Comparing Norm-Based and Role-Based Strategies for Robot Communication of Role-Grounded Moral Norms

RUCHEN WEN, Colorado School of Mines, USA
BOYOUNG KIM, George Mason University, USA
ELIZABETH PHILLIPS, George Mason University, USA
QIN ZHU, Colorado School of Mines, USA
TOM WILLIAMS, Colorado School of Mines, USA

Because robots are perceived as moral agents, they must behave in accordance with human systems of morality. This responsibility is especially acute for language-capable robots because moral communication is a method for building moral ecosystems. Language capable robots must not only make sure that what they say adheres to moral norms; they must also actively engage in moral communication to regulate and encourage human compliance with those norms. In this work, we describe four experiments (total $N = 316$) across which we systematically evaluate two different moral communication strategies that robots could use to influence human behavior: a norm-based strategy grounded in deontological ethics, and a role-based strategy grounded in role ethics. Specifically, we assess the effectiveness of robots that use these two strategies to encourage human compliance with norms grounded in expectations of behavior associated with certain social roles. Our results suggest two major findings, demonstrating the importance of moral reflection and moral practice for effective moral communication: First, opportunities for reflection on ethical principles may increase the efficacy of robots' role-based moral language; and second, following robots' moral language with opportunities for moral practice may facilitate role-based moral cultivation.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → *Discourse, dialogue and pragmatics*; • **Computer systems organization** → *Robotics.*

Additional Key Words and Phrases: human-robot interaction, role ethics, moral communication

ACM Reference Format:
Ruchen Wen, Boyoung Kim, Elizabeth Phillips, Qin Zhu, and Tom Williams. 2018. Comparing Norm-Based and Role-Based Strategies for Robot Communication of Role-Grounded Moral Norms. 1, 1 (October 2018), 26 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Research has shown that people not only perceive robots as social actors [47], but also as moral agents [19]. Accordingly, people expect and demand that robots behave ethically and tend to extend moral judgments and blame to their robot if they do not [31, 45]. Thus, it is crucial for robots to behave in accordance with human systems of morality. To enable moral competence in social robots, Malle and Scheutz proposed the need for four components: [43, 44]:

(1) a moral core, i.e. a system of moral norms; and the ability to use those norms for:

Authors' addresses: Ruchen Wen, Colorado School of Mines, Golden, CO, USA, rwen@mines.edu; Boyoung Kim, George Mason University, Fairfax, VA, USA, bkim55@gmu.edu; Elizabeth Phillips, George Mason University, Fairfax, VA, USA, ephill3@gmu.edu; Qin Zhu, Colorado School of Mines, Golden, CO, USA, qzhu@mines.edu; Tom Williams, Colorado School of Mines, Golden, CO, USA, twilliams@mines.edu.

(2) moral cognition (to generate emotional responses to norm violations and make moral judgements),
(3) moral decision making and action (to conform their own actions to the norm system), and
(4) moral communication (to generate morally sensitive language and to explain their actions).

These components are critical not only for behaving morally and justifying one's own behavior, but also for understanding and responding appropriately to the moral behavior of others, and thus to regulate others' behavior [44]. This is critical because robots have been shown to hold significant persuasive power over humans in previous HRI research [12, 33], capable of influencing, persuading, and coercing humans in a variety of ways [8, 9, 12, 15, 17, 23, 33, 46, 49–51, 55, 59, 60, 67].

Moreover, there is evidence that robots can not only influence interactants' locally contextualized behaviors, but moreover can indeed exert influence over their interactants' compliance with social norms [39, 60] and moral norms [25, 26], presenting the potential not only to influence humans' long-term social and moral behaviors, but to also influence what social and moral behaviors humans choose to condone or sanction in others. The results of which could lead to potential "ripple effects" of moral behavior across robots' social and moral ecosystems.

This potential for large-scale moral influence presents roboticists with new moral responsibilities. Specifically, because robots have this persuasive power, and because moral communication is a crucial part of building a harmonious moral ecosystem in human-robot interaction, roboticists now have a *moral obligation* to ensure that robots use their communication strategies to avoid accidentally condoning immoral behavior, and to speak out against immoral behavior once detected. By doing so we can proactively use robotic persuasion for good by shaping humans toward morally good ends (e.g., [13, 30, 57]). Indeed, robots' potential for influence presents roboticists with moral opportunities, not only to maintain the health of their moral ecosystem, but also to provide opportunities for human teammates to engage in moral self-cultivation [70].

In this work, we evaluate two different robotic moral communication strategies for encouraging human compliance: a norm-based strategy grounded in deontological ethics, and a role-based strategy grounded in role ethics. We then test the effectiveness of these two strategies in encouraging compliance with norms grounded in role expectations. Specifically, we used moral language that either highlighted the norm-related or role-related tenets of these moral principles. We were interested in the effects of robots' use of such moral language on both moral behavior and moral beliefs that motivate such behavior.

Our framework for analyzing moral behavior is grounded in the Theory of Planned Behavior (TPB) [1], which has been extensively used in previous studies to measure norm strength and intent to comply with norms [24, 32, 35]. The TPB claims that a considerable proportion of variance in behavior can be accounted for on the basis of (1) intentions to perform behaviors, (2) attitudes towards those behaviors, (3) perceptions of the subjective norms regarding those behaviors, and (4) one's perceived behavioral control over those behaviors [1].

The TPB framework provided us not only a means to test the effectiveness of moral communication strategies on behavioral compliance and intention for behavioral compliance, but also served as a reflective exercise which elucidated the conditions in which different types of moral language could be most impactful. And this reflective exercise may have increased the efficacy of a robot's role-based moral communication strategy.

In the following sections, we will present an overview of role-based and normative moral communication (Section 2), introduce how we expect robots' use of two different moral communication strategies to differentially effect human behavior (Section 3), and present and discuss the implications of our results and directions for future work (Section 8 and 10).

2 BACKGROUND

There are a number of techniques that robots can use to proactively shape their social and moral ecologies, including both nonverbal behaviors [61] and verbal behaviors [26]. In this work, we investigate robots' use of explicit verbal communication of moral guidance to exert overt moral influence. Verbal communication of moral principles is made challenging by the wide variety of ways that moral guidance can be delivered verbally, and the significant impact that subtle differences in phrasing can make in these efforts. For example, the face threat and blame ascription of moral language must be carefully tailored in order to be effective, as under-harsh language will not be taken seriously, and over-harsh language will be viewed as impolite and likely promote backlash [14, 26, 27, 70]. Similarly, robot moral language can be subtly varied through grounding in different moral frameworks, which may yield different outcomes in terms of both moral influence and in terms of how people perceive the robots attempt to exert that influence [66].

The majority of previous approaches for enabling morally capable robots have been based on deontological principles [7] in which the morality of an action depends solely on its consistency with well-specified moral norms [22]. However, it is well known that norm-based ethical theories (e.g., deontology) have philosophical and computational limitations. Vallor [62] writes, these frameworks often struggle to "accommodate the constant flux, contextual variety, and increasingly opaque horizon of emerging technologies." [62]

Technology ethicists have thus been exploring underrepresented ethical traditions, such as role-based and relational ethical theories, for new perspectives on morally grounded robotics and automation. For instance, Coeckelbergh [16] discusses the need to focus on moral considerations in human-robot relations rather than on the moral status of humans and robots alone, and offers an alternative, social-relational approach to moral consideration, which re-frames the issue by shifting the focus from individual ontology to social-relational ontology of moral principles for human-robot interaction. [16] Thus, in contrast to norm-centering deontological approaches for robot moral communication, we are also interested in communication strategies grounded in these sorts of social-relational perspectives, such as those found in Confucian ethics, relational ethics [36] and early Stoic works [56], which would all suggest centering communication on the role(s) assigned to robots (and humans) [42].

Among these disparate social-relational perspectives, we are particularly interested in *Confucian role ethics*. In this paradigm, moral rules and virtues are derived from the social roles humans assume, and social roles in turn are determined by the social relationships humans have with others [48]. Ethicists have previously argued that deontological ethics may have difficulty anticipating the roles filled by emerging technologies (like robots), and that roboticists should re-focus on human-robot relationships. We believe that Confucian role ethics may be a suitable way to addresses these concerns.

To illustrate, Confucian role ethics advocates for a kind of *relational ontology*, in which an agent never cultivates virtues solely by herself, and instead becomes virtuous while actively living her social roles through everyday interactions with others [5]. The nature of a particular role relationship often evokes feelings and expectations characteristic of that relationship [6]. Therefore, from the Confucian perspective, a major criterion for technology assessment is whether practices generated by a particular technology such as robotics are conducive or detrimental to our performance of social roles [10]. A truly socially integrated robot has a moral obligation to help humans achieve the project of living social roles appropriately and cultivating the moral self. Viewed another way, in comparison to traditional norm-centering approaches, which emphasize epistemological forms of moral action (e.g., what is good or bad), role-based approaches such as Confucian Ethics also emphasize ontological forms of moral learning (e.g., how to become good) [5, 11, 52, 53].

