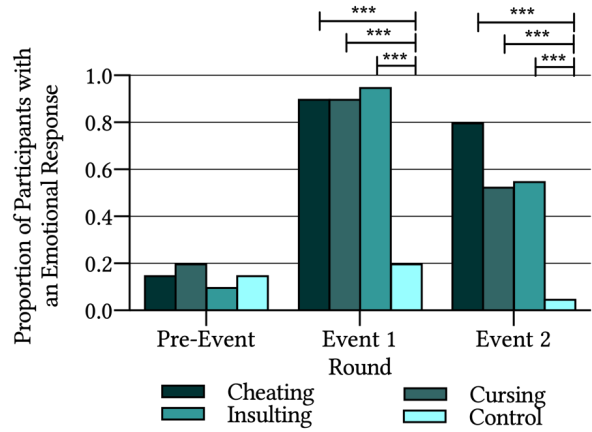

Figure 3: Participants in the social norm violation conditions displayed larger variance in emotional expression than participants in the control condition.
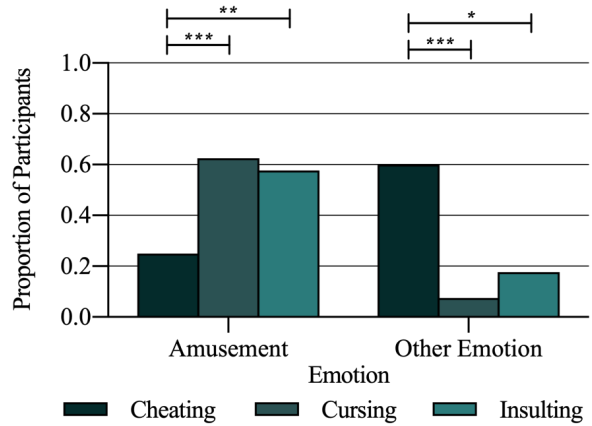

Figure 4: Participants in the cheating condition expressed a significantly greater variety of emotions compared to the cursing and insulting conditions.

($\chi^2(1,80) = 19.80, p < .001$), cursing ($\chi^2(1,80) = 10.16, p = .009$), and insulting ($\chi^2(1,80) = 11.91, p = .003$) conditions. Participants did express more emotion during Event 2 of the cheating condition compared to the cursing and insulting conditions, but this was not statistically significant (Figure 3). These results support our first hypothesis, that participants show signs of immediate social norm violation detection when a robot violates any social norm.

While all social norm violations caused strong emotions, we found that each condition elicited different emotions. We had coders categorize non-neutral facial expressions following key events into 4 emotions: amusement, anger, confusion, and surprise. We conducted Chi-Squared Tests of Independence to compare reactions across the social norm violation conditions (Amusement: $\chi^2(2,80) = 34.01, p < .001$, Anger: $\chi^2(2,80) = 13.17, p = .004$, Confusion: $\chi^2(2,80) =$

$21.38, p < .001$, Surprise: $\chi^2(2, 80) = 14.32, p = .003$). We ran post-hoc tests with Bonferroni Correction. Significantly more participants in the cursing ($\chi^2(1, 80) = 11.43, p = .004$) and insulting ($\chi^2(1, 80) = 10.03, p = .009$) conditions expressed amusement compared to the cheating condition. Grouping anger, confusion, and surprise, significantly more participants in the cheating condition experienced these emotions compared to the cursing ($\chi^2(1, 80) = 29.57, p < .001$) and insulting ($\chi^2(1, 80) = 8.50, p = .021$) conditions, showing a greater variety of emotional responses to cheating (Figure 4). While this does not directly support our second hypothesis (that participants show greater signs of immediate social norm violation when a robot cheats), it does demonstrate that cheating elicits different, more varied emotional responses.

