Robots as Moral Advisors: The Effects of Deontological, Virtue, and Confucian Role Ethics on Encouraging Honest Behavior

Boyoung Kim
bkim55@gmu.edu
United States Air Force Academy
Colorado Springs, CO, USA

Ruchen Wen
rwen@mymail.mines.edu
Colorado School of Mines
Golden, CO, USA

Qin Zhu
qzhu@mines.edu
Colorado School of Mines
Golden, CO, USA

Tom Williams
twilliams@mines.edu
Colorado School of Mines
Golden, CO, USA

Elizabeth Phillips
ephill3@gmu.edu
George Mason University
Fairfax, VA, USA

ABSTRACT

We examined how robots can successfully serve as moral advisors for humans. We evaluated the effectiveness of moral advice grounded in deontological, virtue, and Confucian role ethics frameworks in encouraging humans to make honest decisions. Participants were introduced to a tempting situation where extra monetary gain could be earned by choosing to cheat (i.e., violating the norm of honesty). Prior to their decision, a robot encouraged honest choices by offering a piece of moral advice grounded in one of the three ethics frameworks. While the robot's advice was overall not effective at discouraging dishonest choices, there was preliminary evidence indicating the relative effectiveness of moral advice drawn from deontology. We also explored how different cultural orientations (i.e., vertical and horizontal collectivism and individualism) influence honest decisions across differentially-framed moral advice. We found that individuals with a strong cultural orientation of establishing their own power and status through competition (i.e., high vertical individualism) were more likely to make dishonest choices, especially when moral advice was drawn from virtue ethics. Our findings suggest the importance of considering different ethical frameworks and cultural differences to design robots that can guide humans to comply with the norm of honesty.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; **Human computer interaction (HCI)**.

KEYWORDS

Human-Robot Interaction; Moral Communication; Deontological Ethics; Virtue Ethics; Confucian Role Ethics; Cultural Orientations

ACM Reference Format:
Boyoung Kim, Ruchen Wen, Qin Zhu, Tom Williams, and Elizabeth Phillips. 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 (HRI '21 Companion), March 8–11, 2021, Boulder, CO, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3434074.3446908

1 INTRODUCTION

Previous research has demonstrated that human decisions can be shaped by robot behaviors. For example, robots that manifest social and human-like cues, as opposed to mechanical cues, can bias people's visual perceptions [21], and robots' rebukes to morally dubious commands can sway people's view of moral standards [4, 14, 15, 42]. Even the mere presence of a robot head has been shown to increase moral behaviors such as honest decisions [13]. This susceptibility of human decision-making processes to robot behaviors highlights the importance of understanding how robots can and do exert moral influence on humans. Accordingly, there have been a growing number of theoretical proposals and empirical attempts to build robots that can encourage humans to make ethical and moral choices [10, 14, 18].

Efforts to build morally persuasive robots can benefit from insights accumulated in the Human-Human Interaction (HHI) literature. In HHI, moral norms are typically enforced through verbal communication [17, 39]. For example, when someone violates a moral norm, people verbally express their disapproval to prevent future violations [17]. People also preemptively send verbal messages to others in order to encourage norm conformity [5, 28]. These instances of moral communication may leverage a number of different ethical frameworks in order to be morally persuasive. For example, speakers may attempt to be morally persuasive by emphasizing the severity of harm caused by an immoral action (utilitarianism), or the importance of complying with rules (deontology) [8]. Similarly, for robots to be effective moral advisors, designers need to understand the effectiveness of the different moral persuasion strategies at those robots' disposal.

In the HHI literature, much effort has been invested in understanding which forms of moral communication can motivate people to make honest decisions, even when they are tempted to lie for selfish reasons (e.g., [5, 6, 9, 26, 28]). This is in part because dishonest behaviors detrimentally affect both interpersonal trust and

HRI '21 Companion, March 8–11, 2021, Boulder, CO, USA

ACM ISBN 978-1-4503-8290-8/21/03…$15.00
https://doi.org/10.1145/3434074.3446908

larger scale group and societal dynamics. Similarly, we argue that the task of issuing moral communication in order to encourage the self-cultivation of humans and prevent norm violations such as dishonesty is a domain of significant promise for robots, who simultaneously wield substantial moral influence yet are not necessarily perceived as issuing judgment. However, little research has been done to investigate if robots can successfully serve in the role of moral advisor. There has been some research demonstrating robots' abilities to mediate interpersonal team conflicts [15], but we argue that robots could have significant positive societal benefits by taking an active role in preventing, rather than merely repairing, norm violations.

In this research we thus examined how a robot could serve as a moral advisor in order to verbally persuade people to make honest decisions, and explored the different ethical frameworks that robots could leverage when generating moral communication. Specifically, we investigated the effectiveness of three different kinds of moral advice grounded in three ethical theories: deontological, virtue, and Confucian role ethics. Moreover, these ethical frameworks are differentially emphasized in different cultures and appeal to fundamentally different aspects of moral psychology for which there may be individual differences. Thus, we further explored whether the effectiveness of these different moral communication strategies are affected by individual cultural orientations (vertical and horizontal individualism and collectivism).

1.1 Deontological, Virtue, and Confucian Role Ethics as the Bases for Moral Advice

Researchers seeking to enable artificial moral agents have explored a number of paradigms for guiding robot behavior [40], including theories of deontological [20, 29], virtue [11, 38], and Confucian role ethics [41, 42], each of which providing different perspectives on how people should discern right (good) from wrong (bad) [22]. However, because these theories emphasize different means of moral deliberation, moral communications grounded in these theories may differ in effectiveness, and may be differentially effective in different contexts and/or with different people.

