Journal Pre-proof

Can robot advisers encourage honesty?: Considering the impact of rule, identity, and role-based moral advice

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

PII: S1071-5819(24)00001-6
DOI: https://doi.org/10.1016/j.ijhcs.2024.103217
Reference: YIJHC 103217

To appear in: *International Journal of Human - Computer Studies*

Received date : 30 June 2023
Revised date : 10 December 2023
Accepted date : 1 January 2024

Please cite this article as: B. Kim, R. Wen, E.J. de Visser et al., Can robot advisers encourage honesty?: Considering the impact of rule, identity, and role-based moral advice. *International Journal of Human - Computer Studies* (2024), doi: https://doi.org/10.1016/j.ijhcs.2024.103217.

Highlights

Can robot advisers encourage honesty?: Considering the impact of rule, identity, and role-based moral advice

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

- Social robots may have challenges in encouraging honest behavior
- A robot's moral capacity may trigger psychological reactance
- This reactance can be mitigated by explaining a robot's moral capacity
- Deontology-based or Confucian role-based advice may also mitigate reactance

Can robot advisers encourage honesty?: Considering the impact of rule, identity, and role-based moral advice

Boyoung Kim[a], Ruchen Wen[b], Ewart J. de Visser[c], Chad C. Tossell[c], Qin Zhu[d], Tom Williams[e], and Elizabeth Phillips[f]

[a]*Center for Security Policy Studies-Korea, George Mason University Korea, Incheon, South Korea*
[b]*Department of Computer Science and Electrical Engineering, University of Maryland-Baltimore County, Maryland, USA*
[c]*Warfighter Effectiveness Research Center, United States Air Force Academy, Colorado, USA*
[d]*Department of Engineering Education, Virginia Tech, Virginia, USA*
[e]*Department of Computer Science, Colorado School of Mines, Colorado, USA*
[f]*Department of Psychology, George Mason University, Virginia, USA*

Abstract

A growing body of human-robot interaction literature is exploring whether and how social robots, by utilizing their physical presence or capacity for verbal and nonverbal behavior, can influence people's moral behavior. In the current research, we aimed to examine to what extent a social robot can effectively encourage people to act honestly by offering them moral advice. The robot either offered no advice at all or proactively offered moral advice before participants made a choice between acting honestly and cheating, and the underlying ethical framework of the advice was grounded in either deontology (rule-focused), virtue ethics (identity-focused), or Confucian role ethics (role-focused). Across three studies ($N = 1,693$), we did not find a robot's moral advice to be effective in deterring cheating. These null results were held constant even when we introduced the robot as being equipped with moral capacity to foster common expectations about the robot among participants before receiving the advice from it. The current work led us to an unexpected discovery of the psychological reactance effect associated with participants' perception of the robot's moral capacity. Stronger perceptions of the robot's moral capacity were linked to greater probabilities of cheating. These findings demonstrate how psychological reactance may impact human-robot interaction in moral domains and suggest potential strategies

for framing a robot's moral messages to avoid such reactance.

Keywords: human-robot interaction, robot ethics, honesty, cheating, moral advice, moral capacity

1. Introduction

Social robots refer to robots designed to interact with humans and each other (Duffy et al., 1999; Hegel et al., 2009; Breazeal, 2004). Some of these robots have physical resemblance with humans, can make human-like gestures, and can generate naturalistic speech and engage in conversations with humans. In the Human–Robot Interaction (HRI) research community, there has been increasing interest in understanding the influence that social robots can exert on people in making various judgments and decisions (Dautenhahn, 2007). These tasks include both lower-stake tasks like color-naming (Shinozawa et al., 2005), subjective card games (Salomons et al., 2021), or cognition tests (Saunderson and Nejat, 2021) and higher-stake tasks that fall under the domains of healthcare (Looije et al., 2010) or morality (Cappuccio et al., 2021; DeBaets, 2014; Malle, 2016). These existing studies in the moral domain have examined whether, when, and how social robots' verbal behaviors, non-verbal behaviors, and mere presence, either in isolation or in combination, influence people's moral judgments and decision-making.

The ways that social robots may influence people's behavior have been studied across contexts with a wide range of moral implications, from littering (Maeda et al., 2021) to destructive behaviors (Briggs and Scheutz, 2014) as well as across different dialogue patterns, such as a robot's expressions of uncertainty in its decisions for resolving ethical dilemmas (Straßmann et al., 2020), and robots' responses to morally problematic requests (Briggs et al., 2022; Jackson and Williams, 2019; Jackson et al., 2019) and to other observed norm violations (Winkle et al., 2019, 2021). Evidence accumulated from these studies suggests that the degree of moral influence social robots can exert on people may differ across contexts and according to different communication strategies.

One moral norm that previous researchers have used social robots to promote is the norm of *honesty* (Hoffman et al., 2015; Mubin et al., 2020; Petisca et al., 2020, 2022; Roizman et al., 2016). Specifically, these experiments have sought to deter cheating, with mixed results. For instance, Hoffman et al. (2015) found that merely bringing participants' attention to the presence

of a robot could be effective in discouraging cheating, and Petisca et al. (2022) demonstrated that a robot could succeed in deterring cheating when it repeatedly mentioned its awareness of participants' potential cheating. Meanwhile, other researchers have found other interventions to be largely unsuccessful, regardless of whether a robot was physically co-present with participants (Mubin et al., 2020) or virtually present through video (Petisca et al., 2020). As such, it remains unanswered what factors, including specific persuasive strategies, physical co-location, and broader contextual factors, enable robots to consistently deter cheating.

To address this research gap, we decided to break one of the key assumptions of prior work (Hoffman et al., 2015; Mubin et al., 2020; Petisca et al., 2020, 2022; Roizman et al., 2016). Although previous work has been focused on a robot's *reactive* responses to observed dishonesty, previous work outside the domain of honesty has shown the benefits of artificial moral advisors who, through *proactive* advice before a decision is made, can help people make well-informed, dispassionate, and consistent decisions (Giubilini and Savulescu, 2018; Savulescu and Maslen, 2015). Taking inspiration from that body of work, we thus instead explored *proactive* robot interventions to deter cheating before it had the chance to arise.

This proactive approach has shown to be effective for encouraging honesty in human-human interactions (HHI) (Bryan et al., 2013; Savir and Gamliel, 2019). Bryan et al. (2013), for example, demonstrated that when participants received a message emphasizing the implications of cheating for their moral identity ("cheater") before making a choice where they could cheat, they were less likely to choose to cheat than when the message emphasized the implications of cheating as a wrong action ("cheating is wrong") or when no message at all was offered. These findings from the HHI literature suggest that the impact of a robot's moral advice on discouraging cheating may also depend on how its messages are framed.

Although these results are promising, it is not obvious that they will transfer from human-human interaction to human-robot interaction. People tend to view robots as entities that are mechanistic, devoid of human-like nature (Haslam, 2006) or as entities belonging to an entirely new ontological category distinct from the existing categories of being (Kahn Jr et al., 2004, 2002, 2012; Melson et al., 2009). For instance, prior research has shown that children and adolescents classify humanoid robots as being in between the living and the not living, and are thereby unable to assign them to the existing categories of being (Kahn Jr et al., 2012; Weisman, 2022). Moreover, people

may hold different expectations for robots and humans when making certain moral decisions (Malle et al., 2015). These findings, therefore, present a key tension that we explore in our current research. On the one hand, the norm of honesty is a social norm that a majority of people are willing to adhere to most of the time (Halevy et al., 2014), and it is possible that an artificial moral advisor could improve people's moral decision-making surrounding this norm (Giubilini and Savulescu, 2018; Savulescu and Maslen, 2015). On the other hand, the differences in human perceptions and expectations of robots present uncertainty as to whether moral advice issued by a robot would convince people to forgo benefits that can be gained from cheating.

To resolve this uncertainty, we conducted three experiments across which we examined the effects of differentially-framed moral advice that is *proactively* offered by a social robot (the Aldebaran NAO) on deterring cheating behavior (a violation of the norm of honesty). Specifically, we varied the ethical theories around which the robot's moral advice was framed (deontological, virtue, and Confucian role ethics). While moral advice grounded in deontological and virtue ethics has been closely examined in the HHI literature (Bryan et al., 2013; Savir and Gamliel, 2019), there has recently been a surge of interest in the underexplored Confucian role ethical framework due to its focus on social-relational ontology and moral self-cultivation (Rosemont Jr and Ames, 2016; Williams et al., 2020; Zhu, 2020; Wen et al., 2022a,b). We argue that these key dimensions could be especially helpful for encouraging moral norms like honesty, which are so firmly grounded in the contract of social relations. The central aims of our three experiments were, thus, to understand (1) whether the empirical findings in the HHI literature on promoting honesty (Bryan et al., 2013; Savir and Gamliel, 2019) via differentially-framed messages are transferable to the HRI context, and (2) the particular benefits that underexplored non-Western philosophical traditions, such as Confucian role ethics, could contribute to this goal of encouraging honesty.

2. Background

Before describing our experimental approach, we will provide some additional background on Confucian role ethics and how it differs from Western traditions like deontology and virtue ethics. Deontological ethical theories posit that there are well-established sets of universally applicable moral principles dictating morally right or wrong actions (Briggle and Mitcham, 2012). Deontological ethics thus underscore the importance of aligning one's actions

with these moral principles. In contrast, virtue ethics emphasizes the central role of developing virtues, character, and moral dispositions for living a flourishing life and maintaining a person's social identity (Briggle and Mitcham, 2012; Hursthouse, 1999). Hence, in promoting the norm of honesty, rule-based moral advice grounded in deontological ethics would encourage people to act honestly and to not cheat, while identity-based moral advice grounded in virtue ethics would encourage people to be an honest person and develop a moral tendency to not cheat or be a cheater.

Understanding these two frameworks helps to articulate the novel perspective taken by Confucian role ethics, which instead focuses on one's awareness of their roles *in relation to other people* who also hold specific roles (such as their child, parent, teacher, or spouse). From this perspective, one's responsibilities are prescribed by the moral obligations required for serving well these different societal roles (Ames, 2011; Nuyen, 2007; Rosemont Jr and Ames, 2016; Ni, 2016; Zhu, 2020). The moral character and roles that are defined in relation to *specific* others generate *specific* responsibilities tethered to those relationships (Zhu, 2020).

While both virtue and Confucian role ethics underscore "the self," they do so in distinct ways. Virtue ethics fosters a self-sufficient morally neutral contemplator who lives an ideal life, and may focus on a person's moral character on its own or in relation to *general* others which could be either *non-specific* others or *specific* others. Therefore, one's position in society or social identity, such as a cheater, is often enabled by certain virtues (or vices) and exhibited in demonstrated virtuous dispositions. To some extent, developing virtues or stable virtuous dispositions (or tendencies) eventually leads to the ideal form of social identity.

In contrast, Confucian role ethics fosters an exemplary person who is fully immersed in social relatedness, and a practical and moral life (Ni, 2016). For role ethicists, roles are considered as the source of moral normativity, whereas virtuous tendencies are secondary (Ramsey, 2016). Therefore, to assess whether cheating is wrong or not is to examine whether cheating is instrumental for or detrimental to the fulfillment of one's role in relation to others in the community. In this sense, role-based moral advice grounded in Confucian role ethics and identity-based advice grounded in virtue ethics may have different influences on people's moral decision-making.

We sought to explore the potential differences between moral advice given by a robot to encourage honest behavior grounded in these three ethical theories (deontology, virtue ethics, and Confucian role ethics) across three

studies. In each study, we presented participants an opportunity to earn an extra bonus payoff, according to the self-reported results of the throw of a virtual die. As the amount of the bonus payoff was determined by the number they threw, participants faced a choice between self-reporting the number honestly or cheating by lying about the number they threw to gain a greater bonus payoff. Thus, this task was used as the context to examine whether a robot's advice rooted in different ethical frameworks could proactively impact adherence to the norm of honesty.