To be clear, however, even in role-based moral frameworks, norms are still integral to understanding morality, due to the role they objectively play in human moral cognition. And indeed, from the Confucian role ethics perspective, norms and roles are thus integrally related. Moral rules or rituals accepted in a communal context determine how an agent should act in specific situations, and the cultivation of role-based moralities must be based on the agent's diligent practice of moral rituals [69]. In other words, what is viewed as appropriate for a given role is grounded in norms, and the norms that people are expected to follow are conditioned on their roles. But although norms and roles are closely interconnected, they are nevertheless distinct concepts, and moral language can differentially emphasize norms versus roles even for norms with clear grounding in social roles.

To understand the differences in moral language that would be used under these different frameworks, it may be helpful to consider a few examples. For example, consider a context in which a speaker makes an immoral request to perform an action that could be construed as involving theft. In this case, a listener might respond in different ways grounded in different moral frameworks. A norm-based response might be"*I cannot help you cheat because cheating is wrong,*" while a role-based response might be "*I cannot help you cheat because I am your classmate and a good classmate should not do that.*" Critically, we might expect these approaches to be interpreted in fundamentally different ways. While the norm-based response directly refers to the norm violation (i.e., "*cheating is wrong*"), the role-based response only highlights the role of the speaker (i.e., "*classmate*").

One hypothesis that has been suggested in previous literature [66] is that these two strategies might provoke very different responses for the speaker whose request or command is being rejected. For example, it could be the case that the norm-based response would provoke an immediate emotional reaction due to the threat of sanction associated with norm violations. On the other hand, the lack of direct reference to the norm in the role-based response might require the listener to reflect on the content of the response in order to determine why the request is being rejected, including reflection on their knowledge of the expectations and responsibilities associated with the highlighted role. This is a type of reflection that is centered and valued within the Confucian Role Ethics tradition [68]. If this were the case, it is possible that the role-based approach could be more effective in encouraging positive long-term benefits.

Critically, human moral development is not simply the alignment of behavioral conduct and norms. Instead, what is at stake is whether the practice of norms can lead to instrumental value, that is, a better way of reflecting on our selves, living our communal roles, and cultivating the virtues indispensable to the fulfilment of social roles [4].

In this work, we aimed to test the effectiveness of robot moral language involving norms grounded in role expectations. Specifically, we investigated the differential impacts of robot moral language that highlights either norm-related or role-related tenets, on humans' systems of role-grounded moral norms.

3 METHOD

To achieve our research aims, we conducted a series of four human-subjects studies, modeled on a Solomon four experimental design [58], in which participants were asked to engage in robot-assisted crowdworking scenarios. We chose to use an online crowdworking scenario due in large part to the COVID-19 Pandemic, which prevented in-person experiments at the time of experimental design [18]. Using different moral communication strategies, the robot in each experiment encouraged participants to follow a role-grounded norm: that crowdworkers should strive to attentively engage in the tasks for which they were paid. This role-grounded norm was chosen not because it is the type of norm we foresee robots cultivating in the future, but because it is a norm valued by members of our target community (i.e., crowdworkers). Indeed, in other work we have performed in our lab, crowdworkers have stressed in free responses that they take their jobs seriously, and strive to positively contribute to scientific studies.

In general, we expected that role-based and norm-based interventions would be differentially effective at different time scales, with role-based moral interventions potentially having more long-term impact but norm-based moral interventions having more immediate impact. This suggested to us a variety of experiments would be needed across our research efforts, examining the effectiveness of moral interventions across different time scales. In the experiments presented in this paper, we begin this extended research effort by examining brief interactions with robots (i.e., to assess immediate impact) in which we would expect a norm-based moral intervention provided by a robot to be particularly effective. Our overall hypothesis was that moral interventions may be staged by robots to strengthen human teammates' role-grounded moral norms, and that within the timeframe of brief HRI contexts, norm-based moral interventions will be particularly effective. Specifically, we formulate five sub-hypotheses articulating how we expect this norm strengthening to manifest in observable ways.

3.1 Hypotheses

First, we expected norm-strengthening within our crowdworking context to manifest in two objective measures.

Hypothesis H1 Crowdworkers who received role-based moral communication interventions will perform their tasks more accurately than crowdworkers who received no intervention; crowdworkers who received norm-based moral communication interventions will also perform their task more accurately than crowdworkers who received no intervention; and crowdworkers who received norm-based moral communication interventions will have greater improvement on task performance than crowdworkers who received role-based moral communication interventions.

Hypothesis H2 Crowdworkers who received role-based moral communication interventions will spend more time on their assigned tasks than crowdworkers who received no intervention; crowdworkers who received norm-based moral communication interventions will also spend more time on their assigned tasks than crowdworkers who received no intervention; and crowdworkers who received norm-based moral communication interventions will spend more time on their assigned tasks than crowdworkers who received role-based moral interventions.

Second, we expected norm-strengthening within our crowdworking context to manifest in three ways assessable by subjective (self-reported) measures.

Hypothesis H3 Crowdworkers who received role-based moral communication interventions are more likely to report increases in positive attitudes towards attentive crowdworking behavior than crowdworkers who received no intervention; crowdworkers who received norm-based moral communication interventions are also more likely to report increases in positive attitudes towards attentive crowdworking behavior than crowdworkers who received no intervention; and crowdworkers who received norm-based moral communication interventions are more likely to report increases in positive attitudes towards attentive crowdworking behavior than crowdworkers who received role-based moral interventions.

Hypothesis H4 Crowdworkers who received role-based moral communication interventions are more likely to directly report stronger perceptions of subjective norm strength for attentive crowdworking behavior than crowdworkers who received no intervention; crowdworkers who received norm-based moral communication interventions are also more likely to directly report stronger perceptions of subjective norm strength for attentive crowdworking behavior than crowdworkers who received no intervention; and crowdworkers who received norm-based moral communication interventions are more likely to directly report stronger perceptions of subjective norm

strength for attentive crowdworking behavior than crowdworkers who received role-based moral communication interventions.

Hypothesis H5 Crowdworkers who received role-based moral communication interventions are more likely to express greater intentions to engage in attentive crowdworking behavior than crowdworkers who received no intervention, crowdworkers who received norm-based moral communication interventions are also more likely to express greater intentions to engage in attentive crowdworking behavior than crowdworkers who received no intervention, and crowdworkers who received norm-based moral communication interventions are more likely to express greater intentions to engage in attentive crowdworking behavior than crowdworkers who received role-based moral communication interventions.

3.2 Experimental design

To assess these five hypotheses, we conducted a set of four experiments using a mixed factorial design, in which all participants were asked to complete an experimental task twice (within), and in which participants were given one of three moral communication interventions between those two tasks: a norm-based intervention, a role-based intervention, and a control intervention (between). Participants were randomly assigned to one of the between-subjects conditions.

Using this design, the effects of moral communication interventions could be directly observed by assessing differences in performance between the pairs of tasks. But in addition to examining the effects of different moral interventions on behavior, we were also interested in their effects on inner states' such as the beliefs that predict behavior (e.g., attitudes towards behavior, perceived strength of norms associated with behavior, perceived control over completing behavior, and future intention to perform behavior), including changes to beliefs as a result of the norm-based or role-based interventions. To measure the effects of each type of moral intervention on these beliefs, a *Theory of Planned Behavior (TPB* questionnaire (described below) was designed and deployed. In using this measure, we needed to make decisions as to when to deliver this survey relative to the experimental tasks and experimental intervention.

When considering whether to design a pre-intervention-post-intervention strategy to measure changes in beliefs as a function of our interventions, it became apparent that asking participants to self-report their beliefs prior to the experimental interventions (as a pre-test) could bias them towards thinking about behaviors before engaging in the experimental tasks. Doing so could potentially influence how well participants perform their task even before any intervention. This could cause a situation where the measurement strategy itself serves as an intervention to behavior, and it would be difficult to decipher which (moral intervention or measurement) caused changes to behavior if at all.

To address this concern and investigate whether the placement of the TPB could bias participant behavior, especially for the first task, created four experiments with independent study procedures with respect to both TPB administration (either a pre-test/post-test design or a post-test only design), and with respect to TPB order (either task-after-intervention or survey-after-intervention), to counterbalance the order of presentation of the TPB questionnaire. We modeled our experimental design on a Solomon four group design [58]. The Solomon four group design was originally created to help address concerns over participants becoming sensitized to an experimental treatment by a pre-test before the intervention. Even though we could only use data from the experiments with the pre-intervention/post-intervention measurement to analyze the TPB scores, we still decided to include the post-test only design for consistency.

Fig. 1 provides a summary of the four different procedures used in our four experiments, which we used to counterbalance the order of placement of the TPB. Incremental Bayesian sampling plans [54] were used for these four experiments, resulting in slightly different numbers of participants being run in each experiment.

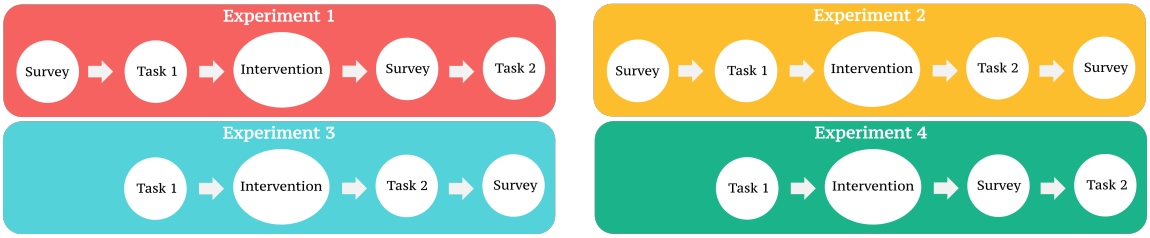

Fig. 1. Experimental procedures in our four experiments. Experiment 1 and 2 had both pre-intervention and post-intervention survey, Experiment 3 and 4 only had post-intervention survey.