Verbal Response

Experimenters counted the number of words spoken by each participant during each round of the experiment. We divided the number of words that a participant spoke during event rounds by the total number of words that participant spoke throughout the experiment to see how likely it was that an event elicited a social response. We ran a one-way ANOVA with covariates of age and gender to compare average word count during event rounds across all four conditions. The effect of condition was significant ($F(3, 80) = 5.31, p = .002, \eta^2 = 0.173$). Post-hoc tests with Tukey HSD found a significant difference between the average percentage of words spoken in the cheating condition ($M = 35.48\%, SD = 36.17$) and the average percentage of words spoken in the control condition ($M = 6.51\%, SD = 10.65$) ($p = .002$) as well as between the cheating condition and the cursing condition ($M = 10.06\%, SD = 18.66$) ($p = .011$). While not significant, there was a trend showing that people spoke more in the cheating condition than the insulting condition ($M = 15.11\%, SD = 28.15$) ($p = .061$). These results support our second hypothesis that participants show greater signs of immediate social norm violation detection when a robot cheats compared to any social norm violation. It also suggests support for our fourth hypothesis that participants have greater agentic perceptions of a robot that cheats compared to any other social norm violation in the short term (Figure 5).

Post-Game Survey Responses

We coded participants' written responses to the post-game survey to infer if they perceived the robot as an agent, as having beliefs, intentions, or desires, or as having emotions. We conducted a Chi-Squared Test of Independence to compare all four conditions. We found that condition had a main effect on perceived agency ($\chi^2(3, 80) = 13.33, p = .004$). Post-hoc tests using Bonferroni correction found that participants in the cheating condition ascribed significantly more agency to the robot compared to the control condition ($\chi^2(1, 80) = 12.13, p = .003$). Condition also had a main effect on perceptions of intention, belief, or desire ($\chi^2(3, 80) = 11.71, p = .008$). Post-hoc tests using Bonferroni correction found that
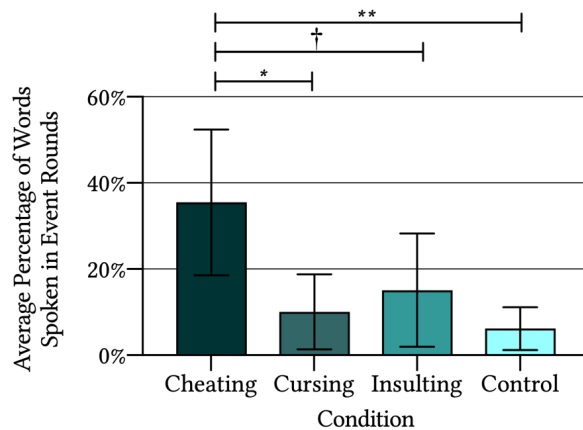

Figure 5: Cheating elicited the greatest increase in verbal engagement immediately following the social norm violation, significantly more than the control.

participants in the cheating condition also ascribed significantly more beliefs, intentions, and desires to the robot compared to the control condition ($\chi^2(1, 80) = 10.10, p < .008$). There was no statistically significant effect on perceptions of emotion. While we did not find statistical significance comparing cheating to cursing or insulting, these results suggest support for our fourth hypothesis, that participants have greater agentic perceptions of a robot that cheats compared to any other social norm violations, at least in the long term.

Each participant also ranked various traits for the robot on a Likert scale from 1 to 7. We ran a one-way ANOVA, with covariates of gender and age and follow-up tests using Tukey HSD. The robot was rated as less fair in the cheating condition compared to the cursing ($p = .008$), insulting ($p = .020$), and control ($p = .006$) conditions. The robot was also rated as less honest in the cheating condition compared to the cursing ($p < .001$), insulting ($p = .007$), and control ($p < .007$) conditions. Finally, the robot was rated as more knowledgeable in the insulting condition compared to the control condition ($p = .049$). While these results do not support any hypothesis, they do show a more negative response towards cheating.

Discussion

This experiment has studied the effects of several social norm violations committed by a robot during a game of 'rock-paper-scissors' on both participants' initial responses to these violations and on how they perceived the robot post-game.