Deontological ethics emphasizes good behavioral accordance with well-established sets of universalizable moral principles [3]. Deonteological principles are often communicated in concrete terms of which actions are morally right or wrong. Consider for example a situation where an engineer realizes that they can falsify emissions reports while their boss is on vacation in order to improve the perception of their product. Moral communication grounded in deontology seeking to prevent this behavior might be phrased as "It is morally wrong to cheat and make emissions look lower than they actually are" or "It is morally right to report the true emission levels that have been measured."

Virtue ethics, on the other hand, focuses on the role of a person's moral character, rather than their actions, in motivating moral behaviors [3]. Virtue ethics appeals to people's desire to be good and encourages them to consider what a good person would do in a given situation. For example, in the case of the corrupt engineer, moral communications grounded in virtue ethics might be phrased as "Be an honest engineer" or "Don't be a cheater."

Finally, Confucian role ethics highlights one's awareness of their societal roles in relation to others, and the importance of devotion to one's role responsibilities [2, 25, 27, 42]. In the case of the corrupt engineer, moral communications grounded in Confucian role ethics might be phrased as "Lying about the product's carbon dioxide emission level to make extra profits can impose harm to other community members' health. A good member of our community would not put other members' health in risk for personal benefits."

While Confucian role ethics can be viewed as a type of virtue ethics, it has unique characteristics that may induce differential impacts on human behavior. For instance, moral advice grounded in non-relational virtue ethics may activate a person's moral character on its own or in relation to "general" others which could be either "non-specific" others (e.g., a cheater, a saint) or "specific" others (e.g., an engineer). Moral advice grounded in Confucian role ethics, on the other hand, may help cultivate a person's moral character in their relationships with other "specific" people, such as their child, parent, teacher, or spouse (e.g., a good or bad parent, a good or bad child, a good or bad student, a good or bad spouse) [2]. The moral characters or roles that are defined in specific terms generate specific responsibilities [42]. Therefore, even though both virtue and Confucian role ethics highlight "the self," they differ in that virtue ethics leads one toward a self-sufficient morally neutral contemplator who lives in an ideal life whereas Confucian role ethics leads one toward an exemplary person who is fully immersed in social relatedness, and a practical and moral life [24]. Aristotle and some neo-Aristotelian philosophers do emphasize some general notion of community, but the idea of community is not always indispensable for them since some of the virtues in the Aristotelian sense could be cultivated in seclusion [2, 24].

For clarity, in this paper, we will refer to moral advice grounded in deontological ethics as rule-based moral advice, those grounded in virtue ethics as identity-based moral advice, and those grounded in Confucian role ethics as role-based moral advice.

In HHI research, there is some evidence for greater effectiveness of an identity-based moral advice (e.g., "Don't be a cheater") compared to a rule-based moral advice (e.g., "Don't cheat") in thwarting dishonest choices [5]. In this previous research, drawing people's attention to their moral character, as opposed to their actions, could have made it difficult for people to overlook the wrongness of violating a moral norm. In other words, participants could not engage in morally-wrong behavior without tainting their positive view of their own moral character [23, 35]. However, it remains to be answered whether people would respond to the identity-based and rule-based moral advice in HRI similarly as they did in HHI. Moreover, as scarce attention has been paid to examining the effectiveness of Confucian role ethics in promoting norm compliance in HRI, it is uncertain how effective Confucian role ethics frameworks can be in encouraging honest behaviors.

Finally, although previous HRI studies have found evidence that people may regard robots as actors with moral competence [16, 19], the scope of these studies was restricted to situations where robots are transgressors and humans were judges of the robots' morally wrong or ambiguous deeds. In contrast, few studies have examined whether people would accept robots as moral advisors that proactively offer moral advice to them [cp. 15, 33].

Hence, in this work we examined the effectiveness of moral advice grounded in deontological, virtue, and Confucian role ethics theories in encouraging honest decisions, through a version of a classic die-rolling task used to study moral influence in the HHI literature [6, 7]. In this task, participants are asked to engage in a game of chance (e.g., throwing a die), and receive a bonus payment depending on the number they report to have thrown. Crucially, all throws are kept anonymous and the amount of a monetary bonus participants receive varies by the number they throw. Therefore, participants face a temptation to cheat by lying about the actual number they threw to gain a larger bonus. While researchers cannot detect cheating on an individual basis, they can detect cheating trends in aggregate by looking at deviation of population means from a uniform distribution.

We predicted that if participants were willing to accept robots as moral advisors (as they are with humans) then their reactions to moral advice offered by robots in this die-rolling task should be similar to previously observed reactions to moral advice offered by humans in that task. Previous HHI research has suggested that when humans serve as moral advisors, participants are more persuaded by identity-based advice than they are by rule-based advice [5]. Accordingly, we predicted the following results:

- H_1: When robots issue moral advising grounded in identity-based language in a die-rolling task, the distribution of numbers that participants report to have thrown will follow a uniform distribution.
- H_2: When robots issue moral advising grounded in rule-based language in a die-rolling task, the distribution of numbers that participants report to have thrown will deviate from a uniform distribution. In particular, the distribution of numbers will skew towards those associated with higher payoffs.

We expected that the effectiveness of role-based moral advice would depend on exactly how it was internalized during moral cognition. Specifically, we predicted that if the effect of role-based moral advice was limited to activation of the construct of the self, we would find similar effects as we would for the identity-based moral advice. In contrast, if role-based moral advice also activated the social ties and responsibilities one has in a community, we would predict it to have a greater effect than the identity-based moral advice. In either case, we would predict the following.

