On Further Reflection... Moral Reflections Enhance Robotic Moral Persuasive Capability

Ruchen Wen[1][0000−0003−1590−1787], Boyoung Kim[2][0000−0003−1214−0169], Elizabeth Phillips[2][0000−0001−9511−3275], Qin Zhu[3][0000−0002−6673−1901], and Tom Williams[1][0000−0001−7921−771X]

[1] Colorado School of Mines, Golden, CO, USA
rwen@mines.edu,twilliams@mines.edu
[2] George Mason University, Fairfax, VA, USA bkim55,ephill3@gmu.edu
[3] Virginia Tech, Blacksburg, VA, USA qinzhu@vt.edu

Abstract. To enable robots to exert positive moral influence, we need to understand the impacts of robots' moral communications, the ways robots can phrase their moral language to be most clear and persuasive, and the ways that these factors interact. Previous work has suggested, for example, that for certain types of robot moral interventions to be successful (i.e., moral interventions grounded in particular ethical frameworks), those interventions may need to be followed by opportunities for moral reflection, during which humans can critically engage with not only the contents of the robot's moral language, but also with the way that moral language connects with their social-relational ontology and broader moral ecosystem. We conceptually replicate this prior work (N=119) using a design that more precisely manipulates moral reflection. Our results confirm that opportunities for moral reflection are indeed critical to the success of robotic moral interventions—regardless of the ethical framework in which those interventions are grounded.

Keywords: human-robot communication, role ethics, moral influence

1 Introduction

Research has shown that language-capable robots hold unique persuasive capability over human interactants. Robots can use their language not only to encourage human compliance with requests and commands [4, 7, 31, 24, 3], but also (intentionally or unintentionally [16, 15]) to influence human systems of moral and social norms [28, 35, 41]. Critically, robots are not only able to influence the behaviors of those they directly interact with, but also may influence others indirectly. That is, robots may influence the behaviors their human interactants choose to perform around other humans [35]. These "ripple effects" thus have the potential to more broadly affect interactants' social and moral ecosystems, above and beyond their immediate interactions with robots. It is thus critical to understand how best to steer robots' potential for moral influence.

Steering Robots' Potential for Moral Influence

To prevent robots from exerting negative moral influence, language-capable robots must avoid violating moral norms or causing moral harm, and be able to explicitly communicate moral conceptions and values. For example, robots must be able to reject unethical commands given by interlocutors, and explain or justify the reason(s) for their non-compliance by highlighting how such commands violate moral principles [42, 33].

Moreover, given robots' unique persuasive power, they must be able to leverage their persuasive capability to exert *positive* moral influence. By demonstrating positive moral tendencies, robots might serve as "moral mediators" that inspire human interactants to cultivate their own moral tendencies. By issuing blame-laden moral rebukes, robots might emphasize the importance of key moral norms, and encourage their adherence by interactants [46]. And just as negative moral influence may cause negative "ripple effects", we must also consider the opportunities for robots to exert *positive influence*, and for this positive influence to similarly result in positive "ripple effects [28]."

A social-relational approach would suggest that robots can encourage, emphasize, and reinforce moral norms within communal contexts by leveraging the power resulting from the normative influences on human-robot relationships. This relational approach can be understood through the lens of a Confucian ethical framework, in which people cultivate self-reflections and virtuous tendencies through daily interaction with others [2], and in which people's moral self-reflection can often be initiated or influenced by other's words and actions [46]. Through their use of moral language, robots may help interactants cultivate their moral selves and contribute to a flourishing moral ecology for human-robot interaction that allows humans to grow.

To best leverage robots' persuasive capability to exert positive moral influence, we need to understand different forms of moral language, and the acute impacts of those different forms. But critically, as we will discuss in the next section, previous work argues that the effectiveness of different types of moral language could be mediated by the structure of the interactions in which they are embedded.

The Structure of Moral Interventions

The impacts of a robot's moral language are mediated by a host of contextual factors. One such contextual factor is the structure of the interaction that surrounds a robot's moral intervention. Previous work from [40], for example, suggested that for certain types of robot moral interventions to be successful, those interventions may need to be followed by opportunities for moral reflection, in which interactants can take their time to examine and digest the information they receive from the robot, and thus more deeply engage with the content of the robot's moral language. If this were the case, it would have significant impacts on the design of the broader interaction structure of robotic moral interventions. This suggestion by [40], however, was based on an emergent observation from an

experiment not explicitly designed to test for the effect of moral reflection. This creates an obvious need to formally verify this suggestion.