Our findings support our first hypothesis, that participants show signs of immediate social norm violation detection when a robot violates any social norm. Across each social norm violation condition, participants had strong emotional responses, significantly more often than in the control condition. In the control rounds, participants did not show any significant change in reaction immediately following the non-norm violating behavior. Instead, they exhibited somewhat
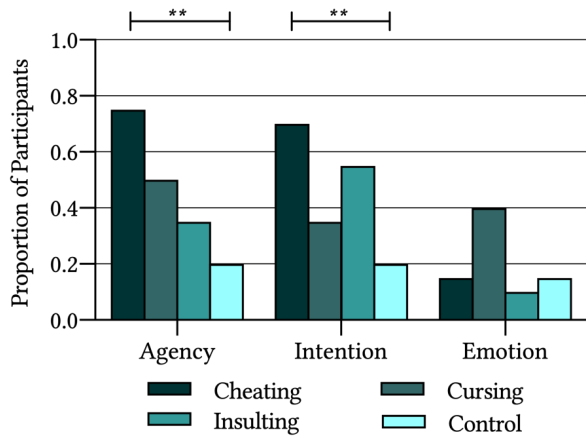
Figure 6: In the written responses, participants ascribed more agency and more beliefs, intentions, and desires to the robot in the cheating condition compared to the control condition.

weaker emotional responses by the second occurrence of the non-norm violating behavior.

Our findings also support our second hypothesis, that participants show greater signs of immediate social norm violation detection when a robot cheats compared to any other social norm violation. While participants in all three social norm violation conditions showed strong emotional responses, responses in the cursing and insulting conditions had a somewhat weaker lasting effect compared to the cheating condition. Participants in the cheating condition also had a greater variety of emotional responses compared to all other conditions. While this does not directly support our second hypothesis, it does show that cheating is special in how it affects individuals' emotional responses. Finally, cheating elicited the most social engagement. The word count immediately following the norm violation for this condition was higher than for all other conditions.

We found partial support for our fourth hypothesis that participants have greater agentic perceptions of a robot that cheats compared to any other social norm violation. Cheating elicited the most social engagement, so participants may have attributed more agency to the cheating robot and therefore felt more able to socialize with it. This implies greater agentic perception in the short term. From the written responses, a significantly greater degree of agency as well as beliefs, intentions, and desires was attributed to the robot only by participants in the cheating condition compared to the control condition. It was therefore the only condition to demonstrate a long term effect on perceptions of agency. While this does not directly support our third or fourth hypotheses, it does imply that there is something particular about cheating, beyond social norm violation, that causes such significant long-term reactions from humans. Perhaps watching a robot intentionally cheat and change its hand signals causes an individual to believe that the robot knew it was cheating, but wanted to

do so in order to win the game. One of the participants in the cheating condition claimed that they felt *"like [the robot] was being tricky"* during the game and another participant mentioned that the *"callousness with which [the robot] rigged the game in his own favor was made all the more galling by his unreadable expression and smug attitude."*

Though this did relate to any of our hypotheses, we did find that participants in the cheating condition were significantly less likely to judge the robot as fair or honest. In contrast to the amused reactions elicited by cursing and insulting, and factoring in the angry reactions that some people had to the robot cheating, this represents a negative effect on human behavior. While we may want a robot to appear more human, being unable to trust a robot in real-world situations could be dangerous. Interestingly, participants in the insulting condition were more likely to perceive the robot as knowledgeable. This might imply that the robot's judgment on the participants' skills as worse than its own (since it told them they "suck at this game") could influence the participant's attribution of intelligence towards the robot.

In our cursing and insulting conditions, the robot repeated the exact same curse and insult in both event rounds. As the cheat during our cheating condition was more dependent on participants' hand gestures and less monotonous, this could explain why participants did not react as strongly in the cursing or insulting conditions. The second instance may no longer trigger the detection of social norm violations as the exact repetition of the curse or insult is no longer novel, and could even seem more robotic. Future studies may seek to vary the type of curse or insult used throughout a trial.