3. Study 1

In Study 1, we examined the effect of different ethical framing of robots' moral advice on the successful deterrence of cheating behavior. Participants were asked to engage in a version of a classic task of (dis)honesty (Fischbacher and Föllmi-Heusi, 2013), in which they threw a six-sided virtual dice to determine the amount of a bonus payoff, and then self-reported the results of their die-throwing. Immediately before they threw the die, a robot communicated with them, either conveying no moral information (control) or offering moral advice grounded in deontological ethics (rule), virtue ethics (identity), or Confucian role ethics (role). This task allowed for group-level assessment of cheating trends, and the effects of a robot's advice strategies on those trends. If the robot's advice had a strong effect on deterring cheating, this should be observed in group-level statistics. That is, the distribution of participants who claim to have thrown each of the six different numbers should follow a uniform distribution of 16.67%, whereas this distribution should be non-uniform (i.e., skewed in favor of higher responses) if people cheated to obtain a higher payoff. While this method, in which *proactively* offering a piece of moral advice to prevent cheating, has been validated in HHI studies (Bryan et al., 2013; Savir and Gamliel, 2019), it has not been previously deployed in HRI studies (except for a subset of Study 1's methods and results that have been reported in Kim et al., 2021). Assuming that these strategies to discourage cheating work in HRI in a manner similar to that in HHI, we would expect to achieve similar findings to those found in the HHI literature (Bryan et al., 2013; Savir and Gamliel, 2019). That is, we would expect identity-based moral advice to successfully discourage cheating, but rule-based moral advice to have little effect. The effect of moral advice grounded in Confucian role ethics has been underexplored, but again drawing from the existing HHI literature (Bryan et al., 2013; Savir and Gamliel, 2019), we hypothesized that

role-based moral advice would also be effective in deterring cheating because it highlights a certain aspect of *self-identity* (in this case, related to one's role and responsibility in relation to others), which is a characteristic that overlaps with the identity-based moral advice.

3.1. Methods

3.1.1. Participants

We recruited participants from Amazon's Mechanical Turk (MTurk) whose past HIT approval rate (i.e., the percentages of their participation in other MTurk studies having been approved) was greater than 90, and who were located in the United States. Our original plan was to examine the difference between the data distributions and the uniform distribution by performing chi-square tests.[1] Assuming a power of 80% and a moderate effect size w of 0.25, we calculated the recommended sample size using the R package pwr (Champely et al., 2018). The suggested sample size was 206, but because we expected a decrease in the final sample size after applying data screening procedures for data quality control (Peer et al., 2022), we aimed to recruit a total of 240 participants (60 participants per condition). In addition to chi-square tests, we used Kolmogorov–Smirnov tests. The sample size was also deemed appropriate for these tests as they are better for testing data distributions than chi-square tests even when the sample size is small (e.g., $N < 50$) and are more sensitive to the shape of distributions (Darling, 1957; Lilliefors, 1967; Engmann and Cousineau, 2011).

A total of 240 participants completed the study. We implemented several steps to verify that the participants understood the task and were attentive enough to follow the instructions. First, we compared the die number participants reported to have thrown and the amount of bonus payoff they claimed to have earned in the die-roll game. If these two responses did not match (i.e., not matching what is dictated by the payoff table), we discarded the responses submitted by those participants. We next removed the responses collected from the participants who did not correctly answer any of the two questions intended to check the audio and video setup of their computers. Among 240 participants, 12 participants were excluded for providing dice number and payoff responses that did not match, and 29 participants were

[1]These chi-square test results for all three studies can be found in the Supplementary Material.

excluded for failing audio and video check tests. This left us with data from 199 participants (M_{Age} = 39.41, SD_{Age} = 12.04, 131 male, 68 female). For race and ethnicity, 10 participants self-categorized as Asian, 13 as Black or African American, 16 as Hispanic or Latino, 157 as White, and 3 as Other. When asked about their prior experience with robots, most participants expressed interest in robotics and/or AI as a hobby but indicated that they had little formal training (M = 3.09, SD = 1.53 on a 7-point rating scale). All participants electronically signed an informed consent form approved by the Colorado School of Mines's Human Subject Research Office. It took about 12 minutes to complete the study. They received a $2.00 base rate in return for their participation, plus up to $0.90 of bonus compensation based on their reported die rolls.

3.1.2. Stimuli

Die-rolling game. To create a situation in which cheating would be a tempting choice, we used a version of the die-rolling game (Fischbacher and Föllmi-Heusi, 2013). Multiple variants of the die-rolling game have been used to examine psychological factors influencing people's (dis)honest behavior in human-human interaction (Bryan et al., 2013; Fischbacher and Föllmi-Heusi, 2013; Halevy et al., 2014; Savir and Gamliel, 2019; Shalvi et al., 2011, 2012) and have recently been applied to human–robot interaction (Petisca et al., 2022; Kim et al., 2021). In our version of the die-rolling game, participants were instructed to visit a die-rolling website, and throw a six-sided die at least twice. Participants were informed that while they could throw the die as many times as they liked, their bonus payment would be determined by the first number they threw, and as such, this was the number they should report. As shown in Table 1, for throwing the numbers between 1 and 5, the bonus increased for each subsequent number by 20 cents, from 10 cents to 90 cents. For a throw of 6, the resulting bonus payment was fixed to zero cents (Fischbacher and Föllmi-Heusi, 2013). As we asked participants to report both the number they threw and the amount of the resulting bonus payment, we expected that we could detect participants who did not understand the payoff rules by checking if they submitted die rolls and rewards that did not match the bonus payoff table and exclude their data from analysis.

Moral advice. There were four different moral advice conditions. In the control condition, the robot provided instructions about the task but did not give moral advice based upon any particular ethical theory. In the other

Table 1: Bonus Payoff Structure

Thrown Number	1	2	3	4	5	6
Bonus Payoff	10c	30c	50c	70c	90c	0c

three different moral advice conditions, the robot offered a piece of moral advice to the participants in addition to giving task instructions. These moral advice conditions were the rule, identity, and role conditions, which were respectively grounded in deontological, virtue, and Confucian role ethical theories.

In the *rule* condition, the moral advice was focused on stating that cheating is a morally wrong behavior. Thus, the robot in the rule condition advised participants as follows:

> "Cheating to maximize your bonus is morally wrong behavior."

In the *identity* condition, the moral advice was focused on communicating how cheating would reflect on participants' moral character. Thus, the robot in the identity condition advised participants as follows:

> "Cheating to maximize your bonus will make you a cheater."

Finally, in the *role* condition, the moral advice was focused on the participants' role as a member of the MTurk community and their responsibility not to unfairly disadvantage their MTurk community members. Since the data collection for all three studies took place while the COVID-19 pandemic was at its peak (time period between September 2020 and January 2022), we targeted participants available in an online crowdsourcing research platform. In preparing the moral advice for the role condition, we thus placed an emphasis on a common background the participants shared, which is holding a membership as a MTurk worker in the MTurk community. Therefore, the robot in the role condition advised participants as follows:

> "A good MTurk community member would not cheat to maximize their bonus at the expense of other MTurkers."

Robot video clips. As this was an online study, we recorded video of a NAO robot (Softbank Robotics) and presented video clips of the robot to participants. In all videos, the upper body of the robot appeared against a black background (Figure 1). As the robot spoke, it made naturalistic gestures by, for example, moving its head and arms. To encourage participants to concentrate on the robot and what it said in the videos, we did not offer any captions in the video clips. Instead, if participants desired to read the scripts of the robot's speeches, they could access the script by clicking on a button below each video. Video clips used for the control, rule, identity, and role conditions are compiled in $https://osf.io/h94qk/?view_only = aace69c33fb44121982982e7c39c20ad$

Figure 1: A screenshot image of video clips used for the robot condition in Studies 1-3

3.1.3. Measures

Cheating behavior. As participants played the die-rolling game on a third-party website, the numbers they threw were kept hidden from the experimenters. Hence, participants had the opportunity to lie about the number they threw in order to acquire larger bonus payments than they actually won from the game, without being caught. Although it was impossible to ascertain cheating behavior at an individual-participant level, we were able to detect cheating at a group level by assessing the overall distribution of the numbers and the bonus payments the participants claimed, which is an approach that also has been adopted in prior work (Fischbacher and Föllmi-Heusi, 2013; Bryan et al., 2013; Shalvi et al., 2012). Assuming a fair die, the distribution of thrown numbers would follow a uniform distribution. In other words, the percentage of each number that would be claimed to have been thrown by participants would converge towards 16.67. Thus, by assessing the

degree to which the distribution of the thrown numbers deviated from the expected uniform distribution of die rolls, we gauged the patterns of cheating in different moral advice conditions.

Robot/AI familiarity. To examine participants' prior knowledge of or experience with robots and AI, we asked participants, "How much prior experience do you have with robots and artificial intelligence (AI)?" and instructed them to answer the question on a scale ranging from 1 (I have no prior experience with robots and/or AI) through 3 (I am interested in robotics and/or AI as a hobby, but have little formal training) and 5 (I have some formal training in robotics and/or AI, e.g., university classes) to 7 (I have a career in robotics and/or AI or an equivalent level of experience).

3.1.4. Design and Procedures

The design of Study 1 was a one-way (moral advice: control, rule, identity, or role) between-subjects design where participants were randomly assigned to one of the four different moral advice conditions, which consisted of control (no advice), rule (deontological ethics), identity (virtue ethics), role (Confucian role ethics) conditions.

Participants were guided through the experiment by watching short video clips of a NAO robot that gave verbal instructions, and by reading written instructions presented alongside the video clips. For instance, when presenting a video clip where the robot delivered the core advice in the three different moral advice conditions (e.g., "Cheating will make you a cheater." in the identity condition), the same message was presented in written text under the video clip. Both video clips and the written instructions were shown on the Qualtrics platform. Upon agreeing to participate in the study, participants were presented with the first video clip in which the robot introduced itself as NAO, a Research Assistant. The robot informed participants that they would be participating in multiple short studies (communicated simply to situate the experiment), and that, for their participation in multiple studies, they would receive a small bonus payment of which the amount would be determined by the first number they throw in a die-rolling game. Specifically, the robot gave the following instructions to the participants:

> "In this experiment, we ask you to participate in multiple short studies. For participating in multiple studies, you will receive a small additional payment. However, this payment will not be the

> same for every participant. The exact amount of your payment will be determined by throwing a virtual die twice. Your first throw will determine how much money you receive. The exact payment amount for a given die throw is listed in the table below (Table 1 was presented under the video). The second throw only serves to make sure that the die is working properly. You may of course throw the die more than twice if you like. However, only the first throw should be used to determine your payment."

The robot then conveyed the following additional piece of information about disadvantages that other MTurk workers would have to bear as a consequence if the participants made dishonest choices because it would unfairly restrict the largest possible bonus payoff the MTurk worker who participates in the study after them.

> "Please note that there is a limited pool of money 90 cents available for you and the next participant. Accordingly, the more money you win in this study, the less will be available for the next participant. Please do not hesitate to claim money that you have rightfully won, but be aware that cheating to maximize your own reward will result in less money being available to the next participant."

Immediately after delivering the above message, the robot in the rule, identity, and role conditions offered moral advice corresponding to each condition, and the robot in the control condition provided no additional message. The moral advice in the rule, identity, and role conditions was delivered following the opening phrase "To be clear." For example, in the rule condition, the robot said "To be clear, cheating to maximize your bonus is morally wrong behavior."

Next, the robot instructed the participants to visit a website where they could throw a virtual die. They were presented with a URL link and, when they clicked on the link, the die roll web page of the random.org's website popped open in a new tab of their web browser window. Participants were asked to return to the original tab where the experiment was in progress after finishing throwing a virtual die and to submit the first number they threw and the resulting bonus payment. Participants submitted these responses by clicking on the options shown in two separate drop-down lists. They

then were asked to answer the questionnaires including the question about their prior experience with robots and AI and questions about their basic demographics information, age, gender, and ethnicity.