In this work, we thus present a conceptual replication and extension of [40] (N=119), following the same study procedures, but designed to systematically investigate the impacts of moral reflection on the effectiveness of robots' moral language on human behavior. Our results indicate that opportunities for moral reflection are indeed critical to the success of robots' moral interventions and their associated perlocutionary goals—regardless of the ethical framework in which those interventions are grounded.

2 Related Work

2.1 Persuasive Robots and Robotic Moral Influence

Human-Robot Interaction (HRI) researchers have demonstrated that interactive robots, especially language-capable robots, have unique potential to influence humans in a variety of ways [4, 7, 31, 24, 3, 32, 6, 12, 30]. Not only are robots able to exert influence over human behaviors, but also researchers have shown that they can exert moral influence, by weakening humans perceptions of certain moral norms [16]. This moral persuasive influence may be especially strong for language-capable robots, due to the uniquely high social agency [19] and moral agency [9] that may be evoked by natural language capabilities [18]. Critically, this exertion of negative moral influence can be unintentional [16], meaning it needs to be watched for and addressed even in contexts where persuasion is not the robot's perlocutionary goal [20]. Moreover, this potential for negative moral influence is particularly concerning due to previous observations that robots can mediate human–human interaction dynamics [10, 36, 11] and create ripple effects [35, 38, 28] in which robots influence over humans carries over into human–human interactions in which the robot is no longer involved.

Yet with this challenge of avoiding negative moral influence comes an opportunity for promoting positive moral influence and helping cultivate moral ecosystems. Research has shown a variety of ways that robots can engage in moral communication, including rejecting inappropriate commands [5, 17, 14, 21, 39, 38], calling out norm violations [43, 23, 43], justifying necessary norm violations [33], and giving moral advice [34, 26, 25, 40]. Although these types of approaches are often motivated by robots' obligation to avoid performing negative moral actions, or to avoid exerting negative moral influence, all of these activities, especially giving moral advice, may also be used to intentionally exert *positive* moral influence. Critically, just as the risks of unintentional exertion of negative moral influence are exacerbated by the possibility for negative ripple effects, we argue that the benefits of intentional exertion of positive moral influence can be accentuated by the possibility for positive ripple effects. Although positive moral influence could be exerted through a variety of moral communicative means, the most obvious and direct way of doing so may be through the use of moral advice that explicitly conveys particular moral principles.

2.2 Confucian Ethics and Moral Reflection

A number of scholars have begun to incorporate Confucian ethics into the philosophical and empirical studies of human–robot interaction, especially as they relate to calls for increased attention to non-Western ethical theories [42] and for ethical pluralism [47]. Self-reflection is of critical importance to moral development in Confucian ethics. It is also worth noting that moral reflection in the Confucian tradition is rarely done by people themselves. Rather, it is an interactive process in concert with others [44]. Such a relational approach can occur in various settings including: (1) observing and reflecting on how others (especially moral exemplars) make decisions in moral situations and how we can improve ourselves by incorporating our reflective thinking into future situations; and (2) exercising and developing moral sympathy toward others in moral thought experiments [45].

2.3 Confucian Ethics for Robot Moral Communication

As part of the recent effort to integrate Confucian ethics into human–robot interaction, some scholars have recently investigated how robotic moral interventions grounded in different moral frameworks might have different moral effects. [42] explored different ways that Confucian ethics could be used to guide the design of language capable robots. [38] explored the ways that Confucian ethics could guide the design and use of knowledge representations for generating robot norm violation responses. [26] [25] investigated the use of role-based, identity-based, and norm-based language in encouraging participant honesty. And most relevant to our work, [40] compared the use of role-based and norm-based language for encouraging interactants to adhere to community-relevant role norms. In that work, [40] used a Theory of Planned Behavior (TPB) [1] questionnaire to measure potential changes to the strength of these role-norms, and systematically varied the timing of this measure to control for potential ordering effects. Curiously, [40] found that role-based moral interventions led to greater observed adherence to the role-norms under investigation, but *only* when immediately preceded by the TPB questionnaire. [40] suggested that this observation may have been due to the TPB questionnaire serving as an opportunity for moral reflection, the importance of which would be well justified from a Confucian ethical perspective, as described above.

