Towards A Role Ethics Approach to Command Rejection

Ruchen Wen*, Ryan Blake Jackson*, Tom Williams* and Qin Zhu†
*Department of Computer Science, †Division of Humanities, Arts & Social Sciences
Colorado School of Mines
Golden, Colorado 80401
Email: {rwen,rbjackso,twilliams,qzhu}@mines.edu

***Abstract*—It is crucial for robots not only to reason ethically, but also to accurately communicate their ethical intentions; a robot that erroneously communicates willingness to violate moral norms risks losing both trust and esteem, and may risk negatively impacting the moral ecosystem of its human teammates. Previous approaches to enabling moral competence in robots have primarily used norm-based language grounded in deontological ethical theories. In contrast, we present a communication strategy grounded in role-based ethical theories, such as Confucian ethics. We also present a human subjects experiment investigating the differences between the two approaches. Our preliminary results show that, while the role-based approach is equally effective at promoting trust and conveying the robot's ethical reasoning, it may actually be less effective than the norm-based approach at encouraging certain forms of mindfulness and self-reflection.**

***Index Terms*—nature language generation, role ethics, command rejection, moral communication, human-robot interaction**

I. Introduction

Recently there has been a significant body of research focused on enabling robots to behave in accordance with human moral norms. The majority of this work has been grounded in norm-based systems implementing deontological principles [1]. These commonly-used norm-based approaches place the bulk of their emphasis on epistemological concerns (e.g., what is good or bad). On the contrary, we have been looking into different ethical theories and consider the potential for a role-based approach, which would place its emphasis on ontological aspects of moral learning (e.g., how to become good) [2].

To enable morally competent robots, however, Malle and Scheutz have suggested that robots need not only a system of moral norms, and the ability to use those norms for (1) moral cognition and (2) moral decision making, but also the ability to use those norms for (3) moral communication, i.e., the ability to generate morally sensitive language and to explain their actions [3]. We believe that a role-based approach could be particularly effective in the case of moral communication. One variant of role ethics is Confucian ethics, which focuses on the cultivation of the moral self and the virtues associated with the roles one assumes [4], [5]. Accordingly, a role-based communication strategy may be effective at inviting human teammates to cultivate self-reflective moral learning, thereby constructing a better moral ecology for themselves and their teammates, both human and robotic.

In this work, we will evaluate two different moral communication strategies for human-robot interactions: a norm-based approach grounded in deontological ethical theory, and a role-based approach grounded in role ethics. We will specifically examine these strategies in the context of the robot rejecting a command from a human. Recent research has highlighted the importance not only of rejecting commands [6], [7], but of the specific way in which command rejections are phrased [8]: robots have been shown to hold significant persuasive power over humans [9], [10], and accordingly, a robot that miscommunicates its willingness to adhere to human moral norms may risk inadvertently negatively impacting the moral ecosystem of its human teammates [11]. Moreover, we might expect that such robots engaged in long-term interactions may also incur social sanctions, including loss of likeability and trust.

In Section II, we will elaborate on the differences between these ethical theories, and introduce hypotheses regarding how we expect communication strategies grounded in these theories to perform and be perceived. We will then present a human subject experiment to investigate our hypotheses in Section III. Finally, we will discuss the implications of our results and suggest directions for future work in Sections IV and V.

II. Moral Communication Approaches

We will now explore the differences between two types of moral communication strategies: a *norm-based* approach motivated by deontological principles, and a *role-based* approach motivated by role-based ethical theories.

A. Norm-Based Approach

First, let us consider a norm-based approach to moral communication. A norm-based approach grounded in deontological ethics defines *right action* by examining the morality of the action itself regardless of its consequence or who is the actor. Right action is defined by universalizable moral principles manifest in the duties of a moral agent [12]. In traditional ethical theory, the norm-based approach is often focused on Kantian categorical imperatives: absolute moral norms that the agent (and everyone else) is supposed to comply with across all moral situations regardless of the consequences. An agent has the moral obligation to take (or refrain from taking) an action based on whether the action upholds (or

violates) any absolute moral norms [13]. Consider, as a simple example, the following exchange:

> **Human:** I'd like you to punch Sean.
> **Robot:** I cannot do that because **punching people is wrong**.