3.2. Data Analysis and Results

Figure 2 shows percentages of the claimed die numbers and bonus payments in increasing order of the bonus payment from 0 cents to 90 cents (the numbers are displayed in the order of 6, 1, 2, 3, 4, and 5 along the x-axis). When throwing a six-sided fair dice, the probability of throwing each of the six possible numbers would remain close to 0.167 ($= \frac{1}{6}$). If honest choices (reporting the actual numbers thrown) had been the dominant trend, the overall distribution of the numbers in Figure 2 would have followed a uniform distribution of 16.17% (i.e., the horizontal dotted line in Figure 2) across the six different numbers. However, Figure 2 suggests deviations from the uniform distribution for all four conditions. We analyzed this response pattern by performing one-sample Kolmogorov–Smirnov tests for discrete data using the empirical distribution function (Massey Jr, 1951) on each condition. We compared the overall distribution of the frequency of throwing the six different numbers with the uniform distribution of the six numbers being thrown with equal frequency. These analyses suggested that the distributions of the numbers in all four conditions significantly deviated from the uniform distribution: $D(51) = 0.40$, $p < .001$ for the control condition; $D(53) = 0.44$, $p < .001$ for the rule condition; $D(48) = 0.46$, $p < .001$ for the identity condition; and $D(47) = 0.46$, $p < .001$ for the role condition. These results indicated the prevalence of dishonest choices in all four conditions.

On a closer look at the response pattern in Figure 2, other noteworthy possible findings became visible. Of the six possible die numbers, 5 was most frequently reported as thrown in all four conditions. Except for 3 in the control and the rule conditions and 4 in the role condition, the summed amount of deviations from 16.67% across the numbers 6, 1, 2, 3, and 4 was close to the exact amount of deviation from 16.67% for the number 5. This suggested that, when participants decided to report a number different from the actual number they had thrown, they tended to report 5, which resulted in a maximum bonus payoff of 90 cents. To examine this specific pattern of cheating, we further analyzed the percentage of participants reporting to have thrown the number 5.

Despite the number 5's frequency across conditions, if identity-based and role-based moral advice had exerted persuasive influence on participants'

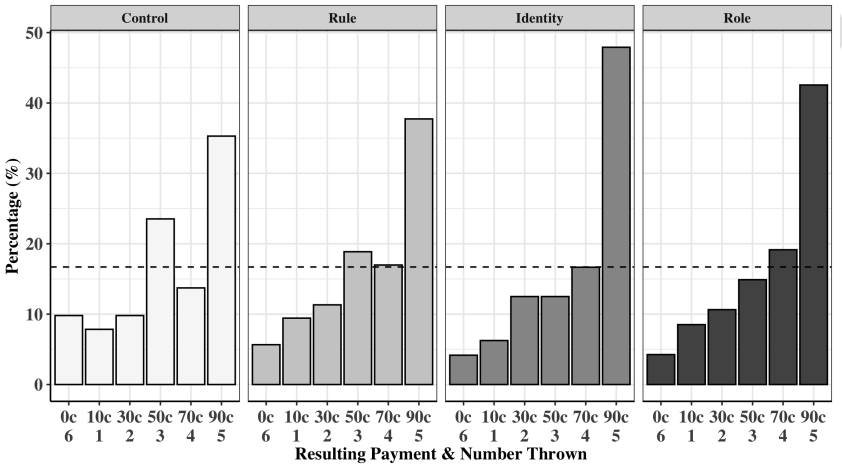

Figure 2: The percentage of the claimed bonus payoff (thrown number) in Study 1. The dotted line represents the expected uniform distribution for a fair six-sided die, 16.67%.

choices and deterred cheating, this frequency should have been smaller in the identity and the role conditions than in the control condition (in which no advice was offered by the robot). However, Figure 2 suggests potentially the opposite response pattern. That is, the percentage of participants claiming to have thrown the number 5 was in fact lower in the control condition (35.29%) than in the identity (47.92%) and role (42.55%) conditions. The percentage of participants reporting five in the control condition was instead more similar to that of the rule condition (37.74%). We performed two-proportion Z-tests to examine whether the proportion of reporting the number 5 in each of the three different advice conditions was significantly different from the proportion of reporting the number 5 in the control condition. However, these differences were not significant ($p_{minimum} = .20$). In conclusion, we did not find a statistically significant difference between the proportion of throwing 5 in the control condition and each of the rule, identity, and role conditions (Data for all three studies can be found at $https://osf.io/h94qk/?view_only = aace69c33fb44121982982e7c39c20ad$).

3.3. Discussion

In Study 1, we found evidence for cheating when a robot offered rule-based moral advice, emphasizing the rule that it is wrong to cheat, or no advice at all. These findings were consistent with the prior research in HHI that found little effect of a moral message emphasizing the wrongness of an action on deterring cheating (Bryan et al., 2013; Savir and Gamliel, 2019). However, in contrast to our expectations based on the prior HHI work (Bryan et al., 2013; Savir and Gamliel, 2019), we also found evidence that cheating was prevalent when a robot issued identity-based moral advice, emphasizing one's moral identity, and role-based moral advice, emphasizing the relationship with others and one's role responsibility. Given that moral advice did not lead to deterrence of cheating (and in fact may have backfired) across all three of these conditions, it is unclear whether the null effect in the rule condition was caused by a lack of persuasiveness in that condition, as it was posited in the existing literature on dishonesty in HHI (Bryan et al., 2013; Savir and Gamliel, 2019), or because the delivery of the moral advice was carried out by a robot. To identify potential reasons for the ineffectiveness of identity-based and role-based advice in discouraging cheating, we conducted Study 2 to discern whether Study 1's findings can be attributed to moral advice itself being ineffective or a robot serving in the role of a moral advisor.

4. Study 2

In Study 2, we examined the effectiveness of an identical set of moral advice when it was delivered by either a human or a robot. We kept the robot condition and added the human condition in which either a female or a male research assistant gave instructions about the task and offered moral advice prior to the die-rolling game, replacing the role a robot played in the robot condition. Also, instead of using the external, third party website for the die-rolling game, we used our custom-made die-rolling game which allowed us to record the actual numbers each participant threw. Therefore, we were able to verify at an individual participant level whether their claimed numbers were the same or different from the numbers they rolled in the die-rolling game. By making this modification, we could directly compare the probability of cheating in each of the three moral advice conditions with the probability of cheating in the control (no advice) condition. If identity-based and role-based moral advice was ineffective in Study 1 solely because it was offered by a robot, the probability of cheating would be lower when the same

advice was offered by a human than when no advice (control) was offered. Alternatively, if identity-based and role-based moral advice was ineffective in Study 1 because the content of the advice itself was not persuasive enough, identity-based and role-based moral advice would remain to be ineffective in deterring cheating, regardless of whether it was a human or a robot that offered the advice.

4.1. Methods

4.1.1. Participants

A total of 967 participants, whose past HIT approval rate was greater than 90 and location was the United States, completed the study via MTurk. There is little consensus over how sample size should be determined for a logistic regression analysis (Demidenko, 2007), and this problem was evident as we planned to conduct a logistic regression analysis that has a categorical predictor with multiple levels. To determine a sample size, we considered that 96 participants are needed to precisely estimate a logistic regression model that only has the intercept (Harrell and Harrell, 2015) and that the general guidelines are to have approximately 20 events (in this case, cheating) per variable (van der Ploeg et al., 2014). We inferred from Study 1's results that, if participants chose to cheat, it was most likely to obtain the largest possible bonus payoff.[2] We thus estimated that approximately 18.62% of the participants in the control condition (9.5 participants out of 51) were likely to have cheated by claiming to have thrown the number 5 (The difference in the observed and the expected percentages of claiming to have thrown five was calculated by subtracting 16.67% from 35.29%, as found in Study 1's control condition). We assumed that to examine the effect of each advice on cheating, at least 20 participants would be required for each of the advice condition. Therefore, we sought to double the sample size we recruited for Study 1. For Studies 2 and 3, we aimed to recruit a total of 960 (120 participants per condition).

We implemented the same data screening steps used in Study 1. We discarded the data collected from 82 participants whose claimed dice numbers and claimed bonus payoffs were mismatched, and the data collected from 214 participants who incorrectly answered any of the audio and video check

[2]Table S1 in the Supplementary Material also suggests that, in Studies 2 and 3, when participants dishonestly reported the numbers they had thrown, the majority claimed to have thrown five (64.42% in Study 1 and 67.92% in Study 2).

tests. After screening the data through these steps, 671 participants were left. Among them, 9 participants claimed to have thrown a number that resulted in a lesser amount of a bonus payment than they had earned. These choices are not cheating, and as it was unclear whether these participants claimed a lesser amount of bonus due to a concern for other participants or a lack of task comprehension, we removed these responses from data analyses.

We conducted analyses on the remaining 662 participants ($M_{Age} = 39.31$, $SD_{Age} = 11.87$, 392 male, 265 female, 2 other, 3 preferred not to say). The composition of race and ethnicity with which participants reported to have identified was 4 American Indian or Alaska Native, 79 Asians, 68 Black or African Americans, 39 Hispanic or Latino, 2 Native Hawaiian or Other Pacific Islander, 456 White, 10 Other, and 4 preferred not to say. Participants indicated that they were on average interested in robotics and/or AI as a hobby but had little formal training ($M = 2.81$, $SD = 1.46$). Participants submitted electronically signed informed consent forms approved by George Mason University's Institutional Review Board Office. It took about 12 minutes to complete the study. Participants received a $1.50 base rate in return for their participation and, regardless of the outcome of the die-rolling game, received $0.90 bonus payments.

4.1.2. Stimuli

Die-rolling game. We created a die-rolling game on Qualtrics, the online study administration platform. For the first two rolls, participants were informed of the trial number (e.g., "This is your second throw.") and instructed to click on a button shown on the bottom right corner of a webpage to throw a die. When they clicked on the button, an image of one of the six sides of a die was presented on the screen with the die number written above in text (See Figure 3). Participants could throw a die up to 10 times by choice. From the third to tenth roll, participants were first asked to indicate whether they would like to throw a die again and, if they clicked on the"yes" button, as opposed to the "no" button, one of the six sides of a die was presented.

Moral advice. The messages used as moral advice in the rule, identity, and role conditions were the same as those used in Study 1.

Robot video clips. For the robot condition, we modified the video clips used in Study 1. As we used a custom-made die-rolling game to record what number participants threw, we removed the parts where the robot gave instructions on how participants could visit a third-party webpage (i.e., random.org) and

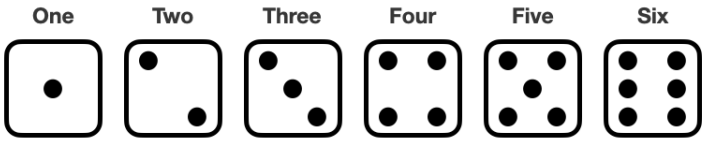

Figure 3: Images of the six sides of a die used in Studies 2 and 3

throw a die. In Study 2, we deleted the core message the robot delivered in the three different moral advice conditions (e.g., "Cheating will make you a cheater." in the identity condition) that had been written in text under the video clip in Study 1. This change was implemented because presenting the core message in written text in addition to having participants listen to what the robot says in the video could interfere with concentrating on the robot's message in the video, potentially weakening the experiences of interacting with a robot (albeit indirectly by watching the video).

Human video clips. For the human condition, the identical video scripts and presentation methods of the video clips and the written instructions were used, except that the female research assistant presented herself as "Eva" and the male research assistant presented himself as "Dan." Same as the videos of the robot condition, their videos were filmed in front of a black background. Their upper body was shown in the middle of the screen and, unlike in the robot condition, a small microphone was placed in front of them (See Figure 4).

Figure 4: Screenshot images of video clips used for the human condition in Study 2

4.1.3. Measures

Cheating behavior. We operationally defined cheating in Study 2 as an act of reporting the number that resulted in the amount of bonus payment larger than the actual amount of bonus payment participants had earned from the die-rolling game.

Robot/AI familiarity. The same question about participants' prior familiarity with robots and AI was used as in Study 1.

4.1.4. Design and Procedures

The design of Study 2 was a 2 (agent: robot or human) by 4 (moral advice: control, rule, identity, or role) between-subjects design where participants were randomly assigned to one of the eight possible conditions. In the human agent condition, the role of a moral advisor was played by a female and a male, but because examining the gender effect was not the main purpose of this research, we sought to randomly assign a half of the participants to the robot condition and the other half to the human condition collapsed across the female and male conditions.