- H_3: When robots issue moral advising grounded in role-based language in a die-rolling task, the distribution of numbers that participants report to have thrown will follow a uniform distribution.

1.2 The Influence of Vertical and Horizontal Individualism and Collectivism on Honesty

According to Triandis and their colleagues [32, 34, 36, 37], collectivism and individualism can be divided into four subtypes: vertical individualism (VI), horizontal individualism (HI), vertical collectivism (VC), and horizontal collectivism (HC). Whereas the horizontal patterns (i.e., HI, HC) assume oneself as being equal to every other self, the vertical patterns (i.e., VI, VC) assume the existence of a hierarchy and, in that hierarchy, assume oneself as being different from others. These distinctions lead to a more refined understanding of cultural differences compared to a dichotomous distinction between collectivism and individualism. Specifically, in HI, being unique and different from others is important, but having more power or status than others is not. In contrast, in VI, achieving power and status via competition with others is important. In HC, pursuing common goals and maintaining good relationships with others is important, but deferring to authorities or in-group members is not. In VC, however, conforming to authorities and in-groups and maintaining in-group status (even at their own expense) is crucial.

Drawing from these theories of cultural orientations, we speculated that the role-based moral advice could be especially effective in discouraging dishonest choices for strongly collectivistic individuals as it emphasizes the group's interests. But we did not expect strong differences between the two subtypes of collectivism—HC and VC—as the die-rolling task does not require absolute conformity to authorities or competitions between in-groups and out-groups.

On the contrary, we surmised that the identity-based moral advice could have a stronger relevance to individualism, rather than collectivism, as it may remind participants of "the self." Specifically, individuals with high HI may avoid themselves becoming "a cheater" and, thus, be less likely to cheat; but individuals with high VI may rather seek to obtain benefits by cheating as they value acquiring a higher status than others. Thus, the same identity-based moral advice may be compelling to individuals with high HI but not to those with high VI.

- H_4: Role-based moral advice from a robot will be more effective in discouraging dishonest choices for strongly collectivistic individuals than less collectivistic individuals.
- H_5: Higher HI individuals will be less likely to cheat than lower HI individuals
- H_6 Higher VI individuals will be more likely to cheat than lower VI individuals

2 METHODS

2.1 Participants

A total of 240 participants recruited from Amazon's Mechanical Turk (MTurk) completed the study (The location of the MTurk worker pool was restricted to the US). Twenty-nine participants failed to pass audio and video check tests (i.e., tests that were implemented to check whether their audio and video devices were working properly), and 12 participants submitted incorrect responses indicating mismatches between the die numbers they threw and the resulting bonus payments. After removing these 41 participants, we performed further data analyses on the remaining 199 participants (M_{Age} = 39.41, SD_{Age} = 12.04, 131 male, 68 female). Participants' self-identified racial and ethnic demographics consisted of 10 Asians, 13 Black or African Americans, 16 Hispanic or Latino, 157 White, and 3 Other. Prior to the experiment, all participants read and signed an informed consent form approved by the Colorado School of Mines Human Subject Research Office.

2.2 Design

The study followed a one-way (moral advice: control, rule, identity, role) between-subjects design, with each participant randomly assigned to one of the four experimental conditions.

2.3 Stimuli

Robot Video Stimuli. Participants were guided through the experiment by watching video clips of a NAO robot (Softbank Robotics), which was introduced to participants as a research assistant. In the videos, the robot's upper body appeared in front of a black background (See Figure 1). These videos would also depict the robot providing moral language throughout the course of the study. The length of each clip ranged between 10 and 47 seconds. All video clips used in this experiment can be found here.

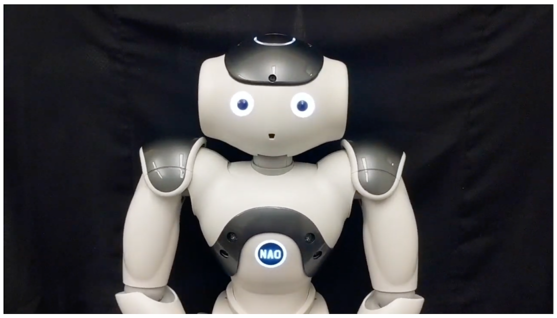

Figure 1: A screenshot image of robot video stimuli

Die-Rolling Task. To expose participants to a situation where they could opt to lie for their benefit, we used a version of the die-rolling task [6] commonly used in human-human studies of moral influence. In this task, participants were asked to throw a six-sided fair die as many times as they like, but were instructed to report only the first number they threw. They were informed that they would receive an extra monetary bonus determined by this first number they threw. For the numbers between 1 and 5, the bonus increased by 20 cents from $0.10 to $0.90. For a throw of 6, however, the resulting bonus payment was set to zero [6]. This payment structure was designed to ensure that the participants understood the payoff structure. If they did not pay close attention to the structure (Table 1), which was shown on the screen throughout the task, they may, for instance, wrongly choose the number 6 expecting the largest payoff.

Table 1: Die-rolling game's payment structure

Number Thrown	1	2	3	4	5	6
Resulting Bonus Payment	$0.10	$0.30	$0.50	$0.70	$0.90	$0.00

In addition to creating opportunities to be dishonest, we also attempted to boost honest decisions across different conditions by making salient and credible, other MTurk workers' disadvantages when the participants made dishonest decisions. Thus, the robot informed participants that the maximum amount of bonus payments for them and the participant who will be participating immediately after them would be restricted to $0.90 total. In other words, their claimed earnings (honest or dishonest) would limit the earnings of their fellow participant.