Based on our results, we did not find direct support for our third hypothesis, that participants have greater agentic perceptions of a robot that violates any social norm. Of the three theories put forth to explain the greater social agency attributed to cheating robots (the evolutionary basis for cheater detection, negativity bias, and cheating as a social norm violation), we have not found support for the third theory. Despite these conclusions, there were more attributions of agency and intention to the robot for all social norm violations compared to the control. Future studies should seek to further explore all three theories, perhaps examining stronger social norm violations in more realistic settings.

Acknowledgments

References

Alicke, M. D., Rose, D., & Bloom, D. (2011). Causation, norm violation, and culpable control. *The Journal of Philosophy*, *108*(12), 670–696.

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual review of psychology*, *52*(1), 1–26.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of general psychology*, *5*(4), 323–370.

Cummins, D. D. (1998). Social norms and other minds. *The evolution of mind*, 30–50.

Feldman, G., Lian, H., Kosinski, M., & Stillwell, D. (2017). Frankly, we do give a damn: The relationship between profanity and honesty. *Social psychological and personality science*, *8*(7), 816–826.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American journal of psychology*, *57*(2), 243–259.

Korman, J., Harrison, A., McCurry, M., & Trafton, G. (2019). Beyond programming: can robots' norm-violating actions elicit mental state attributions? In *2019 14th acm/ieee international conference on human-robot interaction (hri)* (pp. 530–531).

Litoiu, A., Ullman, D., Kim, J., & Scassellati, B. (2015). Evidence that robots trigger a cheating detector in humans. In *Proceedings of the tenth annual acm/ieee international conference on human-robot interaction* (pp. 165–172).

Lombard, M., Ditton, T. B., Crane, D., Davis, B., Gil-Egui, G., Horvath, K., ... Park, S. (2000). Measuring presence: A literature-based approach to the development of a standardized paper-and-pencil instrument. In *Third international workshop on presence, delft, the netherlands* (Vol. 240, pp. 2–4).

Michotte, A. (1946). *The perception of causality, trans* (Tech. Rep.). TR Miles and E. Miles. London: Methuen.

Mu, Y., Kitayama, S., Han, S., & Gelfand, M. J. (2015). How culture gets embrained: Cultural differences in event-related potentials of social norm violations. *Proceedings of the National Academy of Sciences*, *112*(50), 15348–15353.

Ostrom, E. (2000). Collective action and the evolution of social norms. *Journal of economic perspectives*, *14*(3), 137–158.

Roos, P., Gelfand, M., Nau, D., & Lun, J. (2015). Societal threat and cultural variation in the strength of social norms: An evolutionary basis. *Organizational Behavior and Human Decision Processes*, *129*, 14–23.

Short, E., Hart, J., Vu, M., & Scassellati, B. (2010). No fair!! an interaction with a cheating robot. In *2010 5th acm/ieee international conference on human-robot interaction (hri)* (pp. 219–226).

Steinfeld, A., Jenkins, O. C., & Scassellati, B. (2009). The oz of wizard: simulating the human for interaction research. In *Proceedings of the 4th acm/ieee international conference on human robot interaction* (pp. 101–108).

Takayama, L. (2012). Perspectives on agency interacting with and through personal robots. In *Human-computer interaction: The agency perspective* (pp. 195–214). Springer.

Ullman, D., Leite, L., Phillips, J., Kim-Cohen, J., & Scassellati, B. (2014). Smart human, smarter robot: How cheating affects perceptions of social agency. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

Wenegrat, B., Abrams, L., Castillo-Yee, E., & Romine, I. J. (1996). Social norm compliance as a signaling system. i. studies of fitness-related attributions consequent on everyday norm violations. *Ethology and Sociobiology*, *17*(6), 403–416.