As it was in Study 1, we guided participants through the study by asking them to read instructions written in text and watch videos on the Qualtrics platform. In Study 2, we did not use the existing third-party website for the die-rolling game and instead used our custom-made game created on Qualtrics. To reduce a possibility of participants' decisions being affected by their guesses about whether the researchers could later verify the number they obtained from the game, we divided the study into two different phases. In the first phase, participants received instructions about the study and the die-rolling game and, when it was time for the participants to play the die-rolling game, they were redirected to another survey where they could play the game and answer the questionnaires. In the die-rolling game, after throwing the die twice, participants were given the option to throw a die up to eight more times if desired. As in Study 1, participants were asked to report the first number they threw and the amount of bonus they earned respectively by clicking on the options shown in two separate drop-down lists.

4.2. Data Analysis and Results

By checking whether the first number participants threw and the number they *claimed* to have thrown matched or not, we dummy-coded matching answers as '0 (honest answers)' and mismatching answers as '1 (dishonest

answers).' To ensure these dishonest answers only include cases where participants made dishonest choices to gain a larger bonus payoff, we excluded mismatching answers that led to a smaller bonus payoff than participants actually had earned.

Human moral advisor. To understand the effects of moral advice grounded in different ethical frameworks within either a human or a robot moral advisor, we performed a logistic regression analysis separately for the human condition and the robot condition. We evaluated the relative effectiveness of each of rule-based, identity-based, and role-based advice in discouraging cheating by comparing the probability of cheating in each advice condition with the probability of cheating in the control condition where no advice was provided. We thus, in fitting the logistic regression model, specified the contrasts for moral advice in such a way that each of the three different types of moral advice was compared, one by one, with the control condition.

These analyses revealed that, when a human offered moral advice, the probability of cheating was lower in the role condition than the control condition, $b = -0.96$, $SE = 0.48$, $z = -2.00$, $p = .046$, Odds Ratio (OR) = 0.38, 95% Confidence Interval (95% CI) = [0.14, 0.95]. There also was a marginally significant effect suggesting that the probability of cheating was lower in the rule condition than the control condition, $b = -0.83$, $SE = 0.44$, $z = -1.87$, $p = .062$, OR = 0.44, 95% CI = [0.18, 1.02] (See Figure 5).[3]

Robot moral advisor. When a robot offered moral advice, we could not find evidence indicating that any of the advice in the rule, identity, and role conditions reduced the probability of cheating relative to the control condition ($p_{minimum} = .20$, See Figure 5).[4]

[3]To increase the interpretability of these results found in the human condition and complement the traditional null hypothesis testing approach, we conducted a Bayesian logistic regression analysis with a weakly informative prior, as there was insufficient information to specify priors (Gelman et al., 2008), on responses collected for the human conditions. From this supplementary analysis, we found support for the effects of moral advice grounded in Confucian role ethics as well as deontological ethics. The estimated posterior median for the role condition was -0.95 (95% Credible Interval = [-1.93, -0.04]), and the estimated posterior median for the rule condition was -0.81 (95% Credible Interval = [-1.74 and 0.01]).

[4]We performed an exploratory analysis to examine how the relative effectiveness of moral advice grounded in different ethical frameworks differed depending on whether the advice was offered by a human or a robot. We collapsed the data sets of the human and

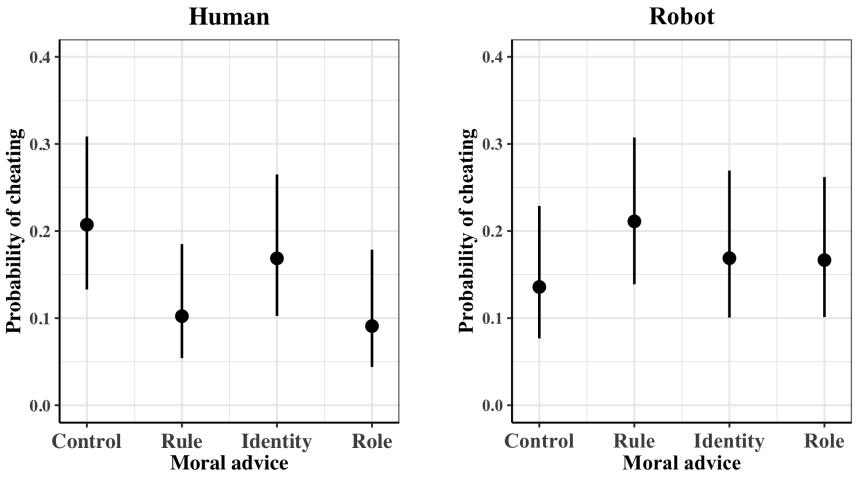

Figure 5: Probabilities of cheating as a function of agent (human or robot) and moral advice (control, rule, identity, or role) in Study 2. The error bars represent 95% confidence intervals.

4.3. Discussion

In Study 2, we found that, when it was a human that offered moral advice, the advice grounded in Confucian role ethics led to less cheating than the control condition where no advice was offered. This effect of role-based advice on deterring cheating was consistent with the hypothesis we derived from the previous findings in the HHI literature (Bryan et al., 2013; Savir and Gamliel,

the robot conditions and fit to this combined data set a logistic regression model where we entered moral advice (the reference level was set to the control condition), agent type (human vs. robot), and the interaction term of these two factors as predictors. We found a significant interaction effect between moral advice (where the contrast was set to compare the rule condition with the control condition) and agent type, $b = 1.36$, $SE = 0.61$, $z = 2.24$, $p = .025$, OR = 3.91, 95% CI = [1.21, 13.29]. Figure 5 shows that, when the agent was a human, the probability of cheating was lower in the rule condition than in the control condition; but in contrast to this pattern, when the agent was a robot, the probability of cheating was higher in the rule condition than in the control condition. This interaction effect complements the findings from our analysis of the human condition data where the rule-based advice was found to have a potential effect on discouraging cheating.

2019). However, Study 2 did not show evidence to suggest that identity-based advice was effective in deterring cheating even when the human served as a moral advisor. This lack of support for moral advice grounded in virtue ethics was inconsistent with the previous HHI research (Bryan et al., 2013; Savir and Gamliel, 2019). We, instead, found an unexpected but noteworthy trend that rule-based moral advice, grounded in deontological ethics, could potentially deter cheating when the advice was offered by a human.

Although the results involving the identity-based moral advice deviated from our predictions, the deterrence effects of role- and rule-based advice on cheating, when the advice-giver was a human, indicate that Study 1's null effects cannot be solely attributed to the content of moral advice lacking persuasive power. Lastly, when a human issued the advice, we found no significant difference in the probability of cheating between the identity condition and the control condition ($p = .526$). Alternatively, it is likely that participants in Studies 1 and 2 were resistant to moral advice offered by a robot because of their perceptions of or attitudes towards the robot given that in Study 2, we again found that, when a robot issued moral advice, it did not reduce cheating behavior regardless of which ethical theory was used to formulate its advice.

At first glance, Studies 1 and 2's results may be interpreted as substantial reasons to conclude that people absolutely refuse to accept robots as an advice-giver and the effects of a robot's advice barely exists. This possibility lacks strong support, however, because there is prior research showing that people adjust their preference for seeking advice from either a human or a robot depending on characteristics of the task at hand, rather than rejecting to seek *any* advice from a robot. Specifically, Hertz and Wiese (2019) found that participants preferred a human advisor for an emotion recognition task but preferred a robot advisor over a human for a math task.

Perhaps, then, people refuse to delegate the role of an advisor only in social and moral domains. For example, one study showed that, while participants found a robot's moral advice trustworthy, they would not necessarily intend to rely on it Momen et al. (2023). However, there also is evidence suggesting the potential for successfully deploying a robot advisor in social and moral domains. Some HRI studies showed that it is possible for a robot to persuade people to comply with a social norm through the means of verbal communication. For instance, previous work showed that a robot's verbal protest can deter participants from engaging in a morally problematic behavior (Briggs and Scheutz, 2014; Briggs et al., 2022) and a robot's message

grounded in Confucian role ethics can encourage participants to improve their task performance by meticulously working on a given task, exhibiting good research participant behavior (Wen et al., 2022b).

We therefore considered two other potential factors. First, in the current research, it could have been the case that the presence of the robot had been too weak to have any impact on participants' decision-making. This is because in the present research, the interaction between the robot and participants was enabled by the participants watching pre-recorded videos of the robot in an online experimental setting. However, the existing findings in HRI research have mixed support for this factor. Some researchers showed that, in an online experiment, playing a video clip of a watchful robot while participants make decisions did not reduce cheating than having participants make decisions alone (Petisca et al., 2020). In contrast, in other work also carried out as an online experiment, researchers found that having participants watch videos of a robot giving a certain verbal message can be effective in encouraging participants to comply with a social norm (Wen et al., 2022b). Therefore, even if the robot had a limited presence as a consequence of only allowing participants to have simulated interactions with a robot, this did not seem to have been a factor that can substantially explain the absence of the robot's influences in Studies 1 and 2.

The second factor we next considered was individual differences in participants' familiarity with and knowledge about robots and their relationship with participants' inferences about the robot's capabilities. When participants were asked about their prior experience with robots and artificial intelligence in the first two studies, the majority of the participants in Study 1 (76.38%) as well as in Study 2 (77.64%) indicated that they had somewhere between *no prior experience with robots and AI (1)* and *interest in robotics and AI as a hobby but had little prior training (3)* on a 7-point rating scale. Since participants had little familiarity with and knowledge about robots in general, they may not have been able to form coherent or informed expectations about any capacities a robot may possess, let alone a social or moral capacity, which could have in turn weakened or blocked the effect of the robot's advice.

This explanation seemed plausible if it were to be assumed that attributions of relevant mental capacities, such as social and moral capacities, are the prerequisite for an artificial agent to have meaningful moral influences on people. To illustrate, when a human plays the role of a moral advisor, it is unlikely that participants wonder whether the human advisor has any mental

capacities of knowing about what is morally right or wrong. Most likely, the human advisor would be viewed as sharing common norms of society as participants themselves do. In other words, an act of giving moral advice would appear natural when it is originated from a human, allowing expectations or heuristics that are aligned with natural human behavior (Banks et al., 2021).

However, when a robot plays the role of a moral advisor, this assumption or belief about commonalities is no longer applicable. The influential power of information is affected by the credibility of the information source (Hovland and Weiss, 1951). This indicates that for information, which would be moral advice in this research, to have influences on people's decision, people should be able to have enough prior knowledge about or experiences with the information source, which would be a robot, to evaluate the credibility of the source.

We thus reasoned that, for a robot to convince people to follow a norm, it may be needed to first demonstrate its social and moral capacities, which are directly pertinent to the advice it is offering. Supporting this possibility, Petisca et al. (2022) showed that, in repeated rounds of a version of the die-rolling game, the probability of cheating was lower after participants heard, multiple times, a robot making remarks signaling its awareness of their possible cheating than after participants played the game alone. Moreover, these researchers found that the robot making remarks implying its awareness of potential cheating was rated as having more social capabilities than the robot making remarks but without any implications of its awareness of participants' potential cheating. Therefore, in the next study, we examined this possible link between participants' perceptions of a robot's moral capacities and the effectiveness of the robot's moral advice in deterring cheating.

5. Study 3

In Study 3, we investigated whether, when participants were informed of a robot's moral capacities before they encountered the robot and received moral advice from it, it would suppress their cheating behavior. We retained the robot agent conditions of Study 2 where a robot offered either no advice or one of the three differentially framed moral advice, and added the parallel conditions with a modification. In these new conditions, before the robot gave instructions about the die-rolling game, participants were provided with a short description of the robot's capacities. In this description, it was explicitly stated that the robot was *programmed by a human* to have

social and moral capacities. We expected that, rather than either omitting or denying the presence of a human programmer, overtly acknowledging the human mind as an enabling factor behind the robot's actions would help participants view the robot's having social and moral capacities as well as offering them moral advice as more plausible. This approach seemed to be more accurately reflecting the current state of robotics technology than an approach that attempts to have participants believe the existence of socially and morally competent robots. Moreover, considering participants' low familiarity with and knowledge about robots, it seemed to be able to provide a basic common foundation for forming what to expect from the robot participants were about to encounter in the study. Therefore, Study 3 tested the effect of a robot moral advisor when it was described as a means (for humans) to promote the norm of honesty (to other humans), without requiring assumptions about the robot's understanding or conscious awareness of the norms.