If this suggestion were to be accepted, it would mean that robots' use of role-based moral language is uniquely impactful for encouraging adherence to community-relevant role-norms, but only if the robot's moral language were followed by an opportunity for reflection on that language. We argue, however, that this suggestion cannot be accepted by the results of [40]'s study alone. First, their study was not explicitly designed to interrogate the role of reflection. Second, while the TPB may indeed have served some reflective role in their study, an intentionally designed reflective exercise would be needed to justify [40]'s suggestion. Finally, it is possible that the TPB questionnaire used by [40] may have overly primed people towards role-oriented modes of reflection, which

could explain the localization of their observed effects to their role-based moral language condition. In this work, we thus conducted a conceptual replication of [40]'s experiment that manipulated the opportunity for reflection in a more controlled, explicit, and intentional manner. This work aims to test the two following experimental hypotheses:

Hypothesis H1 When a robotic moral language intervention is followed by an opportunity for moral reflection, it will lead to greater moral influence, as demonstrated by greater adherence to the moral principles encouraged by that moral intervention.

Hypothesis H2 The increases to moral influence facilitated by opportunities for reflection will be greater when following role-based moral language interventions than when following norm-based moral language interventions.

3 Method

3.1 Experimental Design

To evaluate our hypotheses, we conducted an IRB-approved human-subjects experiment with a mixed factorial design, similar to that used by [40]. This experiment used a 2 (Moral language) $\times$2 (Reflection) $\times$2 (Experimental Task), mixed between-within subjects design with two between-subjects factors and one within-subjects factor. Specifically, participants completed two experimental tasks (order counterbalanced) and were randomly assigned to receive either a *Norm-Based* or a *Role-based* moral language intervention immediately after completing the first experimental performance task. Half of participants then completed a moral reflection exercise after receiving their moral language intervention, and the other half did not complete the moral reflection exercise. After receiving the moral language intervention and/or moral reflection exercise, all participants then completed the experimental task for a second time to allow us to compute pre-intervention to post-intervention performance differences.

3.2 Experimental Task

We chose the experimental tasks used by [40] in order to expand and conceptually replicate the prior study's suggestion that a (certain type of) reflective exercise may have increased the efficacy of (a certain type of) moral language provided by a robot. As in [40]'s study, the experimental task asked participants to count the frequency of three articles ("a", "an", and "the") appearing in book pages.

3.3 Experimental Conditions

Moral Language Interventions We also chose to replicate the two moral language interventions from [40]'s study, and then add an opportunity for reflection to address limitations of the original study and the interpretation of its results.

Thus, for the moral intervention conditions, we used the same two videos which consisted of a norm-based moral language intervention delivered by a Nao robot, and a role-based moral language intervention delivered by a Nao robot. All videos of NAO speaking used NAO's default 'voice' and were coupled with closed captioning located at the bottom center of each video. These videos can be found in this paper's OSF Repository. Participants were first introduced to the study and the task with a video of a NAO robot serving as the experimenter in the study. Participants then completed their first article counting task, and once complete, received either a norm-based moral intervention where the Nao robot said:

> *"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."*

Or, participants received a role-based moral intervention:

> *"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."*

These two videos thus represent two possible robotic moral language interventions which use different moral frameworks (role-based vs. norm based). The effectiveness of these moral language interventions are then measured by assessing the difference in task performance before and after receiving the moral language intervention.

Reflection Exercises The claims of Wen [40] regarding the role of moral reflection were made on the basis of the placement of a Theory of Planned Behavior (TPB) questionnaire they used as a dependent measure. Wen's findings suggested that the act of completing the TPB questionnaire immediately after receiving a moral language intervention may have inadvertently served as an opportunity for moral reflection. Although the TPB may have encouraged reflection, it was not explicitly designed for this purpose. Moreover, [40] identified several items within the TPB questionnaire whose wording may have heightened participants' sensitivity to reflect specifically on the role-based moral language.

We thus developed a two-stage reflection exercise for this experiment. In the first stage, participants were asked to think about the language that the robot used in their moral language intervention (i.e., either the norm-based intervention or the role-based intervention), and write down in a free response text box, what they thought about what the robot said. In the second stage, and immediately after completing their responses to the first stage, participants were then asked to indicate how convincing they found the robot's speech. Participants were also asked to explain why they felt that way.

To ensure that participants considered each stage of the exercise carefully, we required a minimum of three minutes to be spent on each stage, and a minimum of 300 characters to be typed for each free response.