3.3 Experimental task: Counting articles for a robot experimenter

Our four experiments were created and deployed on a custom website using the psiTurk crowdsourcing platform, which randomly batched participants into between-subjects Intervention conditions for each experiment. All procedures were approved by the Colorado School of Mines Institutional Review Board.

The four experiments conducted in this work used a shared experimental task in which participants were required to perform a close reading task in which participants were instructed to read two short passages of text and count the number of grammatical articles (i.e., "a," "an," and "the") contained in each, ostensibly to investigate the use of grammatical articles in text. Modeled after citizen science archiving tasks like those contained in The Smithsonian's Transcription Center (https://transcription.si.edu/), the two passages of text were taken from The Book of Trades published in 1847 by Edward W. Miller. The first passage was taken from a two-page description of a hatter's profession, while the second passage was taken from a a two-page description of a cooper's profession.

Before performing this task, our participants began by watching a video introduction (Fig. 2, in which A NAO robot (Softbank Robotics) was introduced to participants as the study Experimenter, responsible for guiding progression through the study. In the introductory video, the NAO provided introductory information about the study, and instructions for completing experimental tasks. All videos of NAO speaking used NAO's default 'voice' and were coupled with closed captioning located at the bottom center of each video (Fig. 2). All video stimuli can be found in our Open Science Framework Project, available at https://osf.io/6b5ng/?view_only=872834de133748ff8496eb00cc5a8b44. Below is a transcription of NAO's speech from the introductory video.

> "Hello there! Welcome! We are conducting research on examining people's use of articles. Articles are words like 'a,' 'an,' and 'the.' In this project, we plan to examine how often people use articles in various forms of writing. We would like you to help us by counting the number of articles in two passages of text. Later on in this experiment, you will need to complete two tasks. In each task, you will see a text. Please count the number of 'a,' 'an,' and 'the' as accurately as possible, and submit the total number of each of the articles in the box at the end of the text. Please click the 'keep going' button to start the experiment when you are ready."

This task described by the robot was conducted in the web interface shown in Fig. 3. As shown in that Figure, to help increase the realism of interacting with the robot while completing study tasks online, the website containing the study was constructed such that a video of the NAO persisted in the upper left corner as participants completed the article counting tasks. The video depicted NAO's occasional passive movement (e.g., looking around) while silent; no audio was coupled with this video of NAO. Task instructions were persistently displayed in text just below the video of

Fig. 2. Frame from the study's introduction video.

NAO, and on the right side of the page, participants were provided each passage of text as an image file, below which were three boxes where participants could submit their answers for each type of article. All participants completed this task twice, using the two different passages from The Book of Trades.

As previously described in the experimental design section, the periods between Briefing, Task 1, Task 2, and Debriefing, differed between our four experiments, with moral communication interventions and TPB surveys differentially deployed during those periods across our four experiments.

At the end of the experiment, participants answered an attention check question and lead to a page stating that they had reached the end of the study and thanking them for their participation. After completing the experiment, participants were provided with an MTurk payment code. Participants were paid at a rate of approximately $1 per expected 5-minutes in return for their participation.

3.4 Moral communication intervention conditions

Between article counting tasks, participants were provided with a video of the NAO robot providing one of two moral interventions or a control intervention. Descriptions of each intervention are given below.

Condition 1: Control Intervention:

In the Control Intervention condition, the robot guided participants through the experiment without giving an explicit moral intervention between article counting tasks. Instead, after completing the first article counting task, participants in this condition were presented with a video of NAO thanking them and instructing them to continue on to the next phase of the experiment:

> "Thank you! Please click the 'keep going' button to continue this experiment when you are ready."

Condition 2 - Norm-based Moral Intervention:

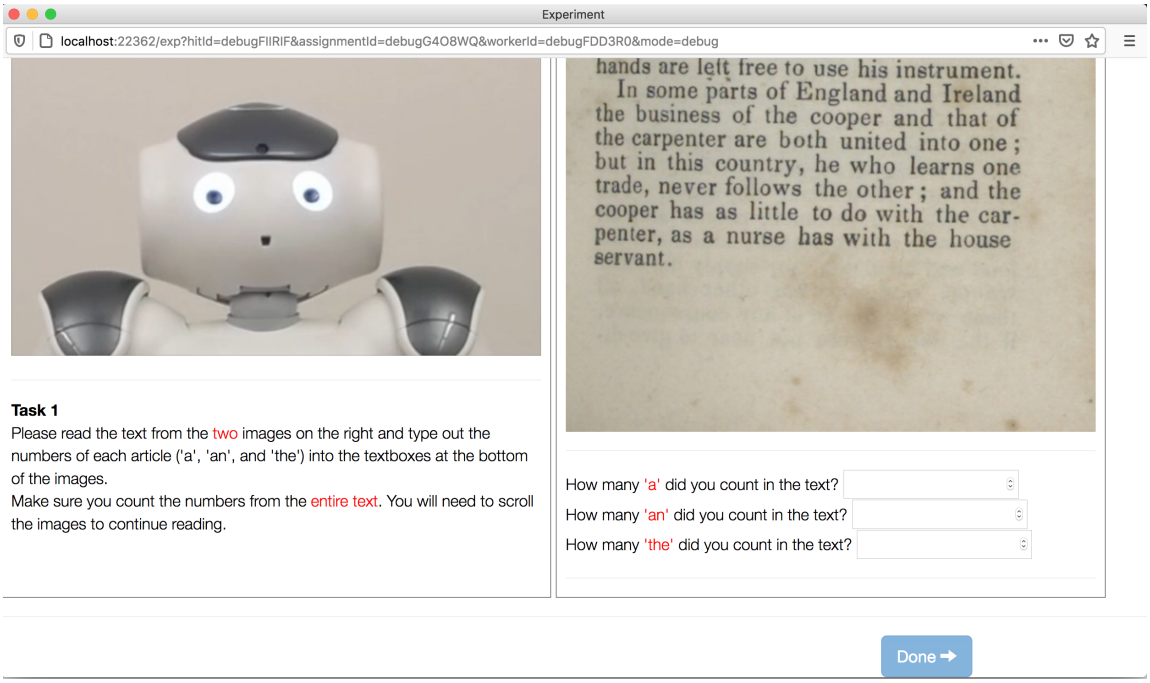

Fig. 3. Webpage of the article counting task used in the study.

In the Norm-based Moral Intervention condition (Norm-based), the robot guided participants through the experiment and gave a norm-based moral intervention between article counting tasks:

> "Thank you! *As a reminder, you are obligated to provide high quality data if you are to accept payment for this task. Therefore, you should find all the articles in the text.* Please click the 'keep going' button to continue this experiment when you are ready."

Condition 3 - Role-based Moral Intervention:

In the Role-based Moral Intervention condition (Role-based), the robot guided participants through the experiment and gave a role-based moral intervention between article counting tasks:

> "Thank you! *As a reminder, you are a paid research participant, and a good paid research participant helps researchers by providing high quality data. Therefore, your responsibility is to find all the articles in the text.* Please click the 'keep going' button to continue this experiment when you are ready."

3.5 Objective measures

Biographical data questionnaire: Participants were asked to provide biographical data including their age, gender, and whether or not English was their native language.

Time on tasks: Throughout the study, we recorded timestamps for when participants began and ended each phase of the study, such as viewing experimental stimuli, completing article counting tasks, and filling out measures. From these timestamps, we calculated how much time it took participants to complete each of the article counting tasks as well

as the overall experiment. For the article counting tasks we calculated a difference score representing the difference between time spent completing the first article counting task and the second article counting task.

Task performance error: In the experimental tasks, participants were asked to count the number of grammatical articles (i.e., "a," "an," and "the") in two small passages of text. Participants then provided the number of each type of article found in each passage, and we calculated two task performance error scores, each representing the difference between the true number of articles in each passage and the participant provided number of articles. From the two task performance error scores, we also calculated the difference between these two scores (i.e., changes in error scores from task 1 to task 2).

3.6 Subjective measures

Theory of Planned Behavior questionnaire: In addition to measuring people's behavior (time on task and performance on task), we also used a Theory of Planned Behavior (TPB) questionnaire to measure people's normative and behavioral beliefs and intentions. According to Ajzen [3], there is no standard TPB questionnaire that can be deployed in all contexts. Rather, formative research is required to construct a questionnaire for specific behavior and population of interest. To construct a TPB questionnaire specifically for this experiment, 15 volunteers were recruited from the Colorado School of Mines campus for a pilot study. During the study, participants were asked to answer a series of free-response questions on their own behavior as a worker who is completing tasks to the best of their ability for payment (see Appendix B). Since this pilot study was designed solely to aid in constructing the TPB questionnaire, no demographic data was collected.

Based on the qualitative data from the pilot study, we identified the top three most frequently responded outcomes from the behavior (*getting approval from your employers and peers*, *senses of satisfaction/fulfillment/pride*, and *spending too much time*), the most frequently responded authoritative and peer normative referent (*my employee* and *my coworkers*), and the top three most frequently responded control factors (*having an enjoyable/interesting job*, *comfortable working environment*, and *having adequate guidance*).