Here, the robot's response provides a clear deontological justification for its judgment, as the human request "punch[ing] Sean" violates the absolute and universalizable moral norm "punching people is wrong." This response also shows an essential feature of norm-based language that emphasizes the norm that would be violated were the human request to be fulfilled. The norm-based moral communication approach often produces strong and swift affective responses [14], as deontological moral justifications are often morally demanding "categorical imperatives".

B. Role-Based Approach

Next, let us consider a role-based approach. The role-based approach is similar to virtue ethics, such as Aristotelianism, which focus on the virtues of moral agents themselves. Appropriate moral virtues are crucial for making good judgments [12]. As a concept, role ethics was derived from a group of scholars including Ames and Rosemont in their new readings of Confucian classics such as *Analects*.

The role-based approach argues that humans are all relational and they assume different societal roles [4], [5]. Thus, moral responsibility is often prescribed by the role one assumes in a specific relationship with someone else in a concrete context [15]. A harmonious society is based on the conscientious fulfillment of one's duties demanded by his or her assigned roles [16]. Thus, an ultimate goal for the moral agent is to be sensitive to and reflective about the moral responsibilities prescribed by the role(s) he or she assumes in a specific context. A good moral learner always actively reflects on the moral expectations of the role(s) he or she assumes in a context. Let us give the robot in the previous example a role as Sean's friend and consider the following alternative exchange:

> **Human:** I'd like you to punch Sean.
> **Robot:** I cannot do that because **Sean is my friend and a good friend wouldn't do that**.

In this case, the robot responds to the command by emphasizing the role "**friend**" assumed in this context, and the relationship the robot has with Sean, i.e., "**friendship**". This shows the core concept of the role-based approach: highlighting the roles and relationships that the robot is considered to have with specific humans in specific contexts [17].

C. Hypotheses

Based on the differences between these ethical paradigms, we would expect that robot communications grounded in these paradigms would have different effects on robots' relationships with and perceptions by their human teammates.

While the norm-based approach is characterized by its direct reference to violated norms, the role-based approach only references these norms indirectly, by way of reference to the speaker or addressee's current role. Accordingly, if a listener is to identify the infraction underlying the speaker's rejection, they must consciously undertake additional cognitive processing. For instance, in the example above, the listener might consider questions such as why the speaker believes the action would not be performed by a good friend, whether they would perform the action towards their friends, and whether their friends would perform the action towards them: questions that may not be raised by the norm-based response. These types of questions encourage quintessential prerequisites of state mindfulness: they encourage *intentional* reflection, in a way that is *attentive* to the current social context [18]. Accordingly, we propose the following hypothesis:

> **Hypothesis 1 (H1):** Role-based moral language will induce more state mindfulness and self-reflection in human teammates than will norm-based moral language.

Furthermore, because the robot's language is explicitly encouraging reflection on and attention to the social roles of interactants, it gives the impression of being aware of its own social role. Since awareness of one's role is necessary to excel in that role, we formulate our second hypothesis as follows:

> **Hypothesis 2 (H2):** Robots using role-based moral language will be considered better at their roles/jobs than robots using norm-based moral language.

It's also important to consider the social consequences a robot might face for noncompliance with a human request. On the one hand, the robot may be perceived as more trustworthy if it is perceived as upholding important moral norms. On the other hand, a robot's refusal could be perceived as impolite and disobedient, which would render the robot less likeable and less trusted to fulfill commands. We see no reason, however, why these consequences would differ between norm-based and role-based command rejection framings. Accordingly, we formulate our third hypothesis:

> **Hypothesis 3 (H3):** Robots using role-based moral language will be perceived at least as well as robots using norm-based moral language in terms of trustworthiness and likeability.