3.4 Experimental Procedures

After completing an audio/video check to ensure that participants could see and hear the videos of the NAO robot, and providing informed consent, participants completed a demographic survey. Participants were then shown the video in which a NAO robot introduced itself and explained the experimental task. Next, participants performed the first article counting task, with a video of the robot continuing to play on the left-hand side of the screen as shown in Figure 1. Once participants completed the first article counting task, they were shown a video in which the NAO robot either gave a norm-based language intervention or a role-based language intervention based on their assigned *Intervention* condition. After watching the intervention video, half of participants completed the reflection exercise and then completed the second experimental task. The other half of participants immediately completed the second experimental task with no opportunity for moral reflection. Finally, participants were paid.

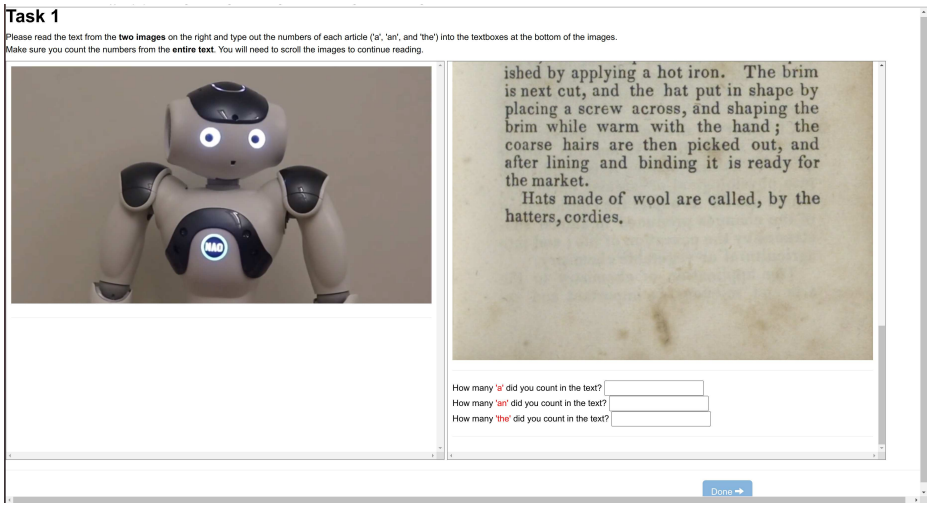

Fig. 1. Screenshot of the experimental task page.

3.5 Dependent Measures

Our key dependent variable in this experiment was *improvement of performance* between article counting experimental task one and article counting task two, which reflected changes in role-norm adherence. Performance was calculated by the difference between the reported counts and the actual counts for all three types of articles. To gain a deeper understanding of how the effectiveness of robots' moral interventions may have been mediated by reflection, we collected

the text participants produced for the reflection exercise. We performed an exploratory content analysis of this text, with respect to: the amount of text participants wrote in their reflection exercises, participants' use of reflection-related verbs and role-related and norm-related words.

3.6 Participants

One hundred and nineteen participants (58F, 60M, 1NB) were recruited from Prolific (www.prolific.co). All participants passed "bot check" procedures. Participants' ages ranged from 19 to 71 years old (M=32.80, SD=10.03).

4 Performance Analysis

4.1 Analysis

We analyzed our data through a Bayesian analysis framework [37], using version 0.16.3 of the JASP statistical software [22]. Within this framework, we conducted a Bayesian Analyses of Variance (ANOVA) with Bayes Factor Analysis to assess (1) the impacts of the moral reflection exercise, (2) the impacts of the type of moral language intervention, and (3) the potential interaction between these two factors. Bayes Factors are odds ratios representing the relative strengths of evidence for and against hypotheses. A Bayes Factor BF_{10} represents the relative likelihood of the collected data under hypothesis H_1 versus hypothesis H_0. We specifically calculated Bayes Inclusion Factors across matched Models [13, 29], which represent, for each candidate main effect and interaction effect, the relative likelihood of models containing that effect versus models not containing that effect, thus providing a measure of the strength of evidence in favor of that effect. Bayes Factors were then interpreted using community standards [27]. All data and analysis scripts can be found in this paper's OSF Repository.

4.2 Results

Our analysis provided very strong evidence in favor of an effect of the reflection exercise (BF 34.783), but moderate evidence against both an effect of intervention type (BF 0.193) and an interaction effect between the reflection exercise and the intervention type (BF 0.255). As shown in Figure 2, participants who received an opportunity for reflection had greater improvement of performance in the second article counting task (M=-3.639, SD=6.718) than did participant who received no such opportunity (M=-0.241, SD=3.461).