Based on this analysis, we followed standard procedures [2] to construct a TPB questionnaire with subscales containing items related to beliefs that predict behavior, like completing tasks well in return for payment. These included both direct and indirect attitudes towards the behavior (e.g., Completing a job that I am paid to do would be good; completing a job I am paid to do, to the best of my ability, will likely result in my own sense of satisfaction/personal fulfillment), the perceived strength of the norm associated with that behavior (e.g., Most people who are important to me approve of completing a job I am paid to do, to the best of my ability), and future intention to complete that behavior (e.g., In the future when being paid to do a job, I intend to complete it to the best of my ability) (see Appendix A).

In total, the TPB questionnaire contained 21 items and participants responded to these items using 7-point semantic differential scales (e.g., good/bad, likely/unlikely, agree/disagree, true/false). Following standard procedures [20], we calculated TPB sub-scales from the 21 items by combining items in each sub-scale. For hypotheses testing, we used scores derived specifically from the following four sub-scales: indirect attitudes, direct attitudes, norm strength, and future intention. Additionally, in a subset of our experiments (see Study design and procedures), participants completed the TPB questionnaire twice. For analyses we calculated difference scores for each of the four sub-scales, which represented changes in sub-scale scores between administration 1 and administration 2 of the TPB questionnaire when applicable.

3.7 Participants overview

We recruited 367 U.S. participants from Amazon's Mechanical Turk (MTurk), across four human-subjects study designs. Overall, 55 participants were run in Experiment 1, 48 in experiment 2, 108 in Experiment 3, and 105 in Experiment 4, with participants randomly and evenly distributed across conditions within each experiment. All participants were native English speakers. Twenty-three participants were excluded from the dataset because they either responded incorrectly to an attention check item. Another twenty-eight participants were excluded because they either (a) spent less than 45 seconds completing all experimental tasks, or (b) provided response to an objective measure of performance for which the ground truth answer to that measure differed by 200% or more (e.g., participants gave an answer that represented a number which was more than double the correct number). After exclusion, we were left with data from N=316 participants (137 female, 177 male, 2 NA), with ages ranging from 19 to 71 years old (M=39.45, SD=10.96). We will report the participants demographics for each experiments in sections 4, 5, 6, and 7.

4 EXPERIMENT 1: SURVEY-AFTER-INTERVENTION PRE-TEST/POST-TEST DESIGN

4.1 Procedure

Experiment 1 used a pre-test/post-test design. Participants first completed the TPB (pre-test Cronbach's α=0.816), followed by the first article counting task (Task 1). Participants then watched the video of the NAO robot associated with their Moral Intervention condition. Finally, participants completed the second administration of the TPB (post-test Cronbach's α=0.845), followed by the second article counting task (Task 2).

4.2 Participants

55 U.S. participants (19 female, 35 male, 1 NA) were recruited from Amazon's Mechanical Turk. Participant ages ranged from 24 to 71 years old ($M = 37.727, SD = 10.559$). These participants were randomly assigned to the three experimental conditions, resulting in 18 participants in the control condition, 18 participants in the Norm-based Moral Intervention condition, and 19 participants in the Role-based Moral Intervention condition.

4.3 Results

The JASP software package [28] was used to perform Bayesian Analyses of Variance (ANOVA) to assess the effect of moral communication intervention conditions on the changes (i.e., difference scores) in task performance error, time on task, and the changes to TPB questionnaire scores between pre-test and post-test.

Because the use of Bayesian statistical analysis is still relatively uncommon within the HRI community, we will briefly provide some helpful information about this approach. Bayesian statistical analysis has been gaining traction within the scientific community due to a wide variety of benefits that it provides over Frequentist Null Hypothesis Significance Testing (NHST). These include more intuitive and common-sense interpretability, robustness to small sample-sizes due to lack of reliance on the Central Limit Theorem, ability to engage in incremental and flexible sampling, and the ability to gather evidence both in favor *and against* hypotheses. This last point is worth special mention. While under the NHST framework, the test results can only be used to reject a null hypothesis, the Bayesian framework allows Bayes Factors (BFs) to indicate the strength of evidence either for or against any hypotheses under consideration.

While not necessary to the Bayesian paradigm, much Bayesian analysis leverages the calculation of *Bayes Factors*. Bayes Factors (BFs) are essentially odds ratios, representing how much more probable the data observed is under one hypothesis than under another. A $BF_{10} = 7$, for example, would indicate that the analyzed data is seven times

more likely to have been observed under hypothesis H_1 (typically the alternate hypothesis) than under hypothesis H_0 (typically the null is).

In our own Bayes Factor analyses, we use the popular interpretation framework proposed by Lee and Wagenmakers [38], which slightly modifies an approach originally proposed by Jeffreys [29]. This approach is summarized in Table ??. Under this framework, a Bayes Factor $BF_{10} = 7$ (also expressable as $BF_{01} = \frac{1}{7}$) would be interpreted as providing moderate evidence in favor of H_1, and moderate evidence *against* H_0. Typically, a BF greater than 3 or less than $\frac{1}{3}$ is considered sufficiently strong to claim evidence for (or against) an effect, while a BF inside $\langle \frac{1}{3}, 3 \rangle$ is considered inconclusive, and an indication that more data should be collected before a conclusion should be drawn (a procedure that while impermissible under a Frequentist paradigm is both allowed and encouraged under a Bayesian framework). We rely upon such frameworks for the interpretation of our results in the following sections.

Descriptive statistics for all results can be found in Appendix C.

Interpretation of Evidence for the Bayes Factor BF_{12}.

Bayes factor BF_{12}			Interpretation
	>	100	Extreme evidence for H1
30	-	100	Very strong evidence for H1
10	-	30	Strong evidence for H1
3	-	10	Moderate evidence for H1
1	-	3	Anecdotal evidence for H1
	1		No evidence
1/3	-	1	Anecdotal evidence for H2
1/10	-	1/3	Moderate evidence for H2
1/30	-	1/10	Strong evidence for H2
1/100	-	1/30	Very strong evidence for H2
	<	1/100	Extreme evidence for H2

Table 1. Reproduced from Lee and Wagenmakers [38].

Change in Task Performance — Participants in both Norm-based and Role-based conditions made less mistakes in the second task while participants in the control group made more mistakes in task 2. However, our Bayesian analysis provided anecdotal evidence against an effect of intervention strategy (BF 0.485), which suggests that there was likely no difference in task performance between intervention strategies, but that more data would be needed to state this definitively.

Change in Time on Task — Participants in all three conditions spent less time on their second task than the first task. Our analysis also provided anecdotal evidence against an effect of intervention strategy (BF 0.806), which suggests there was likely no difference in time on task changes between intervention strategies.

Change in Attitude — Both our analyses of change in direct attitude and indirect attitude towards attentive crowdworking behavior provided anecdotal evidences against an effect of intervention strategy (BF 0.478 for change in direct attitude, BF 0.770 for change in indirect attitude), which suggests there was likely no difference in direct and indirect attitude changes between intervention strategies, but that more data would be needed to state this definitively.

Changes in Subjective Norm Strength — Our analysis of the changes in subjective norm strength for attentive crowdworking behavior provided moderate evidence against an effect of intervention strategy in Experiment 1 (BF 0.241), suggesting there was no difference in subjective perceptions of norm strength changes between intervention strategies.

Changes in Intention — Our analysis provided moderate evidence against an effect of intervention strategy in Experiment 1 (BF 0.289), suggesting there was no difference in intention changes between intervention strategies.

4.4 Discussion

We ized that both norm-based and role-based moral communication interventions would prime people to comply with role-grounded moral norms, which would be reflected not only in their task performances (H1) and task completion time (H2), but also in their attitudes towards attentive crowdworking behavior (H3), perceptions of the subjective norm strength for attentive crowdworking behavior (H4), and intentions to engage in attentive crowdworking behavior (H5). We also predicted that the norm-based moral communication intervention would have stronger impacts than the role-based intervention. *Based on the results of Experiment 1 alone*, our results would refute these hypotheses, by showing anecdotal to moderate evidence against moral intervention being a factor for changes in task performance (H1), completion time (H2), attitudes (H3), subjective norm strength (H4) and intention (H5).

5 EXPERIMENT 2: TASK-AFTER-INTERVENTION PRE-TEST/POST-TEST DESIGN

5.1 Procedure

Participants first completed the TPB (pre-test Cronbach's α=)0.829, followed by the first article counting task (Task 1). Participants then watched the video of the NAO robot associated with their Moral Intervention condition. Finally, participants completed the second article counting task (Task 2), followed by the second administration of the TPB (post-test Cronbach's α=0.816).

5.2 Participants

48 U.S. participants (24 female, 24 male) were recruited from Amazon's Mechanical Turk. Participant ages ranged from 19 to 62 years old (M = 39.771, SD = 11.092). These participants were randomly assigned to the three experimental conditions, resulting in 16 participants in the control condition, 16 participants in the Norm-based Moral Intervention condition, and 16 participants in the Role-based Moral Intervention condition.

5.3 Results

For Experiment 2, we use the JASP software package to perform the same set of Bayesian ANOVA to assess the effect of Moral Intervention conditions on the changes (i.e., difference scores) in task performance error, time on task, and the changes to TPB questionnaire scores.