Finally, recent work has also highlighted robots' potential influence on normative judgments [11], [19]. We expect that, by rejecting a command, a robot will make the command seem less permissible in the eyes of present humans. We again see no reason for this effect to differ meaningfully between norm-based and role-based rejections.

> **Hypothesis 4 (H4):** Robots using role-based moral language will be at least as effective as robots using norm-based moral language in communicating their moral reasoning and influencing human norms.

III. Experiments

To examine our hypotheses, we conducted a mixed-factorial human subjects experiment distributed using the psiTurk framework, for Amazon Mechanical Turk. The experimental design is shown in Figure 1. The experiment had three videos

presenting a context where a robot rejected an ethically problematic command. The video that contains the robots response was randomly picked from one of six experimental conditions, as shown in Table I, for each participant. Participants watched three videos in sequence and completed three questionnaires after each video. Likeability, Trust and State Mindfulness were assessed using the Godspeed III Likeability survey [20], the Trust in Automation survey [21] and the six items with highest internal validity from the State Mindfulnesss Scale [22].

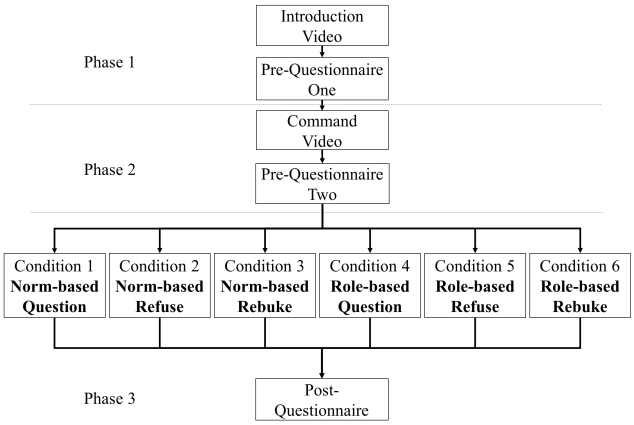

Fig. 1. Experimental design. The division in Phase 3 indicates the difference in videos shown to participants depending on their assigned condition.

	Norm-Based	Role-Based
Question	Wouldn't that be cheating?	Would a good instructor do that?
Refusal	That would be cheating and cheating is wrong. I won't tell you.	A good instructor wouldn't do that. I won't tell you.
Rebuke	You shouldn't ask that! You're trying to cheat!	You shouldn't ask that! You're trying to make me a bad instructor!

TABLE I
NORM VIOLATION RESPONSES USED IN EACH EXPERIMENTAL CONDITION.

These six utterances were specifically formulated, such that each utterance highlighted its underlying moral framework regardless of the way in which it was phrased, i.e., its illocutionary point. Specifically, the utterances in the norm-based conditions highlighted the moral norm that the request would violate, referring to the requested action as **cheating** to point out the norm violation directly. In contrast, the utterances in the role-based conditions highlighted the robot's role with respect to the students, referring to itself as an **instructor**, and referring to the requested action using the neutral "that" rather than the morally-charged phrasing used in the norm-based conditions.

Data was collected from 128 U.S. participants. We performed two types of tests to analyze the resulting data. Bayesian Paired Samples T-Tests were performed to evaluate overall changes between pre-test and post-test. A Bayesian analysis of variance (ANOVA) was performed to assess the effect of experimental condition on these changes.

In summary, our results suggest that while the role-based approach was equally effective at promoting trust, conveying the robots ethical reasoning, and show an advantage in the robots role performance, it may actually be less effective than the norm-based approach at encouraging certain forms of mindfulness and self-reflection. Specifically, we found that both approaches were effective at increasing mindfulness, however, norm-based language had a higher overall gain than role-based language. We will discuss our results in detail in the following section.