5 Content Analysis

We will now discuss the exploratory content analysis that followed our hypothesis-driven statistical analysis.

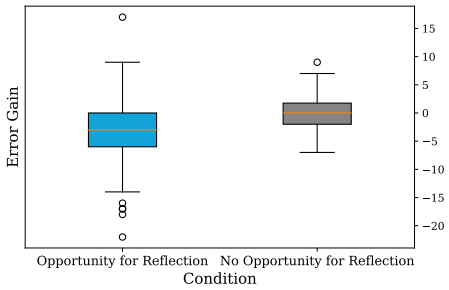

Fig. 2. Error Gain between tasks (errors made post-intervention minus errors made pre-intervention), by experimental condition. Lower numbers indicate better performance on the second task (post-intervention) relative to the first task (pre-intervention).

5.1 Analysis

After screening out one participant whose responses demonstrated a lack of understanding of the task, sixty participants remained who had been completed the reflection exercise. Two authors coded about half of these remaining responses each, looking for keywords that demonstrated either attention to roles, attention to norms, or direct evidence of reflection. After discussion, the two authors agreed on a final set of keywords in each of these three categories.

After defining the norm keywords, role keywords, and reflection verb categories, the two authors each coded a shared set of 20 responses, counting the frequency of words belonging to each of the three categories. We then computed interclass correlation coefficients to assess inter-rater agreement between the two coders, which showed good agreement for all categories (ICCs between 0.8 and 0.89). We thus proceeded to count the frequency of each keyword for each of the three categories for all responses. Analysis on the coded data set was conducted only for participants ($N = 60$) given the opportunity to engage in reflection after the robot's moral intervention. We also computed the total number of words used by each participant in their reflection.

Finally, we conducted a Bayesian repeated measures ANOVA (RM-ANOVA) with Inclusion Bayes Factor Analysis (with moral intervention type as a between-subjects factor and type of vocabulary assessed as a repeated measures factor) to explore differences between norm and role based vocabulary use in each of the two moral interventions, and conducted t-tests with Bayes Factor Analysis to explore the differences in reflection verb use, as well as total number of words typed (as a measure of reflection extensiveness), in each of the two conditions.

5.2 Results

Evidence of General Reflection A t-test revealed anecdotal evidence in favor of an effect of moral intervention type on reflection verb use (BF 1.730) While there was probably no difference in reflection verb use, participants may

have used more reflection verbs after norm-based moral interventions (M=2.000, SD=1.789) than after role-based moral interventions (M=1.172, SD=1.104), A t-test revealed moderate evidence in favor of an effect on reflection as measured by character count (BF 5.907). Participants typed more characters after a norm-based moral intervention (M=375.484, SD=102.355) than after a role-based moral intervention (M=309.414, SD=80.231).

Moral Language Use An RM-ANOVA revealed anecdotal evidence against an effect of robot's moral language intervention type on moral language use in the reflections (BF 0.747) suggesting that there is probably no effect, but participants may have generally used more morally relevant language in their reflection after a role-based moral intervention (M = 1.47,SD = 1.77) than after a norm-based moral intervention (M = 0.97, SD= 1.16). This analysis also revealed moderate evidence against an effect of type of moral language, suggesting that overall, norm-based and role-based moral language were used with relatively equal frequency across reflections (BF 0.297). Finally, this analysis revealed extreme evidence of an interaction between type of robot's moral language intervention and type of moral language use in reflections (BF 348.179), as visualized in Fig. 3.

Post-hoc t-tests provided moderate evidence (BF 3.469) that after the norm-based intervention, norm-based moral language (M=1.323, SD=1.249) was used more frequently than role-based moral language (M=0.613, SD=0.955), and moderate evidence that after the role-based moral intervention, role-based moral language was used even more frequently (M=2.103, SD=2.110) than norm-based moral language (M=0.828, SD=1.037), with very strong evidence (BF 39.871) that role based language was much more strongly encouraged after role-based moral interventions than after norm-based interventions.

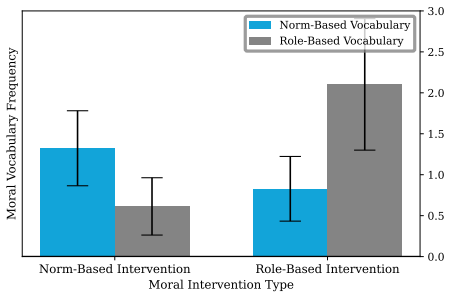

Fig. 3. Vocabulary Use in Moral Reflections by Type of Moral Intervention and Type of Moral Language. Error bars represent 95% Credible Intervals.