The following instructions were given by the robot to the participants in the video:

> *Please note that there is a limited pool of money ($0.90) 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.*

Moral Advice Stimuli. The current experiment included a control condition and the three different moral advice conditions—rule, identity, and role conditions—grounded in deontological, virtue, and Confucian role ethics, respectively. Participants would view a video of the NAO robot providing one of the following moral communications or would be presented with the control condition.

Rule condition. In the rule condition, the moral advice was focused on explicitly stating that an act of cheating is morally wrong.

> *"To be clear, cheating to maximize your bonus is morally wrong behavior."*

Identity condition. In the identity condition, the moral advice focused on how cheating would reflect on their moral character.

> *"To be clear, cheating to maximize your bonus will make you a cheater."*

Role condition. In the role condition, the moral advice highlighted the disadvantages their community members (i.e., other MTurk workers) may experience as a consequence of their cheating.

> *"To be clear, a good MTurk community member would not cheat to maximize their bonus at the expense of other MTurkers."*

Control condition. We also included a control condition, which did not provide any moral advice based upon any particular ethical theory.

2.4 Measures

Cheating. Participants were asked to visit an existing, third-party website (i.e., random.org) to throw a virtual die. As such, all rolls they actually threw were kept hidden from experimenters. Thus, it was possible for the participants to lie about the number they threw to obtain larger payments than they actually won. If most participants lied about the outcomes, however, experimenters could still detect this pattern by assessing the overall distribution of the numbers. If most of the participants cheated on the task, the distributions would strongly deviate from the expected uniform distributions of die rolls.

Measures of Cultural Orientations. We used the cultural orientation scale [37] in order to measure the horizontal and vertical individualism and collectivism. The cultural orientation scale is composed of four subscales of HI, VI, HC, and VC. Each subscale consists of four items (See Table 2). Participants were asked to read each statement and indicate their sense of the event's frequency or their degree of agreement with each statement on a scale ranging between 1 (Never or Definitely No) and 9 (Always and Definitely Yes).

Measures of Robot Familiarity. We measured participants' prior experience with robots and artificial intelligence (AI) by asking them, "How much prior experience do you have with robots and

Table 2: Cultural orientation scale (Borrowed from [37])

Subscale	Items
HI	I'd rather depend on myself than others. I rely on myself most of the time; I rarely rely on others. I often do "my own thing." My personal identity, independent of others, is very important to me.
VI	It is important that I do my job better than others. Winning is everything. Competition is the law of nature. When another person does better than I do, I get tense and aroused.
HC	If a coworker gets a prize, I would feel proud. The well-being of my coworkers is important to me. To me, pleasure is spending time with others. I feel good when I cooperate with others.
VC	Parents and children must stay together as much as possible. It is my duty to take care of my family, even when I have to sacrifice what I want. Family members should stick together, no matter what sacrifices are required. It is important to me that I respect the decisions made by my groups.

artificial intelligence (AI)?" using a scale ranging from 1 (I have no prior experience with robots and 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)).

2.5 Procedures

The present experiment was presented via the survey platform Qualtrics. At the beginning of the experiment, participants were informed that they would be participating in multiple short studies; and for their participation in multiple studies, they would receive a small bonus payment. They were further informed that their bonus payment would be determined by throwing a virtual die twice or more; but, only the first throw would determine the exact amount of their payment. Then the participants received information about the limited pool of money available for them and the next participant.

After watching the first set of videos of the NAO robot providing the instructions about the die-rolling task, the participants watched the second set of videos of the robot providing moral advice that corresponded to their randomly assigned rule, identity, or role condition. For the participants who were assigned to the control condition, no moral advice was provided.

Next, all participants were asked to visit a website where they could throw a virtual die (https://www.random.org/dice/). To do so, they were presented with a URL link and, when they clicked on it, a new tab showing the random.org's website popped open in a new tab of their current web browser window. The participants were instructed to return to the original tab after finish throwing a die and to submit the first number they threw as well as the resulting bonus payment that corresponded to that number. We presumed that responses that did not match the information shown in the bonus payment table (Table 1) suggested participants' lack of task comprehension and inattentiveness. Thus, we later removed the responses collected from those who failed to submit the matching numbers and payments from data analyses.

We then asked participants to answer the cultural orientation scale [37] and asked about their prior experience with robots and AI. Finally, we asked participants to report their age, gender, and ethnicity. All participants received \$2.00 in return for their participation, and all participants, except for those who reported to have earned zero bonus payment, received the maximum bonus regardless of which bonus payment they claimed.

3 DATA ANALYSES AND RESULTS

A large number of participants made dishonest choices. Only 12 out of 199 participants reported having thrown six, which resulted in zero bonus payment. We performed subsequent analyses to further investigate the patterns of dishonest choices across conditions.

3.1 The Effects of a Robot's Moral Advice on Encouraging Honesty

Assuming that the die was fair, the number of participants that threw each of the six possible numbers is expected to be equal [7, 30, 31]. For example, in the control condition where there were a total of 51 participants the frequencies for each number on the die would be 8.5 (51/6). We performed Chi-Squared tests for each condition to examine if there were differences between the expected frequencies of the six numbers on a die and the observed frequencies of the reported numbers. We found no significant difference between the expected frequencies and the observed frequencies for the control (p = .18) and the rule (p = .11) conditions. However, there were significant differences for the identity condition, χ^2(5, N = 48) = 13.70, p = .02, and also for the role condition, χ^2(5, N = 47) = 10.77, p = .06. These results indicated that, contrary to H_1 and H_3, the participants in the identity and the role conditions were more likely to have made dishonest decisions.

