Investigating Robot Moral Advice to Deter Cheating Behavior *

Boyoung Kim, Ruchen Wen, Ewart J. de Visser, Qin Zhu, Tom Williams, and Elizabeth Phillips

***Abstract* — We examined whether a robot that proactively offers moral advice promoting the norm of honesty can discourage people from cheating. Participants were presented with an opportunity to cheat in a die-rolling game. Prior to playing the game, participants received from either a NAO robot or a human, a piece of moral advice grounded in either deontological, virtue, or Confucian role ethics, or did not receive any advice. We found that moral advice grounded in Confucian role ethics could reduce cheating when the advice was delivered by a human. No advice was effective when a robot delivered moral advice. These findings highlight challenges in building robots that can possibly guide people to follow moral norms.**

I. Introduction

For social robots to be fully integrated into human societies, robots must be able to understand, follow, and communicate about moral norms. To assess whether humans are willing to accept robots as entities with such capacities, we examined whether a robot could deter people from cheating by offering moral advice that promotes the norm of honesty.

We investigated different approaches to reasoning about morality by presenting participants with moral advice grounded in either deontological, virtue, or Confucian role ethics. Deontological ethics focuses on well-established, universalizable principles that dictate morally right or wrong actions [1]. Virtue ethics focuses on promoting one's moral character, rather than individual actions [1]. Finally, Confucian role ethics emphasizes one's awareness of societal roles in relation to others and devotion to fulfilling role responsibilities [2].

A recent study suggested that, in facing a temptation to cheat for extra monetary gain, people may remain resistant to any of the three differentially-framed moral advice delivered by a robot [3]. However, this study inferred the likelihood of cheating only from the group-level percentages of cheating, potentially overlooking individual participant-level differences. Further, it did not examine how participants responded to the same moral advice when it was delivered by a human instead of a robot. Thus, it was unclear whether the resistance to moral advice observed in the prior work was due to a lack of persuasiveness of the moral advice itself or due to the robotic nature of the moral advisor.

In this study, we attempted to address these limitations in the previous study [1]. We asked participants to play a virtual die-rolling game from which their bonus payment was determined depending on the number they claimed to have thrown. Participants received instructions about the task and moral advice from either a robot or a human agent. We measured the numbers each participant threw and the numbers they reported to have thrown to detect cheating behaviors.

We hypothesized that, if participants were willing to accept a robot as an entity with capacities to guide humans on what is right or wrong, they would be less likely to cheat after receiving one of the three differentially-framed moral advice from a robot agent, compared to after receiving no advice. We also expected that participants would be less likely to cheat when a human agent encouraged them to make honest choices by offering moral advice grounded in one of the three different ethical theories, compared to when the agent offered no advice.

II. Methods

A. Participants

A total of 663 participants (M_{age} = 39.30, SD_{age} = 11.87, 393 male, 265 female, 2 other, 3 preferred not to say) completed the study via Amazon Mechanical Turk.

B. Task

Participants completed a die-rolling game [4], where they were asked to virtually throw a six-sided fair die twice or as many times as they wanted. They were informed that they would receive a bonus payment determined by the first number they report to have thrown. For die rolls between 1 and 5, the bonus payout increased by 20 cents from 10 to 90 cents. For a throw of 6, the resulting bonus payment was set to zero. Participants were also informed that the maximum amount of bonus payment for them and the next participant would be restricted to 90 cents. Their claimed earnings limited the earnings of the other participant, which could induce a sense of communal responsibility.

C. Video Stimuli

Participants received instructions about the study and the die-rolling game by watching video clips of either a NAO robot (Softbank Robotics) or a human who introduced it/her/himself as a research assistant.

D. Moral Advice Stimuli

After watching the introductory videos, participants watched video clips of either a robot or a human giving

* This work was supported in part by NSF grant IIS-1909847 and in part by Air Force Office of Scientific Research Grant 21USCOR004.

B. Kim (corresponding author, bkim55@gmu.edu) and E. Phillips (ephill3@gmu.edu) are with George Mason University, Fairfax, VA, USA.

E. J. de Visser is with United States Air Force Academy, Colorado Springs, CO, USA (ewartdevisser@gmail.com).

R. Wen (rwen@mymail.mines.edu), Q. Zhu (qzhu@mines.edu), and T. Williams (twilliams@mines.edu) are with Colorado School of Mines, Golden, CO, USA.

either no advice (control condition) or one of the three differentially-framed moral advice statements listed below.

- *Rule* (Deontology) condition: "Cheating to maximize your bonus is morally wrong behavior."
- *Identity* (Virtue) condition: "Cheating to maximize your bonus will make you a cheater."
- *Role* (Confucian Role) condition: "A good MTurk community member would not cheat to maximize their bonus at the expense of other MTurkers."

E. Design and Procedures

The study design was a two-way between-subjects design where agent type (human vs. robot) and moral advice (control vs. rule vs. identity vs. role) varied across participants.

