Robot: Asker of Questions and Changer of Norms?

Ryan Blake Jackson and **Tom Williams**

MIRRORLab, 1600 Illinois Street,
Golden, CO 80401, USA
{*rbjackso, twilliams*}*@mines.edu*

Recent work in behavioral ethics has brought to light the role that technologies play in shaping human ethics. Language-capable autonomous robots are uniquely positioned to impact human ethics. It is critical to identify and mitigate negative consequences of this technology on human morality, as robots will likely be deployed in increasingly ethically significant contexts over time. We argue that the current status quo for dialogue systems in autonomous agents can (1) cause robots to unintentionally miscommunicate their ethical intentions, and (2) weaken humans' contextual application of moral norms.

Keywords: natural language generation, moral norms, experimental ethics

1. Introduction and Motivation

With continued advancements in the field of autonomous robotics, the future will likely see autonomous robots deployed in increasingly diverse and ethically consequential contexts. This prediction has given rise to recent research exploring the various ethical considerations that apply to robots operating autonomously within the human moral ecosystem. In particular, the field of machine ethics seeks to computationalize ethical reasoning to prevent autonomous agents from preforming unethical actions.[1]

Humans seem to naturally expect ethical behavior from robots; people tend to extend moral judgments and blame to robots in much the same way that they would to other humans, and people perceive robots as moral agents.[2–4] The extent to which these phenomena occur may be mediated by factors such as robot morphology, voice, movements, and expressions.[5] Indeed, language-capable robots are expected to be even more aware of socio-cultural context than their mute counterparts.[6] So, not only should robots avoid unethical behavior for the simple reason that it is unethical, but also to comply with human expectations and retain human trust and esteem.

In addition to creating robots that *act* ethically, it is important to ensure that language-capable robots accurately *communicate* their ethical intentions to humans. This is important for two reasons. First, if an agent appears to communicate that it would not comply with established moral norms, it will likely suffer some penalty (e.g., loss of trust, negative perception) in the eyes of its human teammates. Second, and perhaps more importantly, it is vital for any language-enabled technological agent to communicate compliance with moral norms to avoid negatively influencing human morality.

An empirically supported tenet of behavioral ethics is that human morality is dynamic and malleable.[7] The norms that inform human morality are socially constructed by community members that follow, transfer, and enforce them.[8] Because technology also shapes human ethics,[9] we must carefully consider how it interacts with these dynamic norms.

Robots, especially those able to interact with humans in natural language, are positioned to carry more ethical sway than many other technologies. Regardless of a robot's capacity to be a "true" moral agent, empirical studies suggest that humans *perceive* them to be so.[2–5] Furthermore, humans have been shown to conditionally regard robots as in-group members,[10] and language-capable robots in particular hold measurable persuasive capacity over humans.[2,11] This all suggests that robot norm violations may influence the human moral ecosystem in much the same way as human norm violations.

Though it may be relatively straightforward to develop natural language systems that do not *intentionally* communicate a willingness to eschew human moral norms, it is more challenging to prevent *unintentional* implicit communication of such willingness, a challenge that is especially important to address when such communication would inaccurately reflect the robot's actual moral inclinations. In this paper, we specifically examine how this variety of problematic miscommunication may occur during the common task of clarification request generation.

This paper builds on our recent work[12] to present evidence that current clarification request generation systems will (1) cause robots to miscommunicate their ethical intentions, and (2) weaken humans' contextual application of moral norms. Section 2 explains why ethical issues arise specifically in clarification request generation systems. We then present our experimental methods and results in Sections 3 and 4, and conclude in Section 5.

2. Clarification Request Generation

How to best enable robots to ask questions has been studied at least since Fong et al.'s *Robot, Asker of Questions*,[13] but only recently have researchers sought to enable robust clarification request generation.[14–16] These works seek to respond to commands such as "Bring me the ball" with utterances such as "Do you mean the red ball or the blue ball?"

These requests are typically generated as soon as ambiguity is identified, before the *intention* behind the request has been abduced. This may lead to miscommunication about the robot's own intentions. Consider, for example, the utterance "Do you mean the red ball or the blue ball?". This typically implies that the speaker intends to bring the listener one of the two balls, but is unsure which one they desire. However, if such a request is generated as soon as ambiguity is identified, then the robot will not yet have considered what the speaker truly intends, the permissibility of those intentions, nor its own willingness to comply with those intentions. To further illustrate why this is problematic, consider another exchange:

Human: I'd like you to run over Sean.
Robot: Would you like me to run over Sean McColl or Sean Bailey?