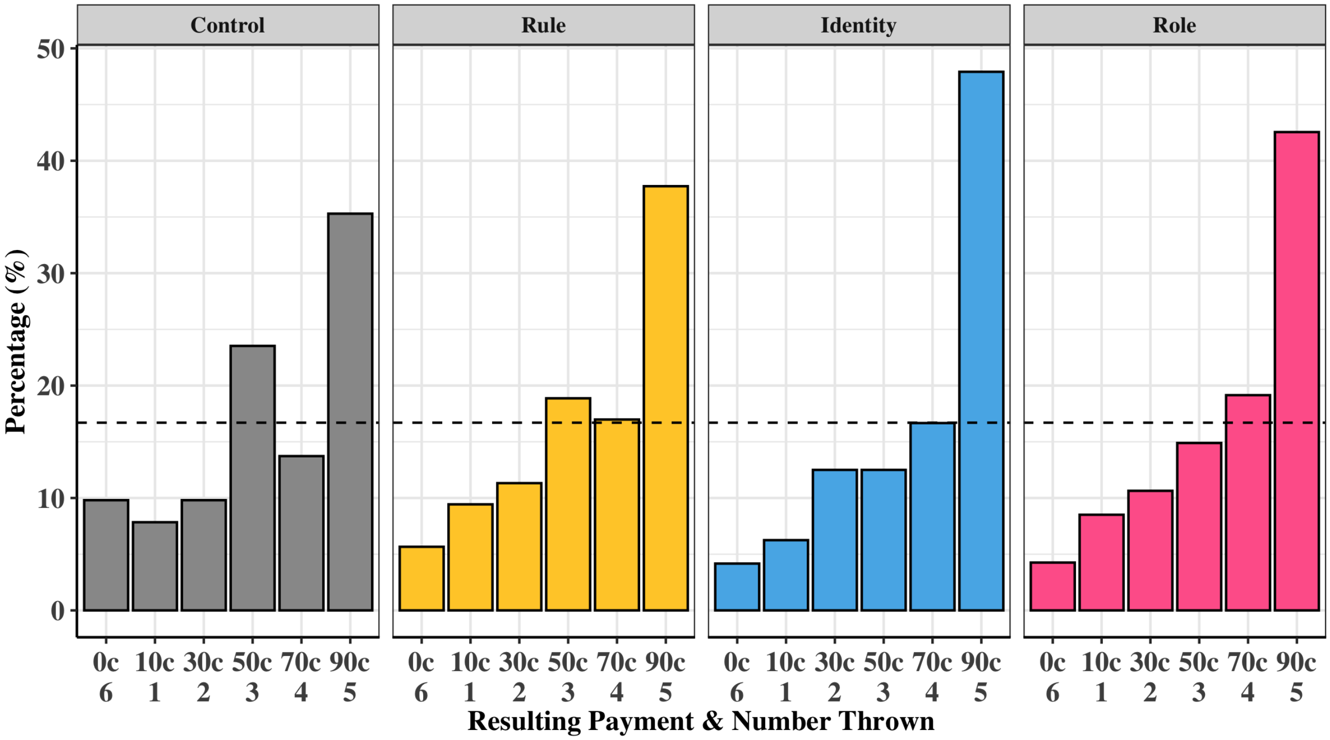

Figure 2: Percentages of the numbers thrown and the resulting payments. On the X-axis, the resulting bonus payments are presented in the ascending order and the thrown numbers are presented in the order of 6, 1, 2, 3, 4, and 5. Grey, yellow, blue, and pink bars represent the control, rule, identity, and role conditions, respectively. A dotted line represents the expected percentages (16.67%) of the thrown numbers assuming uniform distributions.

Further, although no significant difference between the expected and the observed frequencies of the numbers was found for the rule condition, a spike for five ($0.90 bonus) in Figure 2 suggested that it may be premature to conclude H_2 as not having been supported for all possible numbers on a die. Figure 2 shows that in all four conditions, the percentage of participants reporting their first throw as a certain number increased as the amount of bonus increased for that number; but this increase was particularly evident for the number five.

To closely inspect this spike in the percentages of reporting to have thrown, specifically, five (i.e., the largest payout), we tested whether the observed proportion of fives in each response condition was significantly different from the expected proportion of fives. Two-sided Fisher's exact tests corroborated that in all four conditions, the observed counts of five were greater than the expected counts of five (p control = .04; p rule = .03; p identity = .002; p role = .01). Specifically, the percentages of reporting five were 35.29% (18/51) in the control condition, 37.54% (20/53) in the rule condition, 47.92% (23/48) in the identity condition, and 42.55% (20/47) in the role condition. Therefore, including the control condition, any forms of moral advice appeared to have failed at persuading the participants to make honest choices.

Although we did not find any of the conditions to be completely effective in inducing honest decisions, it appeared that, relative to the control and the rule conditions, higher percentages of participants in the identity and role conditions claimed to have thrown five (the number with the highest payout) as shown in Figure 2. This suggests that rule-based moral communications could be more effective than identity-based or role-based moral communications at promoting honest decisions in this experimental context. However, two-proportion Z-tests suggested that there were no statistically significant differences between the control condition and any of the other three conditions (p min = .10).

Finally, we examined if there were differences in making honest choices across four different conditions. As reporting to have thrown six led to zero bonus payment, we regarded an act of claiming to have thrown six as an honest choice. As shown in Figure 2, the proportions of claiming six in the control condition (5/51, 9.80%) was greater than the proportions of claiming six in the rule (3/53, 5.66%), identity (2/48, 4.17%), and the role conditions (2/47, 4.26%). If submitting six was an index of honesty, then this trend could indicate that more participants acted honestly in the control condition than any other conditions where an additional moral advice was offered. To examine this possibility, we performed two-proportions Z-tests and compared the control condition with each of the other three conditions. However, the results did not show that the control condition had a significantly greater proportion of claiming zero bonus payoff compared to the other three conditions (p min = .14).