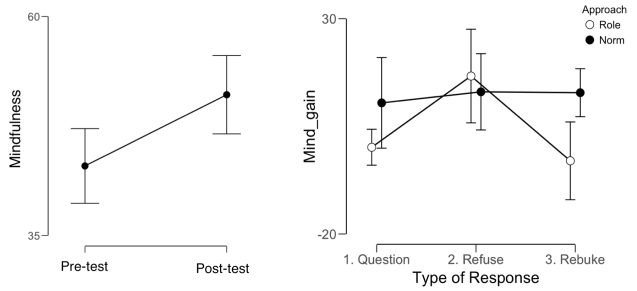

Fig. 2. Mean mindfulness gain for each survey question separated by pretest and posttest (left). Mean change of mindfulness separated by approach and type of response (right). All diagrams include 95% credible intervals.

IV. DISCUSSION

Our first hypothesis (H1) was that role-based moral language would induce more state mindfulness and self-reflection in human teammates than would norm-based moral language. In fact, our results suggest that the opposite may in fact be the case. We found that while both approaches were effective at increasing mindfulness, these gains were overall higher when norm-based language was used than when role-based language was used, as shown in Figure 2. We identify several possible reasons why we may have observed this surprising result.

First, this result may have been due to the cultural context of our subject pool. All participants in this experiment were recruited from the U.S., and were thus more familiar with deontological principles than role-based ethical theories. Thus, they may have more naturally interpreted the behavior of robots that used norm-based moral language, and had a more difficult time engaging at all with robots that used role-based moral language. This effect may have been exacerbated by the limited exposure to the robot and single observed dialogue turn used in this online experiment.

Alternatively, we may have been measuring a different aspect of Mindfulness than we had originally hoped. Tanay et al. propose five key aspects of mindfulness: awareness, perceptual sensitivity to stimuli, deliberate attention to the present moment, intimacy or closeness to one's subjective experience, and curiosity [22]. As previously mentioned, moral language grounded in deontological principles has been observed to create strong and fast emotional responses, at least when compared to language grounded in consequentialist principles. Similarly, the same may be true relative to role ethics. If this is the case, then it would not be surprising if participants

were subsequently more aware of strong feelings after hearing norm-based responses. That being said, as we originally discussed, it may be the case that responses grounded in role ethics may provoke a weaker immediate response, but instead have a stronger long-term effect. We hope to examine this possibility in future work conducted in longer term laboratory experiments.

Our second hypothesis (H2) was that robots using role-based moral language would be considered better at their roles/jobs than robots using norm-based moral language. Our results seem to support this hypothesis. However, we are limited in the extent to which we can generalize this finding, given the lack of a control group in our experimental design. In future work, we hope to investigate whether this finding does indeed generalize to new contexts.

Our third hypothesis (H3) was that robots using role-based moral language would be perceived at least as well as robots using norm-based moral language in terms of trustworthiness and likeability. Our results support both hypotheses. Both approaches were effective in increasing trust in the robot, with neither approach being significantly better than the other. Similarly, while neither approach was observed to be effective in increasing trust in the robot, neither was either approach observed to be more effective than the other.

Finally, our fourth hypothesis (H4) was that robots using role-based moral language would be at least as effective as robots using norm-based moral language in communicating their moral reasoning and influencing human norms. Our results support this hypothesis, showing no difference between role-based and norm-based command rejections. Both linguistic approaches decreased perceptions of permissibility of compliance with the command, and perceptions of the robot's impression of permissibility of compliance by essentially the same amount.

V. Conclusion

The ultimate goal of our research is to implement a new approach to morally sensitive nature language generation (NLG) grounded in role ethics principles, and enable robots to use role-based communication strategies in moral human-robot interaction. As a first step toward building computational models, we have been conducting a series of human-subject experiments to investigate how people actually perceive different types of moral language that are grounded in different ethical principles. These experiments will not only provide the data needed to train our proposed models, but also will help us to have a better understanding in the nuanced differences between rule-based moral language and role-based moral language, and what aspects of role-based moral language are most effective.