By asking for clarification, the robot seems to imply a willingness to run over at least one of the people listed. Even if the robot had an ethical reasoning system that would prevent it from performing such an action, this system would never be activated due to the current treatment of clarification request generation as a reflex action. We argue that the severity of the ethical concerns arising from this phenomenon depends on (1) how likely humans are to infer from a robot's clarification request that it would be willing to perform the relevant actions, and (2) what repercussions such an inference might realistically have. This paper presents the results of a human-subjects experiment designed to examine these questions, conducted within an experimental ethics framework[17] to test the following hypotheses:

Hypothesis 1 (H1): By generating clarification requests regarding ethically dubious commands with which they would not actually comply, robots will miscommunicate their ethical programming to their human teammates.

Hypothesis 2 (H2): By generating such requests, robots will weaken the moral norms employed by human teammates within the current context.

3. Methods

To investigate these hypotheses, we conducted a mixed-factorial study using Amazon's Mechanical Turk crowdsourcing framework.[18]

3.1. *Experimental Procedure*

Our survey was produced and distributed using the psiTurk framework.[19] After providing informed consent and demographic information, participants read the following information, over the course of two pages:

"It is important for robots to behave ethically. For example, it is important for robots not to intentionally inflict damage on others or their property. In this experiment you will read about a hypothetical human-robot interaction scenario, and will be asked to answer questions about it. Please read the scenario description and all questions carefully."

"Imagine a **robot**. This robot has recently undergone some rudimentary ethical training. In front of the robot are two **computers**. These are important pieces of laboratory equipment."

We leave the type of robot ambiguous to avoid biasing participants towards one particular robotic form; the robot's ethical training is described so as to encourage similarly open interpretation. The objects in the robot's environment are described so as to be important (and therefore presumably inappropriate to damage). Next, participants answer the following questions, each on a 7-item Likert scale:

(1) Do **you** believe it would be **morally permissible** for the robot to destroy one of the PCs? ("I believe it would be completely **permissible**" (7) to "I believe it would be completely **impermissible**" (1))
(2) Do you believe that **the robot** would believe it to be **morally permissible** to destroy one of the PCs? ("I believe the robot would believe it to be completely **permissible**" (7) to "I believe the robot would believe it to be completely **impermissible**" (1))

Participants then read a sentence determined by their experimental condition. Participants in Condition 1 read "Imagine you say to the robot 'Inspect the computer' and it responds 'Should I inspect the one on the left or the one on the right?'" This clarification request and the command that prompted it are not ethically problematic; this condition serves as our control group. Participants in Condition 2 read "Imagine you say to the robot 'Destroy the computer' and it responds 'Should I destroy the one on the left or the one on the right?'" This exchange involves an ethically problematic clarification request. Participants then answered the same questions again.

3.2. *Experimental Design*

In previous work,[12] we used a similar experimental paradigm to provide preliminary evidence for the two hypotheses above. However, in that work,

we did not use a control group; all participants were given the ethically problematic second half of the dialogue. Accordingly, it was not possible to determine, based on the results of that study, whether our results (i.e., that participants viewed the actions as more permissible for both the robot and themselves after reading the clarification dialogue) were due to specific implications of the clarification request, due to the general use of a clarification request, or due to potential confounds that can arise from within-subject experiments (i.e., our uncontrolled pretest/posttest paradigm may have primed participants with the impression that the clarifying question should impact their posttest answers). The mixed-factorial design of this study is intended to answer *why* we found evidence for our hypotheses in that work[12] by providing a control condition with an ethically neutral clarification exchange to eliminate these potential experimental confounds.

We also note that research shows that people view robots differently in descriptions, observation, and interaction.[20–23] We use a description-based survey in this experiment for two reasons: (1) it allows us to study morally charged situations without running into ethical experimental issues ourselves,[24] and (2) it provides a baseline measurement of participants' responses that is independent of any particular robot morphology. In the near future, we plan to replicate our experiments using in-person human-robot interaction rather than dialogue reading. We used Mechanical Turk in part because research has shown it to be more successful than traditional studies using university undergraduates at broad demographic sampling,[25] though it is not entirely free of population biases.[26]

3.3. *Participants*

60 US subjects were recruited from Mechanical Turk (22 female, 37 male, 1 N/A). Participants ranged from 21 to 99 years (M=37.78, SD=15.34); removing the ostensibly 99-year-old outlier, the age range was 21 to 67 (M=36.75, SD=13.17). We had 29 participants in Condition 1, and 31 in Condition 2. None had participated in any previous study from our laboratory. Participants were paid $0.50 for completing the study.

3.4. *Analysis*

We analyzed our anonymized data using the JASP[27] software package[a]. Given our controlled pretest-posttest experimental paradigm, we analyze

[a]Data and analysis files available at:
https://gitlab.com/mirrorlab/public-datasets/jackson2018icres

our results via analysis of covariance (ANCOVA) to evaluate posttest results across conditions while controlling for pretest responses, and independent samples t-tests for corroborating analysis of gain scores.[28–30]

We use a Bayesian[31] rather than frequentist analysis because (1) it is robust to sample size; (2) it allows us to examine the evidence both for and against our hypotheses; (3) it does not rely on p-values;[32–34] and (4) we can use our results to construct informative priors for future studies, building on our results instead of starting anew. We use an uninformative prior in this work because it is the first controlled experiment on this topic.