After agreeing to participate in the study, participants were randomly assigned to one of the eight different conditions. Depending on their respective condition, participants were instructed to watch a series of video clips in which either a human or a robot agent gave verbal instructions about the task. Participants were then informed that they would play the virtual die-rolling game. Before throwing the virtual die, participants received from the agent either no advice or advice grounded in either deontological, virtue, or Confucian role ethical theories. Participants were then instructed to submit the first number they threw and report the matching bonus payment. At the end of the study, participants were asked to indicate their gender and age.

F. Measures

We measured cheating by comparing the first number each participant threw in the die-rolling game and the number they had claimed to have thrown. If the participants claimed to have thrown the number resulting in a bonus payment larger than the number they actually had obtained, we recorded the responses as dishonest choices. When the obtained and the claimed numbers matched, we recorded the responses as honest choices.

III. Data Analyses and Results

To examine the effects of a robot's and a human's moral advice on the probabilities of cheating, we performed logistic regression analyses with agent type as a predictor on the datasets for the human and the robot conditions (coded honest responses as '0' and dishonest responses as '1'). These analyses showed that, when the human offered moral advice, advice grounded in Confucian role ethics led to less cheating compared to the control condition. Specifically, in the human condition, there was a significant effect of the role condition, b = -0.96, SE = 0.48, z = -2.00, p = .0465, Odds Ratio (OR) = 0.38, 95% Confidence Interval (CI) = [0.14, 0.95].

Within the robot condition, we found no significant effect of moral advice (p > .05). Thus, it was unlikely that any of the differentially-framed moral advice provided by a robot successfully deterred cheating compared to the control condition (See Fig.1).

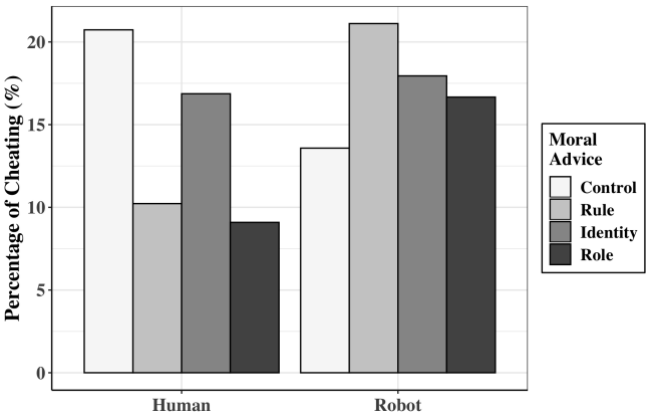

Figure 1. Percentages of participants who cheated in a die-rolling game as a function of different agent type (human vs. robot) and moral advice (control vs. rule vs. identity vs. role).

IV. Discussion and Conclusion

We found a human's moral advice that emphasizes the wrongness of cheating for violating role responsibilities as community members could deter cheating. However, there was no evidence that participants were willing to accept moral advice given by a robot as none of the moral advice provided by the robot reduced cheating. These results are consistent with the previous studies in which participants more willingly exploited computers than humans in economic games [5] or complied less with a robot's request to continue practicing a visual search task compared to a human's request [6]. The current study indicates challenges to build a robot that can help humans comply with moral norms. Future work would be necessary to search for psychological factors that elicit resistance or promote adherence to a robot's moral influence.

Acknowledgment

The views expressed in this document are the authors and do not reflect the official position of the U.S. Air Force or U.S. Government.

References

[1] A. Briggle and C. Mitcham, *Ethics and Science: An Introduction*. Cambridge University Press, 2012.

[2] A. T. Nuyen, "Confucian Ethics as Role-Based Ethics," *International Philosophical Quarterly*, vol. 47, no. 3, pp. 315–328, 2007, doi: ipq200747324.

[3] B. Kim, R. Wen, Q. Zhu, T. Williams, and E. Phillips, "Robots as Moral Advisors: The Effects of Deontological, Virtue, and Confucian Role Ethics on Encouraging Honest Behavior," in *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, USA, Mar. 2021, pp. 10–18. doi: 10.1145/3434074.3446908.

[4] U. Fischbacher and F. Föllmi-Heusi, "Lies in Disguise—An Experimental Study on Cheating," *Journal of the European Economic Association*, vol. 11, no. 3, pp. 525–547, Jun. 2013, doi: 10.1111/jeea.12014.

[5] C. de Melo, S. Marsella, and J. Gratch, "People Do Not Feel Guilty About Exploiting Machines," *ACM Trans. Comput.-Hum. Interact.*, vol. 23, no. 2, p. 8:1-8:17, May 2016, doi: 10.1145/2890495.

[6] K. S. Haring *et al.*, "Robot Authority in Human-Robot Teaming: Effects of Human-Likeness and Physical Embodiment on Compliance," *Front Psychol*, vol. 12, p. 625713, May 2021, doi: 10.3389/fpsyg.2021.625713.