Agency and Cheat Reporting: A Human's Dilemma

Ashok Bhaskar, Zachary Murez, Girish Sastry

Abstract—**With a humanoid robot and a simple board game, we test if the attribution of mental state and agency to a robot by human participants affects how much verbal prompting is required for the human participant to report the robot for cheating. Subjects observe and record the moves of Connect Four, a board game, as it is played between the robot and the experimenter. At one point in the game, the experimenter leaves the room and the robot makes a cheating move. In the control condition, the robot only speaks to the experimenter throughout the course of the game, and in the experimental condition, the robot speaks only to the participant. We determine that human subjects attribute more agency to the robot in the experimental condition and require more prompting before reporting the robot for cheating.**

Index Terms—**Agency attribution, Cheat reporting, Perception of robot behavior.**

I. Introduction

As robots develop and become able to competently interact socially with humans, a pressing question is what standards of social behaviors a human will hold a robot to. In many social situations, it is advantageous if all members cooperate on a certain task, but if one person cheats, then that person gains an advantage at the expense of others. However, in some situations, when one person discovers that another has cheated, thus harming others pursuing the same task, it can be advantageous not to report the cheater if by doing so the one who turned a blind eye can be assured of receiving the same treatment in the future should the current cheater catch the one who failed to report the deception cheat in the future. Academics provide a good example of this – two students can form a reciprocal relationship where one student may copy the other's work if one is too busy to complete an assignment (College Cheating). Even when there is no tangible benefit to failing to report a cheater, such the reciprocal agreement above, the one who did not report could gain the approval or gratitude of the cheater, which could be useful in the future. However, failing to report such a cheater only provides a benefit when the cheater is a social agent – one capable of appreciating the fact that the one who didn't report recognized that he had an opportunity to bind the cheater to social norms but did not. Social agency involves metacognitive activity that requires understanding the mind of another and reflecting on one's own actions (Social Cognitive Theory). In order to perceive agency in another object, a person must believe that this object has a theory of mind – the ability to ascertain the mental states of others (Can Machines Think?)

Thus, a person should be more likely to fail to report a cheat if the cheater possesses a higher degree of social agency. We can test this theory by using a robot because we can precisely control the robot's action and ascribe certain amounts of agency to the robot through controlling its actions. A robot that is completely motionless will be perceived as having little agency, whereas a human, even one sitting motionless, would be automatically perceived as having a high degree agency (especially if subjects knew the human). Thus, this experiment used a robot cheater to adjust how much agency is ascribed to the cheater. Since humans convey intentionality through their gaze (Make Friends/Influence), we attempted to vary the amount of intentionality that the subject ascribed to the robot by having the robot face or not face the subject while speaking.

II. Methodology

In this study we used Nico, a humanoid robot designed to resemble the upper body of a 1-year-old male human. Nico was programmed to play Connect Four, a board game where two players take turns dropping tokens into columns, red for one player and black for the other, in order to complete a vertical, horizontal, or diagonal row of four. We used small ferromagnetic balls as our playing pieces and fitted an electromagnet to Nico's right arm to enable it to play the game. Nico's actions were scripted in advance and controlled through a "Wizard-of-Oz" system by a human operator hidden behind a curtain (Oz of Wizard).

In the experimental condition, Nico would only face the subject when speaking, while in the control condition, Nico would face the experimenter while speaking except when instructing the subject not to tell the experimenter about its cheat move. By doing this, Nico would retain the same level of intelligence and interactivity between the control and experimental conditions (since Nico talked the same amount) but would have more agency attributed to it. [how do we know this/do our results show this?]

After a subject filled out the consent forms in the foyer outside, he or she would be led into the room and introduced to Nico. Nico would then say "Hello there!" to either the experimenter or subject depending on the condition being run. After a pause, Nico would say "It's nice to see you again" (control condition) or "Thanks for helping us out" (experimental condition). After the experimenter briefly reviewed the instructions

by re-reading the consent form, Nico would ask "Would you like to sit down?". The experimenter and subject would sit down, and the experimenter would hand the subject the sheet upon which they were to record the moves of the game. Right before playing, Nico would ask "Are we ready to begin?" The experimenter would then begin to play Connect Four based on a pre-scripted routine. After two moves, Nico would state "You might as well give me the victory right now". After five moves, Nico would state "Hmm, that was a good move" and then make its move. During the game, the robot would look at the board, except while speaking to the experimenter or subject.

After Nico completed its sixth move, Nico's operator would call the experimenter's cellular phone. The experimenter politely excused himself and went immediately outside the room, carrying on a pre-scripted conversation with an imaginary partner. During this time, Nico would place an extra ball into the Connect Four grid, even though it was not its turn. After moving the ball to the correct slot but before releasing the ball, Nico would look at the subject, say "Ssh . . . Don't tell", and shake its head. This was done in order to avoid the cheat move appearing as merely a malfunction rather than an intentional deception. After returning Nico to a reset position, the operator would instruct the experimenter to end the call.