These results suggest that although the norm-based intervention may have been more effective at promoting general reflection, both interventions were ef-

fective at promoting their specific types of *moral* reflection, with the role-based intervention especially effective in this regard.

6 Discussion

We contribute to the growing body of research demonstrating the potential for positive moral influence of language capable robots. Specifically, we investigated the positive impact of robots' moral language, as mediated both by different moral frameworks and by the opportunity for reflection. Our work serves as a successful conceptual replication of [40]'s work, validating their suggestion that opportunities for people to reflect on ethics could increase the effectiveness of role-based moral language delivered by a robot, while building on and providing nuance to this suggestion.

A main contribution of this work was to address the limitations of their study, in which questionnaire placement may have inadvertently acted as an opportunity for moral reflection and potentially for only one type of moral language.

Building on [40]'s work, our first hypothesis was that when a robotic moral intervention was followed by an opportunity for moral reflection, it would lead to a greater moral influence, as demonstrated by greater adherence to the moral principles encouraged by that moral intervention. In this study, the adherence to the moral principles was reflected in the improvement in performance of a citizen science task. Our behavioral data support this hypothesis, with participant error rates decreasing only when provided with opportunities for reflection, regardless of which type of moral language was used to encourage their performance improvement. Thus, Hypothesis H1 was supported.

Our second hypothesis was that the increases to moral influence facilitated by reflection would be greater after role-based moral interventions than after norm-based moral interventions. This hypothesis (H2) was not supported. However, our exploratory content analysis provided preliminary insights that role-based moral interventions were slightly more successful at *encouraging moral reflection*, even if this moral reflection did not directly lead to increased norm adherence as we had expected. Together, these findings support the conclusion that role-based interventions can lead to moral reflection, and that reflection can serve as a means to influence human behavior, but that reflection's positive influence on human behavior is sufficiently strong that it is observed regardless of the ethical framework used to guide a robot's moral language.

Moreover, while our results suggest potential benefits of both types of moral language (especially role-based moral language) they also demonstrate a need to broaden consideration beyond the ethical grounding of robots' moral interventions, as the interaction context in which a moral intervention is embedded may be much more important than the nuances of the moral intervention itself. In our case, the structuring of an intervention to allow for moral reflection was more important than the norm-based or role-based grounding of our moral interventions. Future work should explore other ways that the context surrounding a moral intervention might be structured to best support intervention efficacy, and

other ways that reflection exercises can be intentionally structured to facilitate or reduce the invasiveness of reflection.

7 Limitations and Future Work

Our work has several limitations that motivate future work. First, in conditions where participants were given the reflection opportunity, we required participants to reflect for a minimum amount of time and provide a minimum amount of content. It is possible that different types of reflection exercises, or ones people select themselves, would have different effects or different moderating effects on their behavior. Moreover, while we carefully controlled our reflection exercise, it is possible that reflection exercises that are specifically targeted to different moral frameworks could be beneficial, especially if they are shorter or less invasive. Second, this experiment operated on a brief timescale. As one of the tenets of Confucian ethical principles is that cultivating the moral self requires continued practice over time, change in behavior may require repeated moral reflection or repeated interventions over longer time scales. Third, future work should replicate our study in a context with increased ecological validity. This experiment was conducted using video stimuli of robots as a result of the COVID-19 pandemic [8]. In-person experiments would allow for increased ecological validity and richer qualitative analysis. Finally, just as our work helped to provide more formal basis for [40]'s suggestion through rigorous conceptual replication, so too should the results of our exploratory content analysis be replicated.

8 Conclusion

Robots stand to wield significant positive and pro-social impact through their unique capability for positive moral persuasion. By doing so, robots might help interactants to cultivate their moral selves, and moreover, might cause positive ripple effects that positively effect interactants' broader moral ecosystems. In this work, we explored the ways that robots' moral interventions – and moreover, the contexts into which they are embedded – can be structured to best wield this positive persuasive power. To do so, we conceptually replicated [40]'s prior work, justifying their intuitions that providing opportunities for moral reflection on robot-delivered moral language could be the key to unlocking robots' persuasive capabilities when giving moral advice. Moreover, our work simultaneously sheds light on the unique benefits of role-based moral interventions, while also encouraging HRI researchers to move beyond specific choices of phrasing and focus more attention on the interaction structures that will support such interventions. Our work thus provides substantial nuance to the understanding of this research landscape that was enabled by prior work [40], while opening up promising new directions to further explore that landscape in future work.