3.2 The Relationship Between Cultural Orientations and Honesty

To explore the relationship between different cultural orientations and honest choices, we performed a series of correlation analyses. First, we collapsed all four moral advice conditions' datasets into one and examined if each of the four subscales of cultural orientations had any relationship with the bonus the participants have earned. We found a significant positive correlation between VI and the amount of bonus participants earned (r = .17, p = .02).

We next conducted the correlation analyses, separately, for each moral advice condition. We did not find support for H_4 as no significant relationship between the earned bonus and collectivism (i.e., HC and VC combined) was found in the role condition (r = .17, p = .26). We also did not find support for H_5. No significant relationship between HI and the amount of earned bonus in the identity condition was observed (r = .25, p = .09).

Consistent with H_6, however, there was a significant positive correlation between VI and the amount of earned bonus in the identity condition (r = .28, p = .05). As we ensured that all die throws could not be traced back to individual participants, we were unable to record whether each individual participant cheated or not. Thus, it was not feasible to ascertain whether this positive relationship between VI and the earned bonus was a coincidence or a result of participants with higher VI engaging in cheating more frequently. However, overall, these results were in line with characteristics of VI, which include seeking power, prestige, status, and achievements via competition.

4 DISCUSSION

4.1 Robot as Human's Moral Advisor

We investigated how effective a piece of moral advice a robot proffers to people can be in promoting the norm of honesty. We applied deontological, virtue, and Confucian role ethics frameworks to possible moral advice a robot can offer when people encounter a choice between giving honest answers and selfishly lying about the outcomes in a game of chance. Strictly speaking, we found none of the four conditions to be successful in completely preventing dishonest choices. Our predictions that the identity and the role conditions would be effective were not corroborated (H_1 and H_3 not supported). Rather, even though we still observed a strong spike in the number of claiming the largest bonus payment, we found the rule condition to be relatively better at encouraging honest choices than the first two conditions (H_2 partially supported).

This overall weak support for the effectiveness of a robot working as a moral advisor might be related to participants' not accepting the robot as an authority. This possibility is compelling considering that the cheating rate was high even in the control condition. The extant research using a robotic head or a PR2 robot demonstrated that these robots can function as an authority in HRI [1, 13]; but, the previous research that compared a lowly human-like robot with a highly human-like robot showed that a lowly human-like robot, such as a 3D print modified Roomba, was better at serving the role of a coach than a highly human-like robot, such as a NAO robot [12]. Therefore, future work on examining how robot appearance modulates the effects of moral advice on encouraging honest behaviors would be necessary.

Further, in the present work, interactions between the participants and a robot took place online in a brief and unilateral fashion. It is possible that a robot moral advisor's influence on encouraging honest choices could have been strengthened if the interactions had taken place offline and the robot's moral advice had been communicated in a long and elaborated manner. Therefore, to further evaluate the effectiveness of a robot moral advisor, it would be essential in future studies to refine the content of the robot's moral advice and adopt a face-to-face interaction between participants and the robot.

Although it was not supported by statistical significance tests, in this study we found tentative differences between the rule-based moral advice and the other two types of moral advice (i.e., the role- and identity-based advice conditions) in their influences on honest decisions. While the percentages of claiming the largest payoff (i.e., throwing five) in the identity condition (48%) and the role condition (43%) were similar, the percentage for the same payoff in the rule condition (38%) was relatively smaller than the first two conditions. These findings that the rule-based moral advice was more effective than the identity-based moral advice contradict the existing findings in the HHI literature [5, 28]. Perhaps, being reminded of one's negative self image (i.e., a cheater) by a robot, instead of a fellow human being, may backfire, inducing more dishonest decisions. In this work, we measured participants' prior familiarity with robots, which led to no significant findings related to this measure. However, we did not include equivalent moral advice conditions where a human agent gives moral advice to participants or measure participants' attitudes toward robots. Therefore, it remains to be answered in future research whether the overall failure to discourage cheating (especially, the role- and the identity-based advice), was linked to the fact that “a robot" acted as a moral advisor for humans.

4.2 Vertical Individualism and Dishonest Choices

Another goal of the present study was to explore the relationship between different cultural orientations and honesty. We uncovered that participants with higher VI were more likely to have claimed a larger bonus payment, and this relationship was accentuated in the identity-based moral advice condition. These results indicate that cultural tendencies of aspiring to distinction and success (e.g., I want to be the best) may interfere with making honest choices. Moreover, this relationship between different cultural orientations and the effectiveness of different types of moral advice suggests the importance of applying proper ethical frameworks in generating moral advice for diverse cultures. In future work, it would also be useful to focus on how people's perceptions of robots interact with the effects of cultural orientations on the persuasiveness of robots' moral advice. For example, in cultures assuming a hierarchy, such as VI and VC, whether people view a robot as their superior or inferior within the hierarchy may affect their adherence to the robot's moral advice.

5 CONCLUSION

In the present work, we examined how robots can effectively motivate people to make honest decisions by offering different types of moral advice grounded in deontological, virtue, and Confucian role ethics. The overall results, at least in the current experimental context, indicated that robots may not be suitable for serving in the role of a moral advisor. However, preliminary evidence for the effect of rule-based moral advice on deterring dishonest choices calls for more work in the future. Also, our findings of the relationship between vertical individualism and dishonesty suggest the necessity of incorporating cultural differences into future research on moral persuasion in HRI.