In Study 3, we also included a questionnaire measuring participants' perceptions of the mental capacities of a particular robot to which they were introduced in this experimental context. In Studies 1 and 2, we asked participants about their prior experience with robots and AI. But, because this question was inquiring about their familiarity with robots and AI *in general*, it was not adequate for understanding their impressions of the capacities of a specific robot they were exposed in the experiment. We expected that the inclusion of this additional measure would allow explorations of whether the effect of a robot's moral advice on cheating could be affected by the extent to which participants attributed moral capacities to the robot.

We predicted that when participants were informed of a robot's moral capacities before receiving moral advice from the robot (which we refer to as the background condition), the probability of their cheating would be lower in the role-based advice condition than in the control condition. As the prior HHI research (Bryan et al., 2013; Savir and Gamliel, 2019) suggested, we expected that in the background condition, the probability of cheating would be lower in the identity-based advice condition than in the control condition, but we did not expect this effect of advice for the rule-based advice condition.

5.1. Methods

5.1.1. Participants

In total, 968 participants whose prior HIT approval rate was greater than 90 and location was set to the United States participated in the study via

MTurk. Among them, 15 participants completed the study twice due to a technical error. We kept the first participation from these participants and removed the second participation from the data to ensure their decisions only reflect their initial responses to the experimental manipulations. We then, following Study 2's data screening procedures, discarded data from 81 participants who submitted the numbers and bonus payoffs that did not match the payoff table and 40 participants who claimed to have earned a smaller amount of bonus than the amount they actually had earned in the die-rolling game. We performed analyses on the data obtained from the remaining 832 participants (M_{Age} = 37.87, SD_{Age} = 10.73, 508 male, 323 female, 1 preferred not to say). Participants' self-identified race and ethnicity consisted of 3 American Indian or Alaska Native, 32 Asians, 108 Black or African Americans, 28 Hispanic or Latinos, 1 Native Hawaiian or Other Pacific Islander, 646 White, 6 Other, and 8 preferred not to say. Participants submitted electronically signed consent forms approved by George Mason University's Institutional Review Board Office. To complete the study, it took approximately 15 minutes. Participants received a $1.00 base rate in return for their participation and, regardless of the outcome of the die-rolling game, all of them received $0.90 bonus payments.

5.1.2. Stimuli

The custom-made die-rolling game, the moral advice used in the rule, identity, and role conditions, and the video clips of a robot were identical to those used in Study 2.

Background information about the robot. In a written description of the background information about the robot, we first provided a general definition of social robots by describing them as autonomous robots designed to exhibit human-like behaviors and interact with humans in daily life (Duffy et al., 1999). We then provided a further description of moral capacities programmed onto a specific social robot participants would encounter in the study. Drawing from Malle and Scheutz (2020), we described the robot as being programmed to have a database of moral norms, conform to the norms, and encourage others to conform to the norms.

The background information about the robot was as follows:

> "In this experiment, a social robot will walk you through the tasks and procedures. Social robots are autonomous robots that are designed to interact with humans in everyday life. These

> robots are programmed to perform human-like behaviors. For example, the social robot you will see today is programmed to make gestures like a human does when it speaks. This social robot is also specially programmed with a comprehensive database of social norms that many humans generally follow. For example, the robot has information about things that a person should and should not do. Moreover, the robot is programmed to follow these norms and to encourage others to follow the norms."

5.1.3. Measures

In addition to other measures included in Studies 1 and 2, we administered the Measure of Mind Perception in 3 to 5 Dimensions (MMP35; Malle, 2019) to examine mental capacities people attribute to the robot to which they were exposed in the study. This measure is composed of five subscales, positive affect, negative affect, morality, social cognition, and reality interaction, each of which is represented by the averaged rating of 4 items. For instance, the 4 items for measuring the capacity of morality consisted of *praising moral actions, upholding moral values, telling right from wrong, and disapproving of immoral actions*. We instructed participants to "Tell us about your perception of the social robot NAO you met today. Some judgments might be difficult to make; just give us your best guess." For each statement, the participants' task was to indicate "To what degree do you think this robot is capable of [each statement] (e.g., telling right from wrong)." on the rating scale ranging between 0 (Not at all capable) and 7 (Completely capable). The presentation order of the all 20 items was randomized across participants.

5.1.4. Design and Procedures

Study 3 followed a 2 (background: no background or background) by 4 (moral advice: control, rule, identity, or role) between-subjects design. Participants were randomly assigned to one of the eight conditions. In all conditions, a NAO robot was introduced as a research assistant.

The study procedures were same as those adopted in Study 2's robot conditions except for the following changes. First, participants were randomly assigned to either the no background condition or the background condition. In the no background condition, prior information about the robot's capacities was not presented to participants. Hence, the procedure was the same as in Study 2's robot conditions except that the MMP35 questionnaire was administered at the end of the study. In the background condition, the writ-

ten descriptions of social and moral capacities of the robot were presented after the informed consent form was signed and before the first video clip of a NAO robot was presented. Thus, when participants watched the first video of the robot in which the robot introduced itself as a research assistant, it was after they read about what the robot was capable of doing.

Second, we asked participants to answer the MMP35 questionnaire immediately after they completed the die-rolling game and reported the number they threw and the bonus payoff they earned from the task. The rest of the measures about participants' familiarity with AI and robots and basic demographic questions were presented afterwards.

Finally, we changed how we screened participants based on the audio and video systems of their computers. In Studies 1 and 2, we let all participants complete the study regardless of whether they had successfully passed the audio and video check questions. For this reason, we had to remove a large number of participants who did not pass either or both of the audio and video check questions before analyzing data (29 participants in Study 1 and 214 participants in Study 2). In Study 3, before presenting the consent form, we asked participants to answer those questions along with a statement that incorrect answers would lead to an immediate termination of the study. Therefore, only the participants who passed both questions were able to move on to the page where the consent form was presented.

5.2. Data Analysis and Results

5.2.1. The effects of describing moral capacities on perceptions of a robot

We first examined whether portraying the robot as a social robot programmed to have knowledge of moral norms, follow norms, and encourage others to follow norms influenced participants' perceptions of the robot by analyzing the MMP35 data. We performed a Multivariate Analysis of Variance (MANOVA) with background (with vs. without) and moral advice (control, rule, identity, role) as two between-subject variables on the five subscales of MMP35. We found a significant main effect of Background, Pillai's Trace $= 0.02$, $F(5, 820) = 2.90$, $p = .013$. Specifically, this effect of Background was driven by the differences in perceptions of morality, $F(1, 824) = 4.03$, $p = .045$. Participants reported to have perceived greater moral capacity when the background information about the robot was provided ($M = 4.21$, $SD = 2.02$) than when no background information was provided ($M = 3.93$, $SD = 2.01$). Therefore, descriptions of the robot's moral capacities successfully induced increased ascriptions of moral capacities to the robot.

5.2.2. The effects of a robot's moral advice on cheating

We next performed logistic regression analyses, separately for the no background and the background conditions, based on moral advice in order to predict the probability of cheating ('honest answers' coded as '0' and 'dishonest answers' as '1'). We treated the control condition in which no advice was offered as the reference level for moral advice to examine the relative effect of each of the rule-, identity-, and role-based advice in comparison to the control condition. When the background information about the robot's moral capacity was *not* provided, which is the same as the robot agent condition of Studies 1 and 2, we could not find evidence to suggest the effect of any of the moral advice on deterring cheating. There was no statistically significant difference in the probability of cheating between the control condition and either the rule-, the identity-, or the role-based advice condition ($p_{minimum} = .549$, See Figure 6).

Furthermore, when the background information about the robot's moral capacity was provided, none of the three differentially framed moral advice appeared to have reduced the probability of cheating than the control condition ($p_{minimum} = .361$, See Figure 6). Therefore, regardless of whether the robot was portrayed as having moral capacities or not, the probability of cheating when the robot offered moral advice was not significantly different from when the robot did not offer any advice.

5.2.3. The influence of perceived moral capacity of a robot on deterring cheating

Even when the robot was introduced as having moral capacities, the advice from the robot remained ineffective in discouraging cheating. These findings raised a follow-up question about the relationship between the extent to which participants ascribed moral capacities to the robot and their decisions to cheat or not. Although the nature of this question was exploratory, we had a general expectation that, the stronger participants viewed the robot as having moral capacities, the stronger the impact of the robot's advice would be, reducing the probability of cheating. Thus, a negative relationship between the perceived moral capacity and the probability of cheating was expected.

To address this question, we split the entire dataset into eight subsets by advice type and background information and, to each of the datasets, fit a logistic regression model with perceived moral capacity as a parameter. As a result, we discovered a specific relationship between perceived

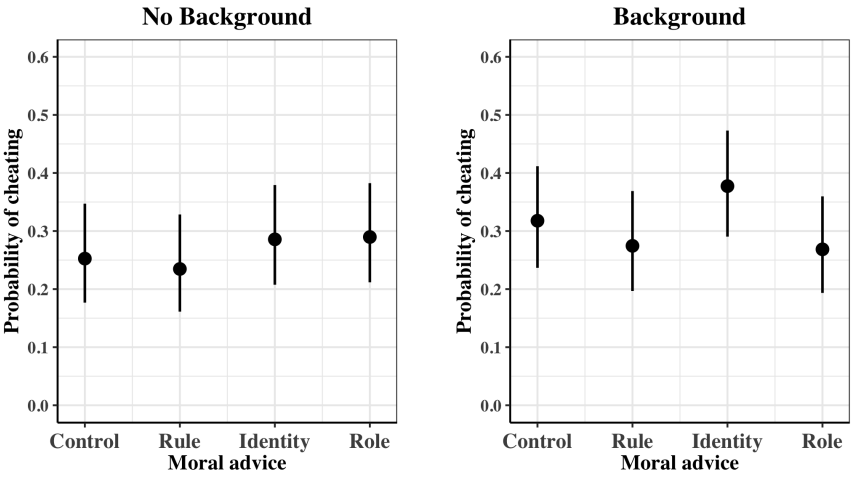

Figure 6: Probabilities of cheating as a function of background information about the robot's moral capacity (no background or background) and moral advice (control, rule, identity, or role) in Study 3. The error bars represent 95% confidence intervals.

moral capacity and the probability of cheating for certain conditions. In the *no background-control* condition, we found that the degree to which participants attributed moral capacities to the robot had a significant effect on their cheating behavior. However, in contrast to our expectation to find a negative relationship between perceived moral capacity and the probability of cheating, we uncovered a positive relationship indicating that, the more participants attributed moral capacities to the robot, the greater the probability of their cheating was, $b = 0.36$, $SE = 0.14$, $z = 2.54$, $p = .011$, OR = 1.43, 95% CI = [1.11, 1.95]. This positive relationship between perceived moral capacity and cheating was also significant when the robot offered identity-based advice. Specifically, we found significant positive relationships between perceived moral capacity and the probability of cheating in the *no background-identity* condition, $b = 0.47$, $SE = 0.15$, $z = 3.03$, $p = .002$, OR = 1.59, 95% CI = [1.21, 2.22] as well as in the *background-identity* condition, $b = 0.26$, $SE = 0.12$, $z = 2.10$, $p = .036$, OR = 1.29, 95% CI = [1.03, 1.67]. In the rest of the conditions — the *no background-rule* condition ($p = .198$), the *no background-role* condition ($p = .849$), the *background-*

control condition ($p = .362$), the *background-rule* condition ($p = .524$), the *background-role* condition ($p = .857$) — we did not find any statistically significant relationship between perceived moral capacity and the probability of cheating (See Figure 7).