Acknowledgements 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. Ajzen, I.: The theory of planned behavior. Organizational behavior and human decision processes **50**(2), 179–211 (1991)
2. Ames, R.T.: Confucian role ethics: A vocabulary (2011)
3. Baroni, I., Nalin, M., Zelati, M.C., Oleari, E., Sanna, A.: Designing motivational robot: how robots might motivate children to eat fruits and vegetables. In: Int'l Symp. Robot and Human Interactive Communication (2014)
4. Bartneck, C., Bleeker, T., Bun, J., Fens, P., Riet, L.: The influence of robot anthropomorphism on the feelings of embarrassment when interacting with robots. Paladyn, Journal of Behavioral Robotics **1**(2), 109–115 (2010)
5. Briggs, G., Williams, T., Jackson, R.B., Scheutz, M.: Why and how robots should say 'no'. International Journal of Social Robotics pp. 1–17 (2021)
6. Chidambaram, V., Chiang, Y.H., Mutlu, B.: Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues. In: International conference on Human-Robot Interaction (HRI). ACM (2012)
7. Cormier, D., Newman, G., Nakane, M., Young, J.E., Durocher, S.: Would you do as a robot commands? an obedience study for human-robot interaction. In: International Conference on Human-Agent Interaction (2013)
8. Feil-Seifer, D., Haring, K.S., Rossi, S., Wagner, A.R., Williams, T.: Where to next? the impact of covid-19 on human-robot interaction research (2020)
9. Floridi, L., Sanders, J.W.: On the morality of artificial agents. Minds and machines **14**(3), 349–379 (2004)
10. Gillet, S., van den Bos, W., Leite, I.: A social robot mediator to foster collaboration and inclusion among children. In: Robotics: Science and Systems (2020)
11. Gillet, S., Cumbal, R., Pereira, A., Lopes, J., Engwall, O., Leite, I.: Robot gaze can mediate participation imbalance in groups with different skill levels. In: Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. pp. 303–311 (2021)
12. Ham, J., Bokhorst, R., Cuijpers, R., van der Pol, D., Cabibihan, J.J.: 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. pp. 71–83. Springer (2011)
13. Hinne, M., Gronau, Q.F., van den Bergh, D., Wagenmakers, E.J.: A conceptual introduction to bayesian model averaging. Advances in Methods and Practices in Psychological Science **3**(2), 200–215 (2020)
14. Jackson, R.B., Wen, R., Williams, T.: Tact in noncompliance: The need for pragmatically apt responses to unethical commands. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. pp. 499–505 (2019)
15. Jackson, R.B., Williams, T.: Robot: Asker of questions and changer of norms? Proceedings of ICRES (2018)
16. Jackson, R.B., Williams, T.: Language-capable robots may inadvertently weaken human moral norms. In: Companion of the 14th ACM/IEEE International Conference on Human-Robot Interaction (alt.HRI). pp. 401–410. IEEE (2019)
17. Jackson, R.B., Williams, T.: Language-capable robots may inadvertently weaken human moral norms. In: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI). pp. 401–410. IEEE (2019)
18. Jackson, R.B., Williams, T.: On perceived social and moral agency in natural language capable robots. In: 2019 HRI workshop on the dark side of human-robot interaction. Jackson, RB, and Williams. pp. 401–410 (2019)