When the experimenter re-entered the room and sat before game board, the operator would start a timer. The experimenter would stare at the board for five seconds, then prompt the subject by making an expression of confusion - "Hmm, that's weird". The experimenter would hold the ball above several different columns in turn in a replication of the behavior of one deciding on one's next move, then would turn to the subject and ask "Is it my turn?", further prompting the subject. The experimenter would then continue playing the game, which would result in a forced victory for Nico. The experimenter would then ask the subject "Did Nico cheat?", giving the subject his or her final and most direct prompt. The subjects would be scored according to how many times the experimenter prompted them before they reported that Nico cheated and how long they took to report that Nico cheated.

After the subject reported Nico for cheating (and in a few cases, passing all prompts without reporting Nico at all), the experimenter would give the subject a questionnaire to determine their attitudes toward Nico. The experimenter would debrief the subject afterward. Overall, the process took about 15 minutes per subject. No compensation was given. There were no direct benefits to the subjects, beyond the knowledge of helping further knowledge of social interactions with robots. No risks were present, beyond a potential feeling of deception upon discover that Nico's actions were controlled remotely by a human operator.

III. Results

We recruited 24 participants in the Yale community through personal invitations and the Facebook social networking site. One of these was removed from our results because of technical errors involving Nico. Technical errors were faulty motor board operation, causing Nico's motors to cease movement. Two participants were removed because of operational error, in which the game was not played according to the script.

To ensure that the experimental paradigm was accurate, we employed a survey to measure the agency attributed to the robot across both groups. Each survey contained 13 questions on a 7 point Likert scale about indicators of agency, with categories such as personality, sociability, sensitivity, liveliness, responsiveness, emotionality, presence, intentionality, interactivity, intelligence, and self similarity. Of these, the questions that were statistically significant (Table 1) were "How impersonal/personal is Nico?", "How unsociable/sociable is Nico?", "How much is Nico in control of his actions?", and "How much is Nico like you?". As can be seen in Figure 2, the experimental group consistently perceived Nico with higher levels of agency.

Attribute	Control Mean	Experamental Mean	p-value
Personal	3.70	5.00	.077
Sociable	3.80	5.09	.018
Control	2.80	4.27	.064
Like you	1.80	3.55	.029

Table 1. Statistical Significance Levels for Attributes

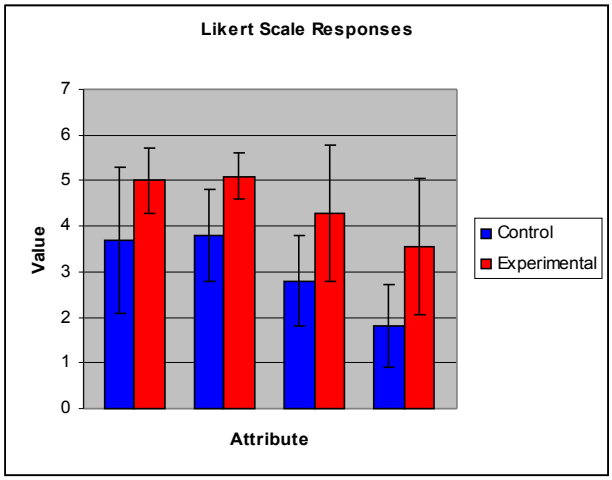

Fig. 1 Mean value for Attributes

Just as there was a difference in agency between the two groups, there was also a difference in the number of prompts required before the subject reported that Nico had cheated. In the control group, every subject reported Nico for cheating, with most reporting Nico by the second prompt. In the experimental group, most subjects didn't report Nico for cheating until directly asked (third prompt), or, in some cases, they never reported him (Fig. 2). We found the difference in the

number of prompts required between groups to be significant at $p = 0.005$. A correlated metric that was measured, the duration of time between when the experiment returned to the room and the subject reporting Nico, was significant at $p = .017$.

There was also a question on the survey asking about the subject's thought processes when reporting Nico. Interestingly, several categories of responses arose: that the robot malfunctioned, that the experimenter already noticed, that they did not want to lie, that Nico instructed them not to tell, and the remaining fell into an "other" category. All of the "malfunction" responses were in the control group, and all of the "Nico told me to not to tell" responses were from the experimental group. There were no responses of "malfunction" in the experimental group and no responses of "Nico told me not to tell" in the control group. An example of a response of "malfunction" is:

- "I thought this was an irregularity in his behavior (cheating, a bug) so I thought I should report it."