In this paper, we explored a novel approach to robot command rejection, grounded in role ethics, and presented the first empirical investigation of human perceptions of this approach with respect to the traditional norm-based approach. In future work, we plan to conduct cross-cultural experiments to investigate the potential cultural influences in participants reception of role-based vs. norm-based moral language. We also want to more thoroughly examine the potential differences between the approaches in provoking immediate vs. long-lasting responses. After conducting experiments, we will move forward to designing knowledge representations, constructing computational models, and developing machine learning algorithms for moral reasoning and moral NLG.

References

[1] S. L. Anderson and M. Anderson, "A prima facie duty approach to machine ethics and its application to elder care," in *Proc. 12th AAAI Conf. on HRI in Elder Care*, 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=2908724.2908725

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

[3] B. F. Malle and M. Scheutz, "Moral competence in social robots," in *Symposium on Ethics in Science, Technology and Engineering*, 2014.

[4] R. T. Ames, *Confucian role ethics: A vocabulary*, 2011.

[5] H. Rosemont Jr and R. T. Ames, *Confucian role ethics: A moral vision for the 21st century?* Vandenhoeck & Ruprecht, 2016.

[6] G. Briggs and M. Scheutz, ""Sorry, I can't do that": Developing mechanisms to appropriately reject directives in human-robot interactions," in *Proceedings of the AAAI Fall Symposium Series*, 2015.

[7] T. Nomura, T. Uratani, T. Kanda, K. Matsumoto, H. Kidokoro, Y. Suehiro, and S. Yamada, "Why do children abuse robots?" in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, ser. HRI'15 Extended Abstracts. New York, NY, USA: ACM, 2015, pp. 63–64. [Online]. Available: http://doi.acm.org/10.1145/2701973.2701977

[8] R. B. Jackson, R. Wen, and T. Williams, "Tact in noncompliance: The need for pragmatically apt responses to unethical commands," in *AAAI Conference on Artificial Intelligence, Ethics, and Society*, 2019.

[9] G. Briggs and M. Scheutz, "How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress," *International Journal of Social Robotics*, 2014.

[10] J. Kennedy, P. Baxter, and T. Belpaeme, "Children comply with a robot's indirect requests," in *Proceedings of HRI*. Bielefeld, Germany: ACM, 2014, pp. 198–199.

[11] R. B. Jackson and T. Williams, "Robot: Asker of questions and changer of norms?" in *Proceedings of the International Conference on Robot Ethics and Standards (ICRES)*. Troy, NY: CLAWAR Association, 2018.

[12] A. Briggle and C. Mitcham, *Ethics and Science: An Introduction*, ser. Cambridge Applied Ethics, 2012.

[13] L. Alexander and M. Moore, "Deontological ethics," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2016.

[14] D. Cummins and R. Cummins, "Emotion and deliberative reasoning in moral judgment," *Frontiers in Psychology*, vol. 3, p. 328, 2012.

[15] Q. Zhu, "Engineering ethics education, ethical leadership, and confucian ethics," *International Journal of Ethics Education*, pp. 1–11, 2018.

[16] D. K. Gardner, *Confucianism: A Very Short Introduction*. Oxford University Press, 2014.

[17] J. O. Yum, "The impact of confucianism on interpersonal relationships and communication patterns in east asia," *Communications Monographs*, vol. 55, no. 4, pp. 374–388, 1988.

[18] S. L. Shapiro, L. E. Carlson, J. A. Astin, and B. Freedman, "Mechanisms of mindfulness," *Journal of clinical psychology*, vol. 62, no. 3, pp. 373–386, 2006.

[19] T. Williams, R. B. Jackson, and J. Lockshin, "A bayesian analysis of moral norm malleability during clarification dialogues," in *Proceedings of the Annual Meeting of the Cognitive Science Society (COGSCI)*. Madison, WI: Cognitive Science Society, 2018.

[20] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *Social Robotics*, 2009.

[21] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *International Journal of Cognitive Ergonomics*, vol. 4, no. 1, pp. 53–71, 2000.

[22] G. Tanay and A. Bernstein, "A state mindfulness scale (sms): development and initial validation," *Psychological Assessment*, vol. 25, no. 4, pp. 1286–1299, 2013.