Change in Task Performance — Our Bayesian analysis provided moderate evidence against an effect of intervention strategy in Experiment 2 (BF 0.303), suggesting there was no difference in task performance changes between intervention strategies.

Change in Time on Task — Our analysis also provided anecdotal evidence against an effect of intervention strategy (BF 0.541), suggesting there was likely no difference in time on task changes between intervention strategies, but that more data would be needed to state this definitively.

Change in Attitude — Our analysis of change in direct attitude towards attentive crowdworking behavior provided moderate evidence in favor of an effect of intervention strategy (BF 3.081). Post Hoc analysis provided moderate evidence specifically for differences between the role-based and norm-based interventions (BF 5.190). Post Hoc analysis also provided moderate evidence specifically for a difference between the role-based and control interventions (BF 4.205). As shown in Fig. 4, the role-based intervention had the best improvement in direct attitude. Our analysis of change in indirect attitude towards attentive crowdworking behavior provided moderate evidence against an effect of intervention strategy (BF 0.194), suggesting there was no difference in indirect attitude changes between intervention strategies.

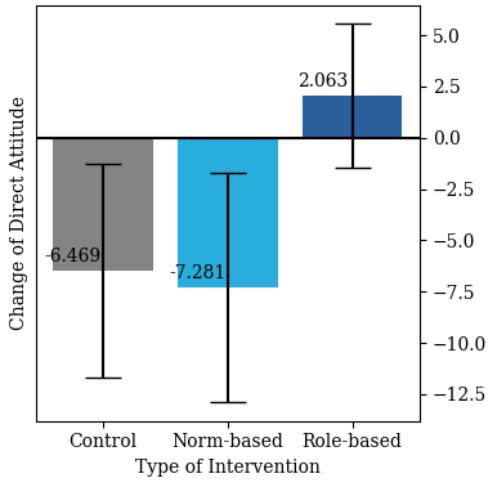

Fig. 4. Change in direct attitude and towards attentive crowdworking behavior by Intervention in Experiment 2. Higher numbers indicate more positive attitudes toward the role-based norms in the second TPB survey (post-intervention) relative to the first TPB survey (pre-intervention). Error bars represent 95% Credible Intervals.

Changes in Subjective Norm Strength — Our analysis of the changes of subjective norm strength for attentive crowdworking behavior provided moderate evidence against an effect of intervention strategy (BF 0.311), suggesting there was no difference in subjective norm strength changes between intervention strategies.

Changes in Intention — Our analysis provided anecdotal evidence against an effect of intervention strategy in Experiment 2 (BF 0.366), suggesting there was likely no difference in intention changes between intervention strategies, but that more data would be needed to state this definitively.

5.4 Discussion

We hypothesized that both norm-based and role-based moral communication interventions would prime people to comply with role-grounded moral norms, which would be reflected not only in their task performances (H1) and task completion time (H2), but also in their attitudes towards attentive crowdworking behavior (H3), subjective perceptions of norm strength for attentive crowdworking behavior (H4), and intentions to engage in attentive crowdworking behavior (H5). We also predicted that the norm-based moral intervention would have stronger impacts than the role-based moral

intervention. *Based on the results of Experiment 2 alone*, our results would demonstrated partial support for hypotheses H3 by providing evidence for the predicted impact of role-based moral interventions on direct attitude towards attentive crowdworking behavior (H3), and our results would refute hypotheses H1, H2, H4 and H5 by demonstrating anecdotal to moderate evidence against effects of moral interventions on changes in task performance (H1), completion time (H2), subjective norm strength (H4) and intention (H5).

6 EXPERIMENT 3: TASK-AFTER-INTERVENTION POST-TEST ONLY DESIGN

6.1 Procedure

Participants first completed the first article counting task (Task 1). Participants then watched the video of the NAO robot associated with their Moral Intervention condition. Finally, participants completed the second article counting task (Task 2), followed by the second administration of the TPB (Cronbach's α not reported as TPB was not a Dependent Variable in this experiment).

6.2 Participants

108 U.S. participants (51 female, 56 male, 1 NA) were recruited from Amazon's Mechanical Turk. Participant ages ranged from 23 to 68 years old ($M = 38.880, SD = 11.379$). These participants were randomly assigned to the three experimental conditions, resulting in 35 participants in the control condition, 35 participants in the Norm-based Moral Intervention condition, and 38 participants in the Role-based Moral Intervention condition.

6.3 Results

For experiments which follow the post-test only design (Experiment 3 and 4), we only used JASP to perform Bayesian ANOVA to assess the effect of Moral Intervention conditions on the changes (i.e., difference scores) in task performance error and time on task.

Change in Task Performance — Our analysis provided moderate evidence against an effect of intervention strategy (BF 0.110), suggesting there was no difference in task performance changes between intervention strategies.

Change in Time on Task — Our analysis provided moderate evidence against an effect of intervention strategy in Experiment 3 (BF 0.195), suggesting there was no difference in time on task changes between intervention strategies.

6.4 Discussion

We hypothesized that both norm-based and role-based moral interventions – especially norm-based interventions – would prime people to comply with role-grounded moral norms, which would be reflected in their task performances (H1) and task completion time (H2), and in particular, that the norm-based moral interventions would be more effective than role-based moral interventions. *Based on the results of Experiment 3 alone*, our results would refute both of these hypotheses by showing moderate evidence against differences in change in task performance (H1) and completion time (H2).

7 EXPERIMENT 4: SURVEY-AFTER-INTERVENTION POST-TEST ONLY DESIGN

7.1 Procedure

Participants first completed the first article counting task (Task 1). Participants then watched the video of the NAO robot associated with their Moral Intervention condition. Finally, participants completed the second administration of the TPB (Cronbach's α not reported as TPB was not a Dependent Variable in this experiment), followed by the second article counting task (Task 2).

7.2 Participants

105 U.S. participants (43 female, 62 male) were recruited from Amazon's Mechanical Turk. Participant ages ranged from 24 to 70 years old ($M = 40.781, SD = 10.637$). These participants were randomly assigned to the three experimental conditions, resulting in 37 participants in the control condition, 33 participants in the Norm-based Moral Intervention condition, and 38 participants in the Role-based Moral Intervention condition.

7.3 Results

For Experiment 4, we used JASP packages to perform the same set of Bayesian ANOVA on the effect of Moral Intervention conditions on the changes (i.e., difference scores) in task performance error and time on task.

Change in Task Performance — The Bayesian ANOVA conducted for Experiment 4 provided strong evidence in favor of an effect of intervention strategy (BF 15.138). Post Hoc analysis provided strong evidence for differences in the change of error between the role-based intervention and control intervention (BF 17.627). Post Hoc analysis also provided moderate evidence for differences in the change between the role-based intervention and the norm-based intervention (BF 7.774). As shown in the Fig. 5, in Experiment 4, the role-based intervention had the best improvement in task performance.

Change in Time on Task — Our analysis provided strong evidence against an effect of communication intervention strategy in Experiment 4 (BF 0.089), suggesting there was no difference in time on task changes between intervention strategies.

7.4 Discussion

We hypothesized that both norm-based and role-based moral interventions – especially norm-based interventions – would prime people to comply with role-grounded moral norms, which would be reflected in their task performances (H1) and task completion time (H2), and in particular, that the norm-based moral interventions would be more effective than role-based moral interventions. *Based on the results of Experiment 4 alone*, our results would provide partial support for hypotheses H1 by demonstrating predicted impact of role-based moral interventions on task performance (H1), but would refute hypothesis H2 by showing strong evidence against differences in change in completion time (H2).

8 GENERAL DISCUSSION

We hypothesized that both role-based and norm-based moral interventions staged by robots, especially norm-based moral interventions, will strengthen participants' role-grounded moral norms, so as to improve task performance (H1), time on task (H2), positive attitude towards attentive crowdworking behavior (H3), subjective norm strength for attentive crowdworking behavior (H4), and intention to engage in attentive crowdworking behavior (H5).

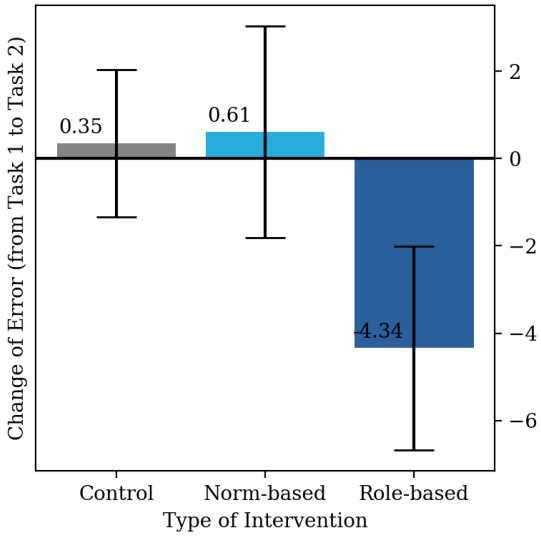

Fig. 5. Change in error (error made in task 2 - error made in task 1) by Intervention in Experiment 4. Lower numbers indicate better performance on the second task (post-intervention) relative to the first task (pre-intervention). Error bars represent 95% Credible Intervals.