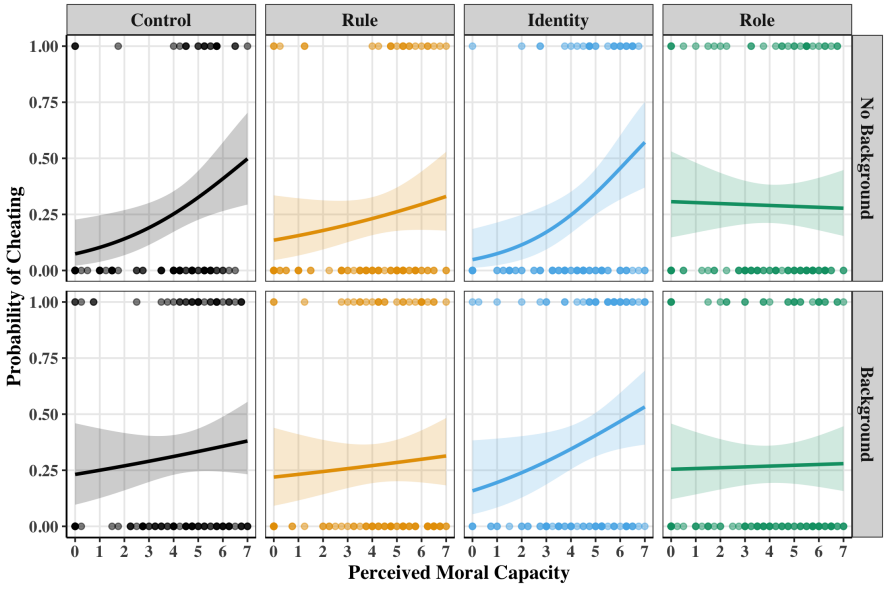

Figure 7: The relationship between participants' perception of the robot's moral capacity and probabilities of their cheating in Study 3. The error bars (shaded areas) represent 95% confidence intervals.

5.3. Discussion

The findings in Study 3 further supported that, when a robot offered moral advice, the advice was unlikely to be effective in discouraging participants from cheating. This lack of influence that the robot's moral advice exerted on people's decisions was present regardless of which ethical framework was underlying the advice and whether participants were informed of the robot's moral capacities. When descriptions of the robot's moral capacities were provided, it led to a stronger attribution of moral capacity to the robot than when no such descriptions were provided. However, even after the

robot was explicitly described as having moral capacities, we found none of the three different advice conditions to be more effective in deterring cheating than the control condition where no advice was offered.

Rather, when we performed exploratory analyses to better understand the relationship between participants' perceptions of the robot's moral capacity and the probability of their cheating, we discovered the findings that contradicted our expectations. We had expected a negative relationship between perceived moral capacity and the probability of cheating, but we found that perceived moral capacity either had a positive relationship with the probability of cheating in different pairings of advice type and background information, or had no effect on the probability of cheating. Specifically, when no prior information about the robot's capacity was available and the robot did not offer any advice, the more participants perceived the robot as having moral capacities, the more likely they were to cheat. This response pattern was also observed when the robot offered identity-based moral advice irrespective of whether prior information about the robot's capacity was available or not.

However, we did not find this resistance against the robot's moral advice when the robot was portrayed as having moral capacities but did not offer any moral advice. It appears that presenting the robot as a morally capable entity may not entirely reverse the reactance effect but at least dwindle its strength. We also did not observe the reactance when the robot offered either rule-based or role-based advice. Therefore, it is likely that, even if offering moral advice grounded in deontology or Confucian role ethics may not result in significantly less cheating than offering no advice, it may still reduce the strength of the resistance participants may express when they attribute moral capacities to the robot.

6. General Discussion

6.1. Deterrence effect of a robot's moral advice on cheating

The purpose of the present research was to investigate to what extent moral advice grounded in different ethical theories, which are deontological ethics, virtue ethics, and Confucian role ethics, could inhibit cheating when a robot proactively provided the advice to participants. Based upon the HHI literature on honesty (Bryan et al., 2013; Savir and Gamliel, 2019) and the recent HRI literature on the use of different ethical frameworks in persuasion (Wen et al., 2022b, 2023), we hypothesized that if participants

were receptive to a robot's moral advice as they were to a human's moral advice, identity-based advice grounded in virtue ethics and role-based advice grounded in Confucian role ethics would deter cheating, but rule-based advice grounded in deontological ethics would not.

Across three studies, however, we could not find evidence that a robot's moral advice, irrespective of its underlying ethical frameworks, deterred cheating. In Study 1, we found that the distributions of the percentages of the numbers participants claimed to have thrown in all three different advice conditions (rule, identity, role) and the control (no advice) condition significantly deviated from the uniform distribution. There also was no evidence to suggest that the percentages of claiming to have thrown the number resulting in the largest bonus payoff were lower in any of the three differentially-framed moral advice conditions than in the control condition. In Study 2's robot condition and Study 3's no background condition, both in which prior information about the robot's capacities was not provided to the participants, we found no evidence indicating that the percentages of cheating in any of the rule-, identity-, and role-based advice conditions were lower than the control condition. This ineffectiveness of a robot's moral advice on discouraging cheating persisted even when in Study 3's background condition, we informed participants about the robot's moral capacities before they encountered the robot. Therefore, the use of a robot moral advisor in taking preemptive measures by giving moral advice to deter cheating appears to require different approaches than the approach taken in the current research.

To better understand this lack of a robot moral advisor's persuasive power, we examined whether the ineffectiveness of a robot's moral advice was due to the content of the advice itself or participants' existing knowledge about a robot's capacities. We found from Study 2 that moral advice from a human agent was effective when it was grounded in Confucian role ethics, and it could also be potentially effective when the advice was grounded in deontological ethics. Thus, it did not appear to be reasonable to attribute the ineffectiveness of a robot's moral advice solely to the message itself lacking persuasive power. Also, in Study 3, we confirmed that when the robot was described as having moral capacities, participants judged the robot as having greater moral capacities than when no such descriptions about the robot were provided. However, despite this change in their perceived moral capacity of the robot, the probabilities of participants' cheating in any of the advice conditions were still not different from the no advice (control) condition. Therefore, it seemed unlikely that participants were not receptive

to the robot's advice solely because of their denial of any moral capabilities the robot may have.

There are other potential factors that could have hindered the influence of the robot's moral advice but left unaddressed as they were beyond the scope of the current experimental design. First, it would be worthwhile to, once again, consider the limitations of using pre-recorded video clips of a robot in an online experimental research to create HRI experiences. A limited presence of the robot in the current experimental setting could have obstructed participants from viewing the robot as a social actor. Considering the existing findings about how, in a laboratory experiment, a mere exposure to a robot that does not communicate verbally but has eyes could discourage cheating (Hoffman et al., 2015), it is possible that a robot's physical presence is crucial for effectively inducing honesty. Moreover, previous work suggests that constraints related to online experiments could be strong when a robot makes minimal interactions with participants or demonstrates minimal social behavior, such as a robot in a video merely looking and blinking its eyes in silence during an experiment (Petisca et al., 2020). However, contrasting these possibilities, other research showed that, with the effective use of a robot's verbal messages, a robot can influence participants' task performance even when its presence is limited to video clips presented in an online experiment (Wen et al., 2022b). We thus argue that a lack of a robot's presence in the current experimental setting may have partially contributed to the absence of the deterrence effect but not be a sufficient explanation for participants' resistance to a robot's moral advice.

Alternatively, it is possible that, when a robot guided participants through the experimental procedures and offered them a piece of moral advice, cheating could have been perceived as a less severe norm violation than in HHI contexts where a human served in the role the robot served. The strength of social norms can vary depending on one's belief about what ought to be done but also one's awareness of what others expect them to do (Gelfand et al., 2017; Bicchieri, 2005). The latter component, in particular, could have led participants in this research to perceive the norm of honesty as having a weaker strength when the experiment was led by a robot, rather than a human. Supporting this possibility, Hoffman et al. (2015) found that participants felt less guilty after completing a task that incentivized them to cheat when they had been monitored by a robot than by a human during the task. In a similar vein, Petisca et al. (2020) showed that participants expected feeling less guilty when they were to be dishonest to a robot than to

their brother, friend, or a stranger. Therefore, when moral advice is offered by a robot, for the advice to have a meaningful impact on cheating behavior, a stronger persuasive strategy may be needed than when moral advice is offered by a human possibly because the strength of the norm of honesty could be weaker in the context of HRI than HHI.

Third, a closer look at the philosophical foundations of the rule-based and the role-based advice conditions might also be helpful for explaining the ineffectiveness of a robot's moral advice in discouraging cheating. One of our major initial goals for Study 3 was to examine whether the perception of the ontological status of the robot like the possession of moral capacities would affect the efficacy of moral advice given by the robot. Returning to the initial goal of setting up a context for influencing how participants may perceive the ontological status of the robot, it is worth asking whether moral capabilities or a different concept, such as moral status, would in fact be more effective. Classic Kantian deontology was constructed based on such a belief that only rational beings can be called persons and their rationality constitutes the foundation of morality and mark them out as ends in themselves (Alexander and Moore, 2007; Briggle and Mitcham, 2012). When deontological principles are introduced by robots, the efficacy of these principles therefore would rely on to what extent humans consider robots as beings having *human alike properties* that will make them have human alike *moral status* (more than just moral capabilities). If the robot is not considered to possess some kind of moral status, then the deontological moral advice provided by the robot may not be convincing to participants, which probably could have led to the null effect of the rule-based advice when it was given by a robot.

Confucian role ethics has a different approach to discussing what counts as a person and the role of personhood in moral development. For Confucian role ethics, unlike Kantian deontology that would consider moral personhood as something depending on whether a robot demonstrates human alike properties, what matters more is whether nonhumans such as robots *work* to achieve their moral status by assuming ethically relevant roles and duties as humans (Wong and Wang, 2021; Cassauwers, 2019; Zhu, 2023). Therefore, for a robot to which participants have a transient exposure in an experiment like the one in the present research, it would not be clear to participants what efforts such a robot has put into living ethically relevant roles and duties as humans. This lack of explicit efforts to personhood development could have led participants to be not convinced by the moral advice given by the robot. Future research may consider exploring how different approaches to provid-

ing background information about the robot, such as moral capability and moral personhood, affect human perceptions of the ontological status of the robot and the effect of its moral advice on encouraging moral behavior.

6.2. Psychological reactance effect associated with the adoption of a robot moral advisor

Although the current research did not lead to evidence supporting the deterrence effect of a robot's moral advice on cheating, it instead led to an unexpected and potentially critical discovery. In Study 3, we found that when participants' exposure to the robot was held to a minimal level in that they received neither background information about the robot's moral capacities nor the robot's moral advice, the more the participants perceived a robot as having moral capacities, the more likely they were to cheat. These choice patterns contradicted the potential accounts we proposed to explain the findings in Studies 1 and 2 (Our initial explanation was that the robot's advice could have had no effect in those two studies because participants' weak perceptions of the robot's moral capacity had interfered with the expected deterrence effect). Further, in Study 3, the increased prevalence of cheating behavior among participants who more strongly attributed moral capacities to the robot was also observed when the robot issued moral advice grounded in virtue ethics, which is the identity-based advice. This trend emerged not only for the participants who did not receive any background information about the robot's capacities but also for the participants who did receive the information.

When there was a minimum exposure to the robot's capacities or when the robot preemptively offered moral advice highlighting the implications of cheating on participants' self-identity and moral character ("cheater"), it appeared that participants' decisions reflected a form of psychological reactance. Psychological reactance theory propounds that, when their freedom of behavior is threatened or eliminated, people become motivated to make attempts to restore freedom or perceive the lost option more attractive (Brehm, 1966; Brehm and Brehm, 2013; Rosenberg and Siegel, 2018; Steindl et al., 2015). When experiencing psychological reactance, people would experience negative emotions (e.g., anger) and thinking, and would respond to these negative experiences by engaging in behaviors opposite of the recommended ones or by counterarguing (Dillard and Shen, 2005; Rains, 2013). Our findings suggest that acknowledging or recognizing a robot as being equipped with moral capacities could induce participants to experience a sense of threat

to their freedom, which in turn motivates them to make choices *opposite* of the choice the robot exhorted. It is noteworthy that, in this research, the choice the robot encouraged participants to make was the *moral* choice (i.e., to be honest) and choosing to cheat implied that participants were willing to make the *immoral* choice in order to not follow the robot's advice. These reactions pointing toward the psychological reactance effect are also in line with the previous literature on ethical concerns in AI and robotics (Pauketat and Anthis, 2022; Złotowski et al., 2017) where it was discussed the potential for heightened perceived threats to the existence or survival of the humanity, as AI and robots become more and more competent. Thus, in the present work, when a robot guided participants through an experiment that had moral implications or when a robot issued *unsolicited* moral advice on how cheating would make them a morally corrupt person, participants could have felt that their uniqueness and identity as humans are threatened or anticipated competitions with robots over safety and resources necessary for human existence.