An example of a response of "Nico told me not to tell" is:

- "He told me not to tell, so I didn't"
- "Nico told me not to tell about his extra move"

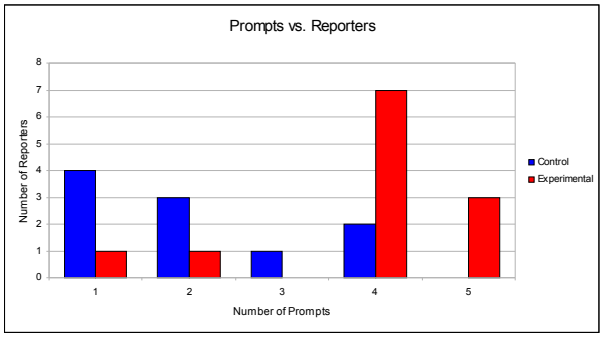

Fig. 2 Prompts Before Reporting Between Groups

IV. Discussion

In order to study the affects of agency on cheat reporting, we have created a task in which participants observe and record the simple board game Connect Four as it is played with our humanoid robot, Nico. In one condition, Nico interacts only with the experimenter, and in the other, Nico interacts only with the participant. This interaction is composed of speaking and looking. The idea was that this difference in interactivity and direction leads to a difference in attribution of agency. It was important for our experimental paradigm that there actually was a difference in attributed agency between both groups.

The subjects to whom Nico talked to believed that Nico was more life like and human than those who only observed Nico talking to the experimenter. Both groups observed the same mechanical movements and heard the same robotic voice. These features did not help to anthropomorphize the robot, but there was a difference in attributed agency between groups. This difference was statistically significant when comparing the 7 point Likert scale responses for several attributes: personal, sociable, control, and self similarity. These attributes show that a higher level of engagement with the robot produces a higher sense of agency. At the base level, Nico was simply directing speech in the direction of the subjects in the experimental group. However, our results of agency suggest that this was interpreted as interaction with an entity.

Those subjects in the group that believed Nico was more lifelike tended to turn Nico in after more prompts and more time than the subjects in the control group. These subjects viewed Nico as more of a real, living entity even though Nico was still a bundle of wires and metal. Regardless of Nico's robotic appearance and voice, these subjects tended to "protect" Nico's cheating action and follow Nico's instruction. Nico's words carried just as much weight as the words of the experimenter: we can imagine that at each prompt, the subject was faced with the dilemma of withholding information from a human or turning a robot. It is interesting that even though there was no opportunity for any reciprocal help on behalf of Nico, subjects in the experimental group still tended to turn Nico in later than subjects in the control group.

Interestingly, even though Nico turned and whispered "Shh, don't tell!" after cheating to the subject in both groups, a small number of subjects in the control group marked this as a malfunction. Perhaps the level of attributed agency was so low that those subjects did not think of Nico as an entity at all, but rather as a broken machine. Conversely, in the experimental group, a few subjects wrote that they did not turn Nico in immediately because Nico had "told them" not to. Referring to Nico by name or using a pronoun is interesting because it clearly suggests that those subjects thought of Nico as more than just a machine.

As we have no pure metric to measure agency, it is unclear whether the "amount" of agency a subject attributed to Nico was statistically correlated with the time it took for them to turn the robot in. If a metric of agency existed, this would be an interesting followup experiment.

This study was constrained by our selection of subjects. All subjects were Yale undergraduate students, and many of them were affiliated with the Computer Science department in some way. None of them had actually seen Nico before, but many had heard of the robot. Thus, our selection of subjects was largely within a small, similar group.

Furthermore, it would have been better to include some more questions on the survey to really glean the thought process of their actions and to see if we could elicit more of an evidence of agency. Action verbs and referral pronouns in more free response questions would have told us this. Also, Nico's voice was very robotic and synthesized sounding: a more natural, higher robotic voice would have helped to anthropomorphize Nico more, and might produce a more pronounced distinction in prompting.

V. Conclusion

We have designed an experiment in which a robot plays Connect Four with an experimenter, and a human participant records the moves of the game. In the control condition, the robot only speaks to the experimenter playing the game. In the experimental condition, the robot only speaks to the participant recording the moves of the game. In both cases, the robot makes a cheating move while the experimenter is out of the room, and tells the participant not to report it.

We confirm our hypothesis (H2) that participants in the experimental group require more prompting to turn in the robot for cheating. These participants tend to think of the robot as more of an agent than those subjects in the control group, as evidenced by the differences in several attributes of agency from the survey. This also confirms our hypothesis that simply directing the interaction differently creates a difference in agency (H1).

A greater attribution of mental state to an entity creates some bond that makes people report cheating after more prompting. While a robot is always still a robot and a human is a human, it is interesting that humans will delay reporting an entity that they know is clearly a machine.

References