19. Jackson, R.B., Williams, T.: A theory of social agency for human-robot interaction. Frontiers in Robotics and AI p. 267 (2021)
20. Jackson, R.B., Williams, T.: Enabling morally sensitive robotic clarification requests. ACM Transactions on Human-Robot Interaction (THRI) **11**(2), 1–18 (2022)
21. Jackson, R.B., Williams, T., Smith, N.: Exploring the role of gender in perceptions of robotic noncompliance. In: Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. pp. 559–567 (2020)
22. JASP Team, et al.: Jasp. Version 0.8. 0.0. software (2016)
23. Jung, M.F., Martelaro, N., Hinds, P.J.: 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. pp. 229–236 (2015)
24. Kennedy, J., Baxter, P., Belpaeme, T.: Children comply with a robot's indirect requests. In: Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction (HRI). pp. 198–199 (2014)
25. Kim, B., Wen, R., de Visser, E.J., Zhu, Q., Williams, T., Phillips, E.: Investigating robot moral advice to deter cheating behavior. In: TSAR Workshop at ROMAN 2021 (2021)
26. Kim, B., Wen, R., Zhu, Q., Williams, T., Phillips, E.: Robots as moral advisors: The effects of deontological, virtue, and confucian role ethics on encouraging honest behavior. In: Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. pp. 10–18 (2021)
27. Lee, M.D., Wagenmakers, E.J.: Bayesian cognitive modeling: A practical course. Cambridge university press (2014)
28. Lee, M.K., Kiesler, S., Forlizzi, J., Rybski, P.: 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. pp. 695–704 (2012)
29. Mathôt, S.: Bayes like a baws: Interpreting bayesian repeated measures in jasp. Cognitive Science and more. Retrieved from: https://www. cogsci. nl/blog/interpreting-bayesian-repeated-measures-in-jasp (2017)
30. Paradeda, R.B., Ferreira, M.J., Dias, J., Paiva, A.: 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. pp. 251–252. ACM (2017)
31. Rea, D.J., Geiskkovitch, D., Young, J.E.: 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. pp. 21–29 (2017)
32. Robinette, P., Li, W., Allen, R., Howard, A.M., Wagner, A.R.: Overtrust of robots in emergency evacuation scenarios. In: The Eleventh ACM/IEEE International Conference on Human Robot Interaction. pp. 101–108 (2016)
33. Scheutz, M., Malle, B.F.: May machines take lives to save lives? human perceptions of autonomous robots (with the capacity to kill). Lethal autonomous weapons: Re-examining the law and ethics of robotic warfare (2021)
34. Strait, M., Canning, C., Scheutz, M.: 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) (2014)
35. Strohkorb Sebo, S., Traeger, M., Jung, M., Scassellati, B.: 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. pp. 178–186 (2018)
36. Tennent, H., Shen, S., Jung, M.: Micbot: A peripheral robotic object to shape conversational dynamics and team performance. In: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI). pp. 133–142. IEEE (2019)
37. Wagenmakers, E.J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q.F., Šmíra, M., Epskamp, S., et al.: Bayesian inference for psychology. part i: Theoretical advantages and practical ramifications. Psychonomic bulletin & review **25**(1), 35–57 (2018)
38. Wen, R., Han, Z., Williams, T.: Teacher, teammate, subordinate, friend: Generating norm violation responses grounded in role-based relational norms. In: HRI. pp. 353–362 (2022)
39. Wen, R., Jackson, R.B., Williams, T., Zhu, Q.: Towards a role ethics approach to command rejection. In: HRI Workshop on the Dark Side of Human-Robot Interaction (2019)
40. Wen, R., Kim, B., Phillips, E., Zhu, Q., Williams, T.: Comparing norm-based and role-based strategies for robot communication of role-grounded moral norms. ACM Transactions on Human-Robot Interaction (T-HRI) (2022)
41. Williams, T., Jackson, R.B., Lockshin, J.: A bayesian analysis of moral norm malleability during clarification dialogues. In: Proceedings of the Annual Meeting of the Cognitive Science Society (COGSCI). Cognitive Science Society, Madison, WI (2018)
42. Williams, T., Zhu, Q., Wen, R., de Visser, E.J.: The confucian matador: Three defenses against the mechanical bull. In: Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (alt.HRI). pp. 25–33 (2020)
43. Winkle, K., Melsión, G.I., McMillan, D., Leite, I.: Boosting robot credibility and challenging gender norms in responding to abusive behaviour: A case for feminist robots. In: Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. pp. 29–37 (2021)
44. Wong, D.B.: Cultivating the self in concert with others. In: Dao companion to the Analects, pp. 171–197. Springer (2014)
45. Zhu, Q.: Confucian moral imagination and ethics education in engineering. Frontiers of Philosophy in China **15**(1), 36–52 (2020)
46. Zhu, Q., Williams, T., Jackson, B., Wen, R.: Blame-laden moral rebukes and the morally competent robot: A confucian ethical perspective. Science and Engineering Ethics pp. 1–16 (2020)
47. Zhu, Q., Williams, T., Wen, R.: Role-based morality, ethical pluralism, and morally capable robots. Journal of Contemporary Eastern Asia **20**(1), 134–150 (2021)