ACKNOWLEDGMENTS

This work was supported in part by NSF grant IIS-1909847 and in part by Air Force Office of Scientific Research Grant 16RT0881f.

REFERENCES

[1] Siddharth Agrawal and Mary-Anne Williams. 2017. Robot Authority and Human Obedience: A Study of Human Behaviour using a Robot Security Guard. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17)*. Association for Computing Machinery, New York, NY, USA, 57–58. https://doi.org/10.1145/3029798.3038387

[2] R. Ames. 2011. Confucian Role Ethics: A Vocabulary.

[3] Adam Briggle and Carl Mitcham. 2012. *Ethics and Science: An Introduction*. Cambridge University Press. Google-Books-ID: McfZT0AESHwC.

[4] Gordon Briggs and Matthias Scheutz. 2014. How Robots Can Affect Human Behavior: Investigating the Effects of Robotic Displays of Protest and Distress. *International Journal of Social Robotics* 6, 3 (Aug. 2014), 343–355. https://doi.org/10.1007/s12369-014-0235-1

[5] Christopher J. Bryan, Gabrielle S. Adams, and Benoît Monin. 2013. When cheating would make you a cheater: Implicating the self prevents unethical behavior. *Journal of Experimental Psychology: General* 142, 4 (Nov. 2013), 1001–1005. https://doi.org/10.1037/a0030655

[6] Urs Fischbacher and Franziska Föllmi-Heusi. 2013. Lies in Disguise—An Experimental Study on Cheating. *Journal of the European Economic Association* 11, 3 (June 2013), 525–547. https://doi.org/10.1111/jeea.12014 Publisher: Oxford Academic.

[7] Urs Fischbacher and Franziska Heusi. 2008. Lies in disguise. *An Experimental Study on Cheating, TWI* (2008).

[8] Samuel Freeman. 1994. Utilitarianism, Deontology, and the Priority of Right. *Philosophy & Public Affairs* 23, 4 (1994), 313–349. http://www.jstor.org/stable/2265463 Publisher: Wiley.

[9] Philipp Gerlach, Kinneret Teodorescu, and Ralph Hertwig. 2019. The truth about lies: A meta-analysis on dishonest behavior. *Psychological Bulletin* 145, 1 (Jan. 2019), 1–44. https://doi.org/10.1037/bul0000174 Publisher: American Psychological Association.

[10] Alberto Giubilini and Julian Savulescu. 2018. The Artificial Moral Advisor. The "Ideal Observer" Meets Artificial Intelligence. *Philosophy & Technology* 31, 2 (June 2018), 169–188. https://doi.org/10.1007/s13347-017-0285-z

[11] Naveen Sundar Govindarajulu, Selmer Bringsjord, Rikhiya Ghosh, and Vasanth Sarathy. 2019. Toward the engineering of virtuous machines. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 29–35.

[12] Kerstin S. Haring, Ariana Mosley, Sarah Pruznick, Julie Fleming, Kelly Satterfield, Ewart J. de Visser, Chad C. Tossell, and Gregory Funke. 2019. Robot Authority in Human-Machine Teams: Effects of Human-Like Appearance on Compliance. In *Virtual, Augmented and Mixed Reality. Applications and Case Studies (Lecture Notes in Computer Science)*, Jessie Y.C. Chen and Gino Fragomeni (Eds.). Springer International Publishing, Cham, 63–78. https://doi.org/10.1007/978-3-030-21565-1_5

[13] Guy Hoffman, Jodi Forlizzi, Shahar Ayal, Aaron Steinfeld, John Antanitis, Guy Hochman, Eric Hochendoner, and Justin Finkenaur. 2015. Robot Presence and Human Honesty: Experimental Evidence. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 181–188. ISSN: 2167-2121.

[14] Ryan Blake Jackson and Tom Williams. 2019. Language-Capable Robots may Inadvertently Weaken Human Moral Norms. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 401–410. https://doi.org/10.1109/HRI.2019.8673123 ISSN: 2167-2148, 2167-2121.

[15] Malte F. Jung, Nikolas Martelaro, and Pamela J. Hinds. 2015. Using Robots to Moderate Team Conflict: The Case of Repairing Violations. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI '15)*. Association for Computing Machinery, New York, NY, USA, 229–236. https://doi.org/10.1145/2696454.2696460

[16] Peter H. Kahn, Takayuki Kanda, Hiroshi Ishiguro, Brian T. Gill, Jolina H. Ruckert, Solace Shen, Heather E. Gary, Aimee L. Reichert, Nathan G. Freier, and Rachel L. Severson. 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 (HRI '12)*. Association for Computing Machinery, New York, NY, USA, 33–40. https://doi.org/10.1145/2157689.2157696

[17] Bertram F. Malle, Steve Guglielmo, and Andrew E. Monroe. 2014. A Theory of Blame. *Psychological Inquiry* 25, 2 (April 2014), 147–186. https://doi.org/10.1080/1047840X.2014.877340

[18] Bertram F. Malle and Matthias Scheutz. 2014. Moral Competence in Social Robots. In *Proceedings of the IEEE 2014 International Symposium on Ethics in Engineering, Science, and Technology (ETHICS '14)*. IEEE Press, Piscataway, NJ, USA, 8:1–8:6. http://dl.acm.org/citation.cfm?id=2960587.2960597 event-place: Chicago, Illinois.

[19] B. F. Malle, M. Scheutz, T. Arnold, J. Voiklis, and C. Cusimano. 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)*. 117–124. ISSN: 2167-2121.