Our results partially support hypotheses H1 and H3 by providing evidence for the predicted impact of role-based moral communication interventions on task performance (H1) and direct attitude towards attentive crowdworking behavior (H3) for specific experimental procedures. Our results refute hypotheses H2, H4, and H5 by providing evidence against a difference in change in time on tasks (H2), change in subjective norm strength for attentive crowdworking behavior (H4) and change in intention to engage in attentive crowdworking behavior (H5) between the moral intervention groups and the control group. Specifically, our results suggest two major findings related to the effects of Moral Interventions as well as the relationship between Moral Intervention and inner states:

(1) Participants' performance became more accurate in the second task *only* after receiving a Role-based intervention and completing a TPB questionnaire between the intervention and the second task (Experiment 4).
(2) Participants gained more positive direct attitudes towards the role-based norm of attentive crowdworking behavior *only* after receiving a Role-based intervention and completing a second task between the intervention and the post-experimental TPB questionnaire (Experiment 2).

8.1 Findings on task performance

For task performance, we found strong evidence for beneficial impact of the Role-based intervention in situations where participants were prompted to consider their beliefs after receiving the Role-based Intervention. Specifically, participants who saw the Role-based Moral Intervention followed by the TPB measure (Experiment 4), showed improved performance between tasks, whereas participants who received the Norm-based Moral Intervention or Control Intervention under the same procedures did not.

This observed improvement in the Role-based Moral Intervention condition may point to the influence of *reflective practice* provided by completing the TPB questionnaire immediately after receiving the Moral Intervention and immediately prior to engaging in the second task. Specifically, the TPB questionnaire included several items with wording that may have heightened sensitivity specifically to the language used in the Role-based Moral Intervention condition. For instance, in several cases, the TPB questionnaire specifically highlighted that participants are paid to do a job (e.g., “Most people who are important to me approve of completing a job I am paid to do to the best of my ability”). Similar language is mirrored in the Role-based Moral Intervention condition, i.e., "As a reminder you are a paid research participant..."

Although participants received a similar reminder in the Norm-based Moral Intervention condition, that condition was specifically designed not to highlight the relationship between their payment and their role as a participant. Thus, completing the TPB questionnaire immediately after receiving the Role-based Moral Intervention may have created a situation in which the questionnaire itself served not only as a measurement of beliefs as intended, but moreover as an exercise to reflect on the role-based norms. This exercise may also have made the role-based treatment more salient in ways that were not applicable to the Norm-based and Control interventions. Additionally, we did not see this same effect of TPB exercise on task performance for the other experimental procedures which included the Role-based Moral Interventions because either (a) the TPB exercises did not precede the second task , or (b) the TPB exercises preceded both tasks which would create a "ceiling" for change between the two tasks if the first task was influenced by the TPB exercise in the same way.

From the Confucian role ethics perspective, moral development in a specific context critically depends on whether the practice of norms can lead to a better way of living one’s communal roles and reflecting on oneself. As suggested earlier, it is likely that the TPB questionnaire used in this study provided an opportunity for participants to reflect on their professional roles in the crowdsourcing community and their relationships to other crowdworkers and requesters. However, it is worth noting that a critical criterion for the effectiveness of the Role-based Moral Intervention is whether participants have developed reflective awareness of the social roles they assume in the communal context. In other words, when the robot initiated the Role-based Moral Intervention, its effectiveness would be different between participants, as participants may have different levels of awareness of their social role as crowdworkers, based on factors such as how long participants have worked as crowdworkers. Future studies could investigate whether participants’ perception of their role as crowdworkers and their professional experience would make a difference for the effectiveness of Role-based Moral Interventions.

8.2 Findings on direct attitude

The second major set of findings are related to our subjective measures and changes in direct attitudes towards crowdworking behaviors. Specifically, we found evidence again for the effects of the Role-based Moral Intervention conditions under the procedures of Experiment 2. Specifically, participants who received the Role-based Moral Intervention condition and completed the study in Experiment 2 (i.e., who completed the post-intervention TPB questionnaire after the second task) reported positive changes in attitudes towards attentive crowdworking behaviors from the first TPB questionnaire to the second questionnaire, while participants in the Norm-based and Control intervention conditions reported negative changes between the first and the second TPB questionnaire.

These findings may be related to the effects of performing immediate moral practice. In Confucian role ethics, moral development includes three components: observation, reflection, and practice [68]. Accordingly, humans not only need to observe others act and interact in society and reflect on themselves, but also need to integrate and practice moral

principles in actions, and reiterate the process of observation, reflection, and practice [37, 66]. If we link this moral development model to our experiment, in Experiment 2, when participants received the Role-based Moral Intervention highlighting their role as an attentive crowdworker and then immediately had the opportunity to enact that role described in the intervention by completing the second task, the role-based norm may have been strengthened. This could also explain why the positive change was only observed after Role-based Moral Intervention in Experiment 2 but not in Experiment 1 (in which the second task was completed after the post-intervention questionnaire).

The effect of the combination of Role-based Moral intervention, role-based practice, and the self-reflective activity (e.g., TPB exercises) discovered in this study also provides empirical evidence for a crucial philosophical statement in Confucian role ethics: effective moral growth requires the interactive association between practice and self-reflective learning [63]. If an agent only practices without reflecting on their roles and associated moral obligations, it is a waste of labor for the agent in their moral development. If the agent only reflects but without any attempt to put reflective learning experience into practice, then the agent can never understand the true meaning of morality or improve their moral expertise.

From the Confucian role ethics perspective, such reiterative processing is critical for moral development from the moral beginner and the developing learner to the *junzi* (i.e., morally superior person). However, to be able to achieve at the level of *junzi*, the agent needs to participate in self-reflective practice continuously in a much longer or even lifelong term just as emphasized by Confucius, "*It (the task of self-cultivation) might be compared to the task of building up a mountain: if I stop even one basketful of earth short of completion, then I have stopped completely*" (*Analects*, 9.19).

9 LIMITATIONS

Before concluding, we will briefly discuss some of the limitations of our approach, which motivate possible directions for future work and points for further reflection. One methodological limitation of this study is its nature as a video-based online experiment with crowdworkers. This necessarily meant that participants were not given a chance to directly interact with the robot. Previous research has shown that in advice-giving scenarios, significant differences can arise between observation and interaction [59]. While our use of a crowdsourcing platform was necessary due to the COVID-19 pandemic [18], it would be a natural direction for future work to follow-up our experiment with an in-person version to confirm our results. Such a follow-up would also have a number of other benefits. For example, while in this work we chose a particular crowdworking norm that matched our subject community, it would be valuable to study a wider variety of norms and contexts. In our previous work, we have shown that the effectiveness of robot explanations grounded in different aspects of role-grounded moral norms are highly dependent on nuanced aspects of robots' environmental and social contexts [64]. An in-person experiment would necessarily require adapting our task and norm-of-interest to align with the community norms of our in-person population. And, as with any experimental effort involving a large suite of tests, it is possible that some of our findings may have been false positives; an in-person experiment confirming our results would help to alleviate this concern.

It would also be interesting for future work to consider cross-cultural differences in the effectiveness of the types of moral language we consider in this work. In our past work, we have observed that individuals' cultural orientations mediate the impacts of robots' moral advice [34]. Similarly, as people from different cultural backgrounds are also likely to have different moral traditions which lead to different sensitivities to different ethical systems. For example, people from Eastern Asia could be more receptive to role-based ethical traditions (as they have been deeply influenced by Confucianism for a long time), while people in the United States are likely not familiar with such ideas conveyed by role-based ethical traditions. On the other hand, unfamiliarity with a moral system could also provide potential learning

opportunities, as the ensuing novelty and foreign experience could be more stimulating for moral reflection. While we would argue against any attempt to automatically "perceive" the cultures of interactants (cf. [65], cross-cultural differences could lead to inform the design of moral robotic technologies by helping designers better attend to the values and priorities of the communities they are designing for and with [21, 40, 41].

Finally, it is worth examining ethical risks imposed by our experiment. Had our examined moral norm (working to ensure high-quality scientific data in the course of crowdworking) not been a self-expressed standard and value of many crowdworkers, it would be reasonable to raise ethical concerns over what could be perceived as an attempt to persuade workers to maximize their efforts in the name of productivity. On the one hand, our choice of population and norm is not necessarily representative of the type of future domain in which we see social robots being used to help teammates cultivate their moral selves. On the other hand, as with most social robotics technologies, there is always a risk of unforeseen or unintentioned dual-use. Indeed, most of the work in social HRI tries to persuade, steer behavior, encourage engagement, or otherwise encourage positive perceptions and interactions in some way, all of which could be misused as part of other design efforts. As HRI researchers, it is important that we remain cognizant that not all future uses of the technologies we are developing will align with our personal motivations. Persuasive robots could indeed be used by future corporations to encourage worker compliance and overwork in toxic ways, or used by governments to manufacture or encourage compliance with socially detrimental or inequitable norms and laws. While our aim in this work is to use robots' persuasive power to help people cultivate their moral selves, these risks nevertheless remain.

10 CONCLUSION

In this work, we evaluated two intervention strategies for robot moral communication: a norm-based strategy grounded in deontological ethics, and a role-based strategy grounded in role ethics. Our results suggests two major findings: (1) reflective exercises may increase the efficacy of role-based moral language and (2) performing immediate moral practice after receiving role-based moral interventions could help peoples' role-centric moral development by promoting positive attitudes towards behaviors emphasised by the role-grounded moral norms used in such interventions. Our findings suggest that our TPB self-report measurement provided an opportunity for role-based reflection, leading to increased efficacy of role-based interventions. Accordingly, future work should investigate other reflective exercises that may facilitate norm-based interventions.