There have been reports on the psychological reactance effect in the existing HRI studies (Boos et al., 2022; Roubroeks et al., 2011; Giroux et al., 2022). For example, Roubroeks et al. (2011) observed an increased social agency of a robot resulting in a stronger, rather than weaker, psychological reactance. Specifically, these researchers showed that, when participants received advice on energy conservation written in text along with either a picture or a short video clip of a robot, their self-reports suggested stronger psychological reactance as measured as anger and negative thinking than when they received the same written advice but without any picture or video of a robot. Therefore, the present research expands the previous findings in HRI literature on the psychological reactance by demonstrating that people's psychological reactance against a robot's influence can also be applied to a basic social and moral norm, which is the norm of honesty. Moreover, these findings suggest that the psychological reactance effect can be strong to the extent that it cannot be easily swayed by moral advice grounded in specific ethical theories.

6.3. Potential strategies to alleviate the psychological reactance effect

When robots are deployed to persuade people to change their behavior in various HRI contexts, the current findings and the extant findings (Boos et al., 2022; Roubroeks et al., 2011; Giroux et al., 2022) together raise the psychological reactance effect as a potential barrier in facilitating successful

persuasive communication in HRI. Besides extending the findings about the psychological reactance effect to moral contexts in HRI, the present research also led to a discovery of possibly promising ways to mitigate these reactance effects. In Study 3, when the robot was described as having moral capacities (and did not proceed to offer moral advice grounded in virtue ethics), the psychological reactance effect associated with the perceived moral capacity was not observed. The psychological reactance effect was also absent when the robot offered moral advice grounded in certain ethical theories, which were Confucian role ethics and deontological ethics, regardless of the availability of the information about the robot's moral capacities. These findings suggest that participants' possible approval of rule-based values and the awareness of their moral relationship toward others in the community may potentially mitigate psychological reactance even when the role of an advice-giver was served by a robot. Admittedly, none of these factors that could have inhibited the psychological reactance effect seems to have been able to *reverse* the observed relationship between the perceived moral capacities and cheating. We did not find evidence for psychological reactance under certain conditions but also did not find that those conditions induce a lower probability of cheating for an increased perception of a robot's moral capacity.

However, the current findings of potential alleviation effects suggest that participants' receptivity to a robot's moral advice on the norm of honesty is malleable and that there is a positive outlook of the existence of effective approaches to promoting the norm of honesty via a robot's moral persuasion. Specifically, when deploying a robotic moral advisor in HRI, potential strategies to prevent the psychological reactance effect would be to prioritize first, communicating the robot's moral capacities than having it deliver moral advice and second, to carefully take into account specifically which ethical theories are chosen to formulate the message. In doing so, a caveat would be to avoid presuming that the findings from the extant HHI research (Fischbacher and Föllmi-Heusi, 2013; Bryan et al., 2013; Savir and Gamliel, 2019) would be readily extended to HRI contexts (e.g., moral advice grounded in virtue ethics was ineffective in the present research). Although these potential strategies are derived from the current research setting where the focus was on the moral domain, we view that these strategies could also be useful in future research for generating strategies to prevent or mitigate the psychological reactance effect when robots attempt to induce changes in people's thoughts and behaviors in other domains besides the moral domain.

6.4. The discrepancy between the current and the existing findings about the effects of moral advice grounded in virtue ethics

There were findings in the present research that were the exact opposite of the existing findings in the psychological literature (Fischbacher and Föllmi-Heusi, 2013; Bryan et al., 2013; Savir and Gamliel, 2019). In these previous studies, identity-based advice was shown to be effective in deterring cheating while rule-based advice was not, and this difference was interpreted as evidence for the effectiveness of highlighting one's self-identity in persuading people to act honestly. In the present work, however, when a human delivered the advice, it was rule-based advice that showed possible effectiveness, not identity-based advice.

To understand this discrepancy between the current and the prior findings, we first closely compared the experimental setups used in this work and the prior work (Bryan et al., 2013; Savir and Gamliel, 2019). In the current research, the advice was delivered to participants by an agent that was a visible and specific person, and this delivery was done in spoken language. In previous studies (Bryan et al., 2013; Savir and Gamliel, 2019), on the contrary, the advice was delivered to participants in written texts, and the agent that is delivering the message was not specified. In those studies, participants completed the study via either in-person or online but in either case, they received task instructions and moral messages on cheating by reading texts printed on paper or online webpage. We expected that an implicit mutual understanding between the research team and the participants in these previous studies was that the messages were originated from the researchers who are humans. Although this possibility was never verified, it would be unlikely that when reading those messages, participants imagined them coming from a robot or any other agents that are not human. For instance, in Experiment 2 of the prior work (Bryan et al., 2013), the message participants read included "NOTE: Please don't be a cheater...," which seems to rather emphasize the voice of the researchers.

These differences in the experimental setup and the findings between the current work and the previous studies (Fischbacher and Föllmi-Heusi, 2013; Bryan et al., 2013; Savir and Gamliel, 2019) suggest that it is necessary to refine the existing approaches to framing persuasive messages. Specifically, receiving a message like "lying will make you a cheater" in written format could be a subtle reminder to oneself of the implication of cheating with regards to their identity, but receiving the identical message in speech format from discernible others could invoke defensive reactions. It is possible that in

the current research, even though the messages were delivered via prerecorded video, having a specific agent that conveyed moral advice in spoken language could still have induced negative reactions when the advice was grounded in virtue ethics.

These possibilities would need to be examined in future work, but there has been some relevant evidence in the linguistics and communication literature. The existing socio- and psycho-linguistics literature on differences between spoken and written language (Akinnaso, 1982) has discussed a variety of aspects in which the two modes of language are different. For instance, in spoken language, emotions can be communicated not only through verbal components but also through non-verbal components, such as tone and prosody (Ross, 1981; Liebenthal et al., 2016). Moreover, the impact of language on emotion in communication can be different depending on whether the communication was done face-to-face or mediated through computers, such as email or videoconferencing (Kappas and Krämer, 2011). In this work, the video stimuli were used, instead of the face-to-face interaction, but it may be reasonable to suggest that being able to see a person or a robot that delivered the messages in video elicited subjectively different experiences than only being able to read the messages printed on paper or computer screen. Therefore, the present findings indicate that simply highlighting the relevance to the self in persuading people does not always lead to positive outcomes. In implementing theories of moral philosophy to promote social and moral norms, it would be critical to consider the mode of communication and its influences on interactants' emotional responses.

Additionally, we considered another explanation for the inefficacy of identity-based advice from theoretical perspectives. Theories of virtue ethics, specifically agent-based virtue ethics, encourage decision-makers observing and judging the human traits that are either admirable or abominable in particular *other* moral agents (Athanassoulis, 2013, n.d.). In the current experimental context, however, these relevant moral traits about others, such as other MTurkers, were not observable. Although this information was also not available in prior studies where identity-based advice was found to be effective (Bryan et al., 2013; Savir and Gamliel, 2019), our findings suggest the need to enhance the existing approaches to framing persuasive messages grounded on virtue ethics. To make the message like "it will make you a cheater" more convincing and relatable, in future work, we recommend adding more information about how honesty (or dishonesty) of a particular MTurker, who could be someone widely known in the community, was judged

to be admired (or abominable).

7. Conclusion

Social robots' capacities to communicate with humans via natural language potentiate the power of those robots' persuasive messages on human behavior. In this research, we examined whether social robots can successfully convince people to follow the norm of honesty by offering them moral advice of which the underlying ethical theories varied. We did not find empirical support for the positive effects of a robot's moral advice on discouraging cheating behavior. We instead found that, when a robot was perceived as having moral capacities, it can induce psychological reactance, but this psychological reactance effect may potentially be mitigated when the background information about the robot's capacities relevant to moral contexts was provided or the robot offered moral advice of which the framing was focused on deontology or Confucian role ethics. The current findings indicate that, despite their great potential to influence people's behavior, social robots and their language capabilities can engender unanticipated resistance that may induce people to make unethical choices like cheating. In future HRI research on moral persuasion, careful consideration would be required when introducing robots and their capabilities to people and selecting the ethical framing of the robots' messages.

Acknowledgement

This work was supported in part by the U.S. National Science Foundation Grant IIS-1909847 and in part by the U.S. Air Force Office of Scientific Research Grant 16RT0881f and FA9550-23-1-0036.

References

Akinnaso, F.N., 1982. On the differences between spoken and written language. Language and speech 25, 97–125.

Alexander, L., Moore, M., 2007. Deontological ethics .

Ames, R.T., 2011. Confucian role ethics: A vocabulary. The Chinese University of Hong Kong Press.

Athanassoulis, N., 2013. Virtue ethics. A&C Black.

Athanassoulis, N., n.d. Virtue ethics the internet encyclopedia of philosophy. https://iep.utm.edu/virtue/. Accessed: 2023-04-10.

Banks, J., Edwards, A.P., Westerman, D., 2021. The space between: Nature and machine heuristics in evaluations of organisms, cyborgs, and robots. Cyberpsychology, Behavior, and Social Networking 24, 324–331.

Bicchieri, C., 2005. The grammar of society: The nature and dynamics of social norms. Cambridge University Press.

Boos, A., Herzog, O., Reinhardt, J., Bengler, K., Zimmermann, M., 2022. A compliance–reactance framework for evaluating human-robot interaction. Frontiers in Robotics and AI 9.

Breazeal, C., 2004. Designing sociable robots. MIT press.

Brehm, J.W., 1966. A theory of psychological reactance. .

Brehm, S.S., Brehm, J.W., 2013. Psychological reactance: A theory of freedom and control. Academic Press.

Briggle, A., Mitcham, C., 2012. Ethics and science: An introduction. Cambridge University Press.

Briggs, G., Scheutz, M., 2014. How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress. International Journal of Social Robotics 6, 343–355.

Briggs, G., Williams, T., Jackson, R.B., Scheutz, M., 2022. Why and how robots should say 'no'. International Journal of Social Robotics 14, 323–339.

Bryan, C.J., Adams, G.S., Monin, B., 2013. When cheating would make you a cheater: implicating the self prevents unethical behavior. Journal of Experimental Psychology: General 142, 1001.

Cappuccio, M.L., Sandoval, E.B., Mubin, O., Obaid, M., Velonaki, M., 2021. Can robots make us better humans? International Journal of Social Robotics 13, 7–22.

Cassauwers, T., 2019. How confucianism could put fears about artificial intelligence to bed. Accessed: 2022-12-27.

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., Ford, C., Volcic, R., De Rosario, H., De Rosario, M.H., 2018. Package 'pwr'. R package version 1.

Darling, D.A., 1957. The kolmogorov-smirnov, cramer-von mises tests. The Annals of Mathematical Statistics 28, 823–838.

Dautenhahn, K., 2007. Methodology & themes of human-robot interaction: A growing research field. International Journal of Advanced Robotic Systems 4, 15.

DeBaets, A.M., 2014. Can a robot pursue the good? exploring artificial moral agency. Journal of Ethics and Emerging Technologies 24, 76–86.

Demidenko, E., 2007. Sample size determination for logistic regression revisited. Statistics in medicine 26, 3385–3397.

Dillard, J.P., Shen, L., 2005. On the nature of reactance and its role in persuasive health communication. Communication monographs 72, 144–168.

Duffy, B.R., Rooney, C., O'Hare, G.M., O'Donoghue, R., 1999. What is a social robot?, in: 10th Irish Conference on Artificial Intelligence & Cognitive Science, University College Cork, Ireland, 1-3 September, 1999.

Engmann, S., Cousineau, D., 2011. Comparing distributions: the two-sample anderson-darling test as an alternative to the kolmogorov-smirnoff test. Journal of applied quantitative methods 6.