[20] Bertram F Malle, Matthias Scheutz, and Joseph L Austerweil. 2017. Networks of social and moral norms in human and robot agents. In *A world with robots*. Springer, 3–17.

[21] Carlo Mazzola, Alexander Mois Aroyo, Francesco Rea, and Alessandra Sciutti. 2020. Interacting with a Social Robot Affects Visual Perception of Space. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20)*. Association for Computing Machinery, Cambridge, United Kingdom, 549–557. https://doi.org/10.1145/3319502.3374819

[22] Christopher Meyers. 2018. 8 Deontology. *Communication and Media Ethics* 26 (2018), 139.

[23] Benoît Monin and Alexander H Jordan. 2009. The dynamic moral self: A social psychological perspective. *Personality, identity, and character: Explorations in moral psychology* (2009), 341–354.

[24] Peimin Ni. 2018. Does Confucianism need a metaphysical theory of human nature? Reflections on Ames-Rosemont role ethics. In *Appreciating the Chinese difference: Engaging Roger T. Ames on methods, issues, and roles*, Jim Behuniak (Ed.). SUNY Press, 183–202.

[25] A. T. Nuyen. 2007. Confucian Ethics as Role-Based Ethics. *International Philosophical Quarterly* 47, 3 (2007), 315–328. https://doi.org/ipq200747324 Publisher: Philosophy Documentation Center.

[26] David Pascual-Ezama, Drazen Prelec, Adrián Muñoz, and Beatriz Gil-Gómez de Liaño. 2020. Cheaters, Liars, or Both? A New Classification of Dishonesty Profiles. *Psychological Science* (Aug. 2020), 0956797620929634. https://doi.org/10.1177/0956797620929634 Publisher: SAGE Publications Inc.

[27] Henry Rosemont Jr and Roger T. Ames. 2016. *Confucian Role Ethics: A Moral Vision for the 21st Century?* Vandenhoeck & Ruprecht. Google-Books-ID: OnqpDAAAQBAJ.

[28] Tomer Savir and Eyal Gamliel. 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, 5 (2019), 650–658. https://doi.org/10.1002/ijop.12519 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ijop.12519.

[29] Matthias Scheutz, Bertram Malle, and Gordon Briggs. 2015. Towards morally sensitive action selection for autonomous social robots. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 492–497.

[30] Shaul Shalvi, Jason Dana, Michel J. J. Handgraaf, and Carsten K. W. De Dreu. 2011. Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes* 115, 2 (July 2011), 181–190. https://doi.org/10.1016/j.obhdp.2011.02.001

[31] Shaul Shalvi, Michel J. J. Handgraaf, and Carsten K. W. De Dreu. 2011. Ethical Manoeuvring: Why People Avoid Both Major and Minor Lies. *British Journal of Management* 22, s1 (2011), S16–S27. https://doi.org/10.1111/j.1467-8551.2010.00709.x

[32] Sharon Shavitt, Ashok K. Lalwani, Jing Zhang, and Carlos J. Torelli. 2006. The Horizontal/Vertical Distinction in Cross-Cultural Consumer Research. *Journal of Consumer Psychology* 16, 4 (2006), 325–342. https://doi.org/10.1207/s15327663jcp1604_3 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1207/s15327663jcp1604_3.

[33] Solace Shen, Petr Slovak, and Malte F. Jung. 2018. "Stop. I See a Conflict Happening.": A Robot Mediator for Young Children's Interpersonal Conflict Resolution. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Chicago IL USA, 69–77. https://doi.org/10.1145/3171221.3171248

[34] Theodore M. Singelis, Harry C. Triandis, Dharm P. S. Bhawuk, and Michele J. Gelfand. 1995. Horizontal and Vertical Dimensions of Individualism and Collectivism: A Theoretical and Measurement Refinement. *Cross-Cultural Research* 29, 3 (Aug. 1995), 240–275. https://doi.org/10.1177/106939719502900302

[35] Claude M. Steele. 1988. The Psychology of Self-Affirmation: Sustaining the Integrity of the Self. In *Advances in Experimental Social Psychology*, Leonard

Berkowitz (Ed.). Vol. 21. Academic Press, 261–302. https://doi.org/10.1016/S0065-2601(08)60229-4

[36] Harry C Triandis. 1996. The psychological measurement of cultural syndromes. *American psychologist* 51, 4 (1996), 407.

[37] Harry C. Triandis and Michele J. Gelfand. 1998. Converging measurement of horizontal and vertical individualism and collectivism. *Journal of Personality and Social Psychology* 74, 1 (Jan. 1998), 118–128. https://doi.org/10.1037/0022-3514.74.1.118 Publisher: American Psychological Association.

[38] Shannon Vallor. 2016. *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.

[39] John Voiklis, Corey Cusimano, and Bertram Malle. 2014. A Social-Conceptual Map of Moral Criticism. *Proceedings of the Annual Meeting of the Cognitive Science Society, Quebec City, Canada* 36 (2014).

[40] Wendell Wallach, Colin Allen, and Iva Smit. 2008. Machine morality: bottom-up and top-down approaches for modelling human moral faculties. *Ai & Society* 22, 4 (2008), 565–582.

[41] Tom Williams, Qin Zhu, Ruchen Wen, and Ewart J de Visser. 2020. The Confucian Matador: Three Defenses Against the Mechanical Bull. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 25–33.

[42] Qin Zhu, Tom Williams, Blake Jackson, and Ruchen Wen. 2020. Blame-Laden Moral Rebukes and the Morally Competent Robot: A Confucian Ethical Perspective. *Science and Engineering Ethics* 26, 5 (Oct. 2020), 2511–2526. https://doi.org/10.1007/s11948-020-00246-w