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] Icek Ajzen. 1991. The theory of planned behavior. *Organizational behavior and human decision processes* 50, 2 (1991), 179–211.
[2] Icek Ajzen. 2006. Constructing a theory of planned behavior questionnaire.
[3] Icek Ajzen. 2013. Theory of planned behaviour questionnaire. *Measurement instrument database for the social science* (2013), 2–9.
[4] R. T. Ames. 2010. Achieving personal identity in Confucian role ethics: Tang Junyi on human nature as conduct. *Oriens Extremus* (2010), 143–166.
[5] R. T. Ames. 2011. *Confucian role ethics: A vocabulary.*
[6] R. T. Ames. 2016. Theorizing "person" in Confucian ethics: a good place to start. *Sungkyun journal of east asian studies* 16 (2016), 141–162.
[7] Susan Leigh Anderson and Michael Anderson. 2011. A Prima Facie Duty Approach to Machine Ethics and Its Application to Elder Care. In *Proc. 12th AAAI Conf. on HRI in Elder Care*. 6 pages. http://dl.acm.org/citation.cfm?id=2908724.2908725

[8] Ilaria Baroni, Marco Nalin, Mattia Coti Zelati, Elettra Oleari, and Alberto Sanna. 2014. Designing motivational robot: how robots might motivate children to eat fruits and vegetables. In *Int'l Symp. Robot and Human Interactive Communication*.

[9] Christoph Bartneck, Timo Bleeker, Jeroen Bun, Pepijn Fens, and Lynyrd Riet. 2010. The influence of robot anthropomorphism on the feelings of embarrassment when interacting with robots. *Paladyn, Journal of Behavioral Robotics* 1, 2 (2010), 109–115.

[10] Daniel Bell and Pei Wang. 2020. *Just hierarchy: Why social hierarchies matter in China and the rest of the world.* Princeton University Press.

[11] Adam Briggle and Carl Mitcham. 2012. *Ethics and Science: An Introduction.* https://doi.org/10.1017/CBO9781139034111

[12] Gordon Briggs. 2014. Blame, What is it Good For?. In *RO-MAN WS:Phil.Per.HRI.* Edinburgh, Scotland.

[13] 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 (2014), 343–355.

[14] Penelope Brown, Stephen C Levinson, and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*. Vol. 4. Cambridge university press.

[15] Vijay Chidambaram, Yueh-Hsuan Chiang, and Bilge Mutlu. 2012. Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues. In *International conference on Human-Robot Interaction (HRI)*. ACM.

[16] Mark Coeckelbergh. 2010. Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology* 12, 3 (2010).

[17] Derek Cormier, Gem Newman, Masayuki Nakane, James E Young, and Stephane Durocher. 2013. Would you do as a robot commands? An obedience study for human-robot interaction. In *International Conference on Human-Agent Interaction*.

[18] David Feil-Seifer, Kerstin S Haring, Silvia Rossi, Alan R Wagner, and Tom Williams. 2020. Where to next? The impact of COVID-19 on human-robot interaction research. , 7 pages.

[19] Luciano Floridi and Jeff W Sanders. 2004. On the morality of artificial agents. *Minds and machines* 14, 3 (2004), 349–379.

[20] Jillian Francis, Martin P Eccles, Marie Johnston, AE Walker, Jeremy M Grimshaw, Robbie Foy, Eileen FS Kaner, Liz Smith, and Debbie Bonetti. 2004. Constructing questionnaires based on the theory of planned behaviour: A manual for health services researchers.

[21] Batya Friedman. 1996. Value-sensitive design. *interactions* 3, 6 (1996), 16–23.

[22] Bertram Gawronski and Jennifer S Beer. 2017. What makes moral dilemma judgments "utilitarian" or "deontological"? *Social Neuroscience* 12, 6 (2017), 626–632.

[23] Jaap Ham, René Bokhorst, Raymond Cuijpers, David van der Pol, and John-John Cabibihan. 2011. Making robots persuasive: the influence of combining persuasive strategies (gazing and gestures) by a storytelling robot on its persuasive power. In *International conference on social robotics*. Springer, 71–83.

[24] Marija Ham, Marina Jeger, and Anita Frajman Ivković. 2015. The role of subjective norms in forming the intention to purchase green food. *Economic research-Ekonomska istraživanja* 28, 1 (2015), 738–748.

[25] Ryan Blake Jackson and Tom Williams. 2018. Robot: Asker of questions and changer of norms? *Proceedings of ICRES* (2018).

[26] Ryan Blake Jackson and Tom Williams. 2019. Language-capable robots may inadvertently weaken human moral norms. In *Companion of the 14th ACM/IEEE International Conference on Human-Robot Interaction (alt.HRI)*. IEEE, 401–410.

[27] Ryan Blake Jackson, Tom Williams, and Nicole Smith. 2020. Exploring the Role of Gender in Perceptions of Robotic Noncompliance. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 559–567.

[28] JASP Team et al. 2016. Jasp. *Version 0.8. 0.0. software* (2016).

[29] Harold Jeffreys. 1961. *Theory of Probability.* Clarendon Press, Oxford.

[30] 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*. 229–236.

[31] Peter H Kahn, Takayuki Kanda, Hiroshi Ishiguro, Brian T Gill, Jolina H Ruckert, Solace Shen, Heather 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 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Boston, MA, 33–40.

[32] Florian G Kaiser, Gundula Hübner, and Franz X Bogner. 2005. Contrasting the theory of planned behavior with the value-belief-norm model in explaining conservation behavior 1. *Journal of applied social psychology* 35, 10 (2005), 2150–2170.

[33] James Kennedy, Paul Baxter, and Tony Belpaeme. 2014. Children comply with a robot's indirect requests. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction (HRI)*. 198–199.

[34] 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*. 10–18.

[35] Ann Knabe. 2012. Applying Ajzen's theory of planned behavior to a study of online course adoption in public relations education. (2012).

[36] Christine Korsgaard. 1993. The reasons we can share: An attack on the distinction between agent-relative and agent-neutral values. *Social Philosophy and Policy* (1993).

[37] Karyn Lai. 2007. Understanding Confucian ethics: Reflections on moral development. *Australian Journal of Professional and Applied Ethics* 9, 2 (2007), 21–27.

[38] Michael D Lee and Eric-Jan Wagenmakers. 2014. *Bayesian cognitive modeling: A practical course.* Cambridge university press.

[39] Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, and Paul Rybski. 2012. Ripple effects of an embedded social agent: a field study of a social robot in the workplace. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 695–704.

[40] Jon A Leydens and Jessica Deters. 2017. Confronting intercultural awareness issues and a culture of disengagement: An engineering for social justice framework. In *2017 IEEE International Professional Communication Conference (ProComm)*. IEEE, 1–7.

[41] Jon A Leydens and Juan C Lucena. 2017. *Engineering justice: Transforming engineering education and practice*. John Wiley & Sons.

[42] JeeLoo Liu. 2017. Confucian robotic ethics. In *International Conference on the Relevance of the Classics under the Conditions of Modernity: Humanity and Science*.

[43] Bertram F Malle. 2016. Integrating Robot Ethics and Machine Morality: The Study and Design of Moral Competence in Robots. *Ethics and Info. Tech.* (2016).

[44] Bertram F Malle and Matthias Scheutz. 2014. Moral Competence in Social Robots. In *Symposium on Ethics in Science, Technology and Engineering*. IEEE.

[45] Bertram F Malle, Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano. 2015. Sacrifice One for the Good of Many?: People Apply Different Moral Norms to Human and Robot Agents. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. Portland, OR, 117–124.

[46] Cees Midden and Jaap Ham. 2012. The illusion of agency: the influence of the agency of an artificial agent on its persuasive power. In *International Conference on Persuasive Technology*. Springer, 90–99.

[47] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 72–78.

[48] A. T. Nuyen. 2007. Confucian ethics as role-based ethics. *International philosophical quarterly* 47 (2007), 315–328.

[49] Raul Benites Paradeda, Maria José Ferreira, João Dias, and Ana Paiva. 2017. How Robots Persuasion based on Personality Traits May Affect Human Decisions. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 251–252.

[50] Daniel J Rea, Denise Geiskkovitch, and James E Young. 2017. Wizard of awwws: Exploring psychological impact on the researchers in social HRI experiments. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 21–29.

[51] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard, and Alan R Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. 101–108.

[52] H. Rosemont Jr. 2015. *Against Individualism: A Confucian Rethinking of the Foundations of Morality, Politics, Family, and Religion (Philosophy and Cultural Identity)*.

[53] Henry Rosemont Jr and Roger T Ames. 2016. *Confucian role ethics: A moral vision for the 21st century?* Vandenhoeck & Ruprecht.

[54] Jeffrey N Rouder. 2014. Optional stopping: No problem for Bayesians. *Psychonomic bulletin & review* 21, 2 (2014), 301–308.

[55] Eduardo Benítez Sandoval, Jürgen Brandstetter, and Christoph Bartneck. 2016. Can a robot bribe a human?: The measurement of the negative side of reciprocity in human robot interaction. In *Int'l Conf. on Human Robot Interaction (HRI)*.