Fischbacher, U., Föllmi-Heusi, F., 2013. Lies in disguise—an experimental study on cheating. Journal of the European Economic Association 11, 525–547.

Gelfand, M.J., Harrington, J.R., Jackson, J.C., 2017. The strength of social norms across human groups. Perspectives on Psychological Science 12, 800–809.

Gelman, A., Jakulin, A., Pittau, M.G., Su, Y.S., 2008. A weakly informative default prior distribution for logistic and other regression models .

Giroux, M., Kim, J., Lee, J.C., Park, J., 2022. Artificial intelligence and declined guilt: Retailing morality comparison between human and ai. Journal of Business Ethics 178, 1027–1041.

Giubilini, A., Savulescu, J., 2018. The artificial moral advisor. the "ideal observer" meets artificial intelligence. Philosophy & technology 31, 169–188.

Halevy, R., Shalvi, S., Verschuere, B., 2014. Being honest about dishonesty: Correlating self-reports and actual lying. Human Communication Research 40, 54–72.

Harrell, Jr, F.E., Harrell, F.E., 2015. Binary logistic regression. Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis , 219–274.

Haslam, N., 2006. Dehumanization: An integrative review. Personality and social psychology review 10, 252–264.

Hegel, F., Muhl, C., Wrede, B., Hielscher-Fastabend, M., Sagerer, G., 2009. Understanding social robots, in: 2009 Second International Conferences on Advances in Computer-Human Interactions, IEEE. pp. 169–174.

Hertz, N., Wiese, E., 2019. Good advice is beyond all price, but what if it comes from a machine? Journal of Experimental Psychology: Applied 25, 386.

Hoffman, G., Forlizzi, J., Ayal, S., Steinfeld, A., Antanitis, J., Hochman, G., Hochendoner, E., Finkenaur, J., 2015. Robot presence and human honesty: Experimental evidence, in: 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE. pp. 181–188.

Hovland, C.I., Weiss, W., 1951. The influence of source credibility on communication effectiveness. Public opinion quarterly 15, 635–650.

Hursthouse, R., 1999. On virtue ethics. OUP Oxford.

Jackson, R.B., Wen, R., Williams, T., 2019. Tact in noncompliance: The need for pragmatically apt responses to unethical commands, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 499–505.

Jackson, R.B., Williams, T., 2019. Language-capable robots may inadvertently weaken human moral norms, in: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE. pp. 401–410.

Kahn Jr, P.H., Friedman, B., Hagman, J., 2002. " i care about him as a pal" conceptions of robotic pets in online aibo discussion forums, in: CHI'02 Extended Abstracts on Human Factors in Computing Systems, pp. 632–633.

Kahn Jr, P.H., Friedman, B., Perez-Granados, D.R., Freier, N.G., 2004. Robotic pets in the lives of preschool children, in: CHI'04 extended abstracts on Human factors in computing systems, pp. 1449–1452.

Kahn Jr, P.H., Kanda, T., Ishiguro, H., Gill, B.T., Ruckert, J.H., Shen, S., Gary, H.E., Reichert, A.L., Freier, N.G., Severson, R.L., 2012. Do people hold a humanoid robot morally accountable for the harm it causes?, in: Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction, pp. 33–40.

Kappas, A., Krämer, N.C., 2011. Face-to-face communication over the Internet: emotions in a web of culture, language, and technology. Cambridge University Press.

Kim, B., Wen, R., Zhu, Q., Williams, T., Phillips, E., 2021. 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, pp. 10–18.

Liebenthal, E., Silbersweig, D.A., Stern, E., 2016. The language, tone and prosody of emotions: neural substrates and dynamics of spoken-word emotion perception. Frontiers in neuroscience 10, 506.

Lilliefors, H.W., 1967. On the kolmogorov-smirnov test for normality with mean and variance unknown. Journal of the American statistical Association 62, 399–402.

Looije, R., Neerincx, M.A., Cnossen, F., 2010. Persuasive robotic assistant for health self-management of older adults: Design and evaluation of social behaviors. International Journal of Human-Computer Studies 68, 386–397.

Maeda, R., Brščić, D., Kanda, T., 2021. Influencing moral behavior through mere observation of robot work: Video-based survey on littering behavior, in: Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction, pp. 83–91.

Malle, B., 2019. How many dimensions of mind perception really are there?, in: CogSci, pp. 2268–2274.

Malle, B.F., 2016. Integrating robot ethics and machine morality: the study and design of moral competence in robots. Ethics and Information Technology 18, 243–256.

Malle, B.F., Scheutz, M., 2020. Moral competence in social robots, in: Machine ethics and robot ethics. Routledge, pp. 225–230.

Malle, B.F., Scheutz, M., Arnold, T., Voiklis, J., Cusimano, C., 2015. Sacrifice one for the good of many? people apply different moral norms to human and robot agents, in: 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE. pp. 117–124.

Massey Jr, F.J., 1951. The kolmogorov-smirnov test for goodness of fit. Journal of the American statistical Association 46, 68–78.

Melson, G.F., Kahn Jr, P.H., Beck, A., Friedman, B., Roberts, T., Garrett, E., Gill, B.T., 2009. Children's behavior toward and understanding of robotic and living dogs. Journal of Applied Developmental Psychology 30, 92–102.

Momen, A., De Visser, E., Wolsten, K., Cooley, K., Walliser, J., Tossell, C.C., 2023. Trusting the moral judgments of a robot: perceived moral competence and humanlikeness of a gpt-3 enabled ai .

Mubin, O., Cappuccio, M., Alnajjar, F., Ahmad, M.I., Shahid, S., 2020. Can a robot invigilator prevent cheating? AI & SOCIETY 35, 981–989.

Ni, P., 2016. Confucius: the Man and the Way of Gongfu. Rowman & Littlefield.

Nuyen, A.T., 2007. Confucian ethics as role-based ethics. International philosophical quarterly 47, 315–328.

Pauketat, J.V., Anthis, J.R., 2022. Predicting the moral consideration of artificial intelligences. Computers in Human Behavior 136, 107372.

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., Damer, E., 2022. Data quality of platforms and panels for online behavioral research. Behavior Research Methods , 1.

Petisca, S., Leite, I., Paiva, A., Esteves, F., 2022. Human dishonesty in the presence of a robot: The effects of situation awareness. International Journal of Social Robotics , 1–12.

Petisca, S., Paiva, A., Esteves, F., 2020. The effect of a robotic agent on dishonest behavior, in: Proceedings of the 20th ACM international conference on intelligent virtual agents, pp. 1–6.

van der Ploeg, T., Austin, P.C., Steyerberg, E.W., 2014. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC medical research methodology 14, 1–13.

Rains, S.A., 2013. The nature of psychological reactance revisited: A meta-analytic review. Human Communication Research 39, 47–73.

Ramsey, J., 2016. Confucian role ethics: A critical survey. Philosophy Compass 11, 235–245.

Roizman, M., Hoffman, G., Ayal, S., Hochman, G., Tagar, M.R., Maaravi, Y., 2016. Studying the opposing effects of robot presence on human corruption, in: 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE. pp. 501–502.

Rosemont Jr, H., Ames, R.T., 2016. Confucian role ethics: A moral vision for the 21st century? V&R unipress GmbH.

Rosenberg, B.D., Siegel, J.T., 2018. A 50-year review of psychological reactance theory: Do not read this article. Motivation Science 4, 281.

Ross, E.D., 1981. The aprosodias: Functional-anatomic organization of the affective components of language in the right hemisphere. Archives of neurology 38, 561–569.

Roubroeks, M., Ham, J., Midden, C., 2011. When artificial social agents try to persuade people: The role of social agency on the occurrence of psychological reactance. International Journal of Social Robotics 3, 155–165.

Salomons, N., Sebo, S.S., Qin, M., Scassellati, B., 2021. A minority of one against a majority of robots: Robots cause normative and informational conformity. ACM Transactions on Human-Robot Interaction (THRI) 10, 1–22.

Saunderson, S.P., Nejat, G., 2021. Persuasive robots should avoid authority: The effects of formal and real authority on persuasion in human-robot interaction. Science robotics 6, eabd5186.

Savir, T., Gamliel, E., 2019. To be an honest person or not to be a cheater: Replicating the effect of messages relating to the self on unethical behaviour. International Journal of Psychology 54, 650–658.

Savulescu, J., Maslen, H., 2015. Moral enhancement and artificial intelligence: moral ai?, in: Beyond artificial intelligence. Springer, pp. 79–95.

Shalvi, S., Dana, J., Handgraaf, M.J., De Dreu, C.K., 2011. Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. Organizational behavior and human decision processes 115, 181–190.

Shalvi, S., Eldar, O., Bereby-Meyer, Y., 2012. Honesty requires time (and lack of justifications). Psychological science 23, 1264–1270.

Shinozawa, K., Naya, F., Yamato, J., Kogure, K., 2005. Differences in effect of robot and screen agent recommendations on human decision-making. International journal of human-computer studies 62, 267–279.

Steindl, C., Jonas, E., Sittenthaler, S., Traut-Mattausch, E., Greenberg, J., 2015. Understanding psychological reactance. Zeitschrift für Psychologie .

Straßmann, C., Grewe, A., Kowalczyk, C., Arntz, A., Eimler, S.C., 2020. Moral robots? how uncertainty and presence affect humans' moral decision making, in: International Conference on Human-Computer Interaction, Springer. pp. 488–495.

Weisman, K., 2022. Extraordinary entities: Insights into folk ontology from studies of lay people's beliefs about robots, in: Proceedings of the Annual Meeting of the Cognitive Science Society.

Wen, R., Han, Z., Williams, T., 2022a. Teacher, teammate, subordinate, friend: Generating norm violation responses grounded in role-based relational norms, in: 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE. pp. 353–362.

Wen, R., Kim, B., Phillips, E., Zhu, Q., Williams, T., 2022b. Comparing norm-based and role-based strategies for robot communication of role-grounded moral norms. ACM Transactions on Human-Robot Interaction .

Wen, R., Kim, B., Phillips, E., Zhu, Q., Williams, T., 2023. On further reflection... moral reflections enhance robotic moral persuasive capability, in: International Conference on Persuasive Technology, Springer. pp. 290–304.

Williams, T., Zhu, Q., Wen, R., de Visser, E.J., 2020. The confucian matador: three defenses against the mechanical bull, in: Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, pp. 25–33.

Winkle, K., Lemaignan, S., Caleb-Solly, P., Leonards, U., Turton, A., Bremner, P., 2019. Effective persuasion strategies for socially assistive robots, in: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE. pp. 277–285.

Winkle, K., Melsión, G.I., McMillan, D., Leite, I., 2021. Boosting robot credibility and challenging gender norms in responding to abusive behaviour: A case for feminist robots, in: Companion of the 2021 ACM/IEEE international conference on human-robot interaction, pp. 29–37.

Wong, P.H., Wang, T.X., 2021. Harmonious technology: A Confucian ethics of technology. Routledge.

Zhu, Q., 2020. Ethics, society, and technology: A confucian role ethics perspective. Technology in society 63, 101424.

Zhu, Q., 2023. Just hierarchy and the ethics of artificial intelligence: Two approaches to a relational ethic for artificial intelligence. Ethical Perspectives 30, 59–000.

Złotowski, J., Yogeeswaran, K., Bartneck, C., 2017. Can we control it? autonomous robots threaten human identity, uniqueness, safety, and resources. International Journal of Human-Computer Studies 100, 48–54.

Declaration of Interest Statement

Declaration of interests

☐The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☒The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Boyoung Kim reports financial support was provided by National Science Foundation. Tom Williams reports financial support was provided by National Science Foundation. Qin Zhu reports financial support was provided by National Science Foundation. Ruchen Wen reports financial support was provided by National Science Foundation. Boyoung Kim reports financial support was provided by Air Force Office of Scientific Research. Ewart J. de Visser reports financial support was provided by Air Force Office of Scientific Research. Elizabeth Phillips reports financial support was provided by Air Force Office of Scientific Research. Tom Williams reports financial support was provided by Air Force Office of Scientific Research. Qin Zhu reports financial support was provided by Air Force Office of Scientific Research.