[56] Keith H Seddon. 2003. Epictetus. In *International encyclopedia of philosophy*. https://www.iep.utm.edu/epictetu/

[57] 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*. 69–77.

[58] Richard L Solomon. 1949. An extension of control group design. *Psychological bulletin* 46, 2 (1949), 137.

[59] Megan Strait, Cody Canning, and Matthias Scheutz. 2014. Let me tell you! investigating the effects of robot communication strategies in advice-giving situations based on robot appearance, interaction modality and distance. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction (HRI)*.

[60] Sarah Strohkorb Sebo, Margaret Traeger, Malte Jung, and Brian Scassellati. 2018. The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 178–186.

[61] Hamish Tennent, Solace Shen, and Malte Jung. 2019. Micbot: A peripheral robotic object to shape conversational dynamics and team performance. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 133–142.

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

[63] Fengyan Wang. 2004. Confucian thinking in traditional moral education: key ideas and fundamental features. *Journal of Moral Education* (2004), 429–447.

[64] Ruchen Wen, Zhao Han, and Tom Williams. 2022. Teacher, Teammate, Subordinate, Friend: Generating Norm Violation Responses Grounded in Role-based Relational Norms.. In *HRI*. 353–362.

[65] Tom Williams. 2022. Race in the Eye of the Robot Beholder: Against Racial Representation, Recognition, and Reasoning in Robotics Research. In *Proceedings of the 2022 Inclusive HRI Workshop on Equity and Diversity in Design, Application, Methods, and Community*.

[66] 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 (alt.HRI)*. 25–33.

[67] Katie Winkle, Séverin Lemaignan, Praminda Caleb-Solly, Ute Leonards, Ailie Turton, and Paul Bremner. 2019. Effective persuasion strategies for socially assistive robots. In *International Conference on Human-Robot Interaction (HRI)*.

[68] Qin Zhu. 2018. Engineering ethics education, ethical leadership, and Confucian ethics. *International Journal of Ethics Education* (2018), 1–11.

[69] Qin Zhu. 2020. Ethics, society, and technology: A Confucian role ethics perspective. *Technology in society* (2020).

[70] 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* (2020), 1–16.

A SUBJECTIVE MEASURE: THEORY OF PLANNED BEHAVIOR QUESTIONNAIRE

(1) Completing a job I am paid to do, to the best of my ability, would be (1=very bad, 5=very good).
(2) Completing a job I am paid to do, to the best of my ability, would be (1=very unpleasant, 5=very pleasant).
(3) Most people who are important to me approve of completing a job I am paid to do, to the best of my ability. (1=Strongly disagree, 5=Strongly agree)
(4) Most people like me when being a paid employee would complete a job they are paid to do, to the best of their abilities. (1=Highly unlikely, 5=Highly likely)
(5) I am confident that I can complete a job I am paid to do, to the best of my ability. (1=True, 5=False)
(6) My ability of completing a job I am paid to do, to the best of my ability, is up to me. (1=Strongly disagree, 5=Strongly agree)
(7) In the future when being paid to do a job, I intend to complete it to the best of my ability. (1=True, 5=False)
(8) Completing a job I am paid to do, to the best of my ability, will likely result in my own sense of satisfaction/personal fulfillment. (1=Highly unlikely, 5=Highly likely)
(9) Completing a job I am paid to do, to the best of my ability, will likely result in my getting approval from my employers and peers. (1=Highly unlikely, 5=Highly likely)
(10) Completing a job I am paid to do, to the best of my ability, will likely result in my spending too much time. (1=Highly unlikely, 5=Highly likely)
(11) My own sense of satisfaction/personal fulfillment is (1=very bad, 5=very good).
(12) My getting approval from my employers and peers is (1=very bad, 5=very good).
(13) In general, my spending too much time is (1=very bad, 5=very good).
(14) My employer thinks that I (1=should, 5=shouldn't) complete a job I am paid to do, to the best of my ability.
(15) Most of my coworkers will complete a job they are paid to do, to the best of their abilities. (1=True, 5=False)
(16) I expect that I will have an enjoyable/interesting job when I need to complete a job I am paid to do, to the best of my ability. (1=Highly unlikely, 5=Highly likely)
(17) I expect that I will have a comfortable working environment when I need to complete a job I am paid to do, to the best of my ability. (1=Highly unlikely, 5=Highly likely)
(18) I expect that I will have adequate guidance when I need to complete a job I am paid to do, to the best of my ability. (1=Highly unlikely, 5=Highly likely)
(19) Having an enjoyable/interesting job would enable me to complete a job I am paid to do, to the best of my ability. (1=Strongly disagree, 5=Strongly agree)
(20) Having a comfortable working environment would enable me to complete a job I am paid to do, to the best of my ability. (1=Strongly disagree, 5=Strongly agree)
(21) Having adequate guidance would enable me to complete a job I am paid to do, to the best of my ability. (1=Strongly disagree, 5=Strongly agree)

B THEORY OF PLANNED BEHAVIOR PILOT QUESTIONNAIRE

Part 1

(1) What do you see as the advantages of completing a job you are paid to do, to the best of your ability?
(2) What do you see as the disadvantages of completing a job you are paid to do, to the best of your ability?
(3) What else comes to mind when you think about completing a job you are paid to do, to the best of your ability?

Part 2

When it comes to your completing a job you are paid to do, to the best of your ability, there might be individuals or groups who would think you should or should not perform this behavior.

(4) Please list the individuals or groups who would approve or think you should complete a job you are paid to do, to the best of your ability.
(5) Please list the individuals or groups who would disapprove or think you should not complete a job you are paid to do, to the best of your ability.

Sometimes, when we are not sure what to do, we look to see what others are doing.

(6) Please list the individuals or groups who are most likely to complete a job they are paid to do, to the best of their ability.
(7) Please list the individuals or groups who are least likely to complete a job they are paid to do, to the best of their ability.

Part 3

(8) Please list any factors or circumstances that would make it easy or enable you to complete a job you are paid to do, to the best of your ability.
(9) Please list any factors or circumstances that would make it difficult or prevent you from complete a job you are paid to do, to the best of your ability.

C DESCRIPTIVE STATISTICS

	Control	Norm-based	Role-based
Change in Error	M=2, SD=7.53	M=1, SD=6.1	M=-1.47, SD=4.68
Change in Completion Time (in seconds)	M=-187.0, SD=176.1	M=-38.24, SD=214.8	M=-114.91, SD=215.5
Change in Direct Attitude	M=-6.47, SD=10.71	M=-7.28, SD=11.47	M=2.06, SD=7.17
Change in Indirect Attitude	M=0.99, SD=8.65	M=6.2, SD=13.43	M=-1.13, SD=9.04
Change in Subjective Norm Strength	M=-4.08, SD=14.39	M=0.28, SD=5.71	M=-2.08, SD=10.82
Change in Intention	M=1.33, SD=26.75	M=14.89, SD=36.98	M=4.16, SD=27.90

Table 2. Means and standard deviations for: change of error (error made in task 2 - error made in task 1), change in time on task in seconds (time spent on task 2 - time spent on task 1), change in direct attitude towards attentive crowdworking behavior, change in indirect attitude towards attentive crowdworking behavior, change in subjective norm strength for attentive crowdworking behavior, and change in intention to engage in attentive crowdworking behavior in Experiments 1.

	Control	Norm-based	Role-based
Change in Error	M=-6.06, SD=17.13	M=0.25, SD=8.32	M=-5.31, SD=15.88
Change in Completion Time (in seconds)	M=-74.3, SD=109.3	M=-171.56, SD=231.2	M=-175.92, SD=156.7
Change in Direct Attitude	M=0.81, SD=7.21	M=0.44, SD=3.29	M=-3.08, SD=9.48
Change in Indirect Attitude	M=-0.07, SD=8.93	M=-0.94, SD=8.92	M=1.79, SD=12.31
Change in Subjective Norm Strength	M=-3.63, SD=9.9	M=-4.91, SD=7.14	M=-0.59, SD=9.72
Change in Intention	M=4.19, SD=35.62	M=-2.63, SD=38.19	M=17.44, SD=37.88

Table 3. Means and standard deviations for: change of error (error made in task 2 - error made in task 1), change in time on task in seconds (time spent on task 2 - time spent on task 1), change in direct attitude towards attentive crowdworking behavior, change in indirect attitude towards attentive crowdworking behavior, change in subjective norm strength for attentive crowdworking behavior, and change in intention to engage in attentive crowdworking behavior in Experiments 2.

	Control	Norm-based	Role-based
Change in Error	M=0.89, SD=9.6	M=-0.34, SD=11.02	M=-0.76, SD=7.32
Change in Completion Time (in seconds)	M=-73.83, SD=175.3	M=-167.16, SD=451.9	M=-88.89, SD=182.6

Table 4. Means and standard deviations for: change of error (error made in task 2 - error made in task 1) and change in time on task in seconds (time spent on task 2 - time spent on task 1) in Experiments 3.

	Control	Norm-based	Role-based
Change in Error	M=0.35, SD=5.24	M=0.61, SD=7.09	M=-4.34, SD=7.04
Change in Completion Time (in seconds)	M=-93.21, SD=226.9	M=-101.3, SD=221.0	M=-92.09, SD=245.4

Table 5. Means and standard deviations for: change of error (error made in task 2 - error made in task 1) and change in time on task in seconds (time spent on task 2 - time spent on task 1) in Experiments 4.