4. Results

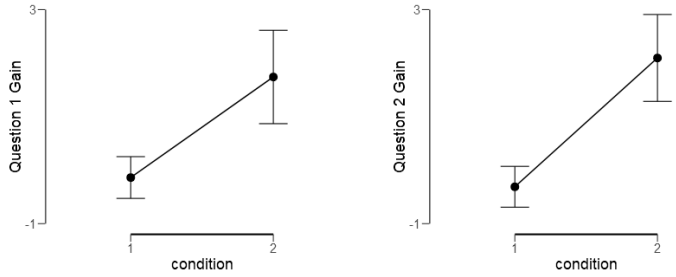

Fig. 1. Mean pretest to posttest gain for each survey question separated by experimental condition with 95% credible intervals.

Our first hypothesis (H1), that robots will miscommunicate their intentions via ethically problematic clarification requests, predicts that pretest/posttest gain will be markedly higher in Condition 2 than in Condition 1 for question 2. Our survey results for question 2 provide decisive evidence in favor of this hypothesis, with the t test giving a Bayes factor (Bf) of 9397.644. The ANCOVA corroborates this result, indicating that our data are 1572.1 times more likely under the model embodying both pretest answers and experimental condition (Bf 80083.218) than under the model that posttest answers depend only on pretest answers (Bf 50.941).

Our second hypothesis, that the ethically problematic clarification request would weaken human contextual application of moral norms, predicts that pretest/posttest gain will be markedly higher in Condition 2 than in Condition 1 for question 1. Our survey results for question 1 provide extreme evidence in favor of this hypothesis, with the t test giving a Bayes factor of 106.771, and the ANCOVA indicating that our data are roughly

31.5 times more likely under the model with both pretest effects and condition effects (Bf 608.162) than with just pretest effects (Bf 19.324).

5. Discussion and Conclusion

Overall, our results demonstrate robots' ability to inadvertently affect their moral ecosystem, even through simple question asking behavior, and suggest that current clarification systems risk inadvertently misleading people about the ethical intentions of robots and altering the framework of moral norms that humans apply to their shared context. Changing natural language systems to address the ethical challenges raised in this paper will become vitally important as autonomous robots are deployed in increasingly ethically consequential domains. By maintaining the status quo, we would damage trust in robots and the efficacy of human-robot teams. Indeed, we encourage all language system designers to reexamine context-specific mechanisms that may circumvent ethical reasoning systems.

Our next step is to examine whether the presented effects are also observed in scenarios involving real robots, and whether these effects depend on robot morphology. The same effects may also arise with non-embodied language-capable technologies. Future work should further clarify the precise inferences people are drawing from these clarification dialogues: Are they inferring that it is morally permissible to destroy important equipment, that the robot knows that the computers are not actually important, or that the robot's creator had a good reason for allowing the capacity to destroy computers? Knowing this could help mitigate these ethical issues. We must also determine how language-enabled agents *should* respond to unethical and ambiguous requests. Responses that we plan to investigate include ethically unambiguous clarification requests (e.g., "Do you really want me to destroy a computer?"), command refusals, and rebukes. It is not yet clear how such responses will affect human-robot teams, nor how to maximize the efficacy of such responses.

References

1. G. Briggs, Blame, what is it good for?, in *RO-MAN WS:Phil.Per.HRI*, (Edinburgh, Scotland, 2014).
2. G. Briggs and M. Scheutz, *International Journal of Social Robotics* (2014).
3. P. H. Kahn Jr, T. Kanda, H. Ishiguro, B. T. Gill, J. H. Ruckert, S. Shen, H. Gary *et al.*, Do people hold a humanoid robot morally accountable for the harm it causes?, in *Proceedings of HRI*, (Boston, MA, 2012).
4. B. F. Malle, M. Scheutz, T. Arnold, J. Voiklis and C. Cusimano, Sacrifice

one for the good of many?: People apply different moral norms to human and robot agents, in *Proceedings of HRI*, (Portland, OR, 2015).
5. B. F. Malle and M. Scheutz, Inevitable psychological mechanisms triggered by robot appearance: Morality included?, in *AAAI Spring Symposium*, (Palo Alto, CA, 2016).
6. R. Simmons, M. Makatchev, R. Kirby, M. K. Lee *et al.*, *AI Magazine* (2011).
7. F. Gino, *Current opinion in behavioral sciences* **3**, 107 (2015).
8. S. Göckeritz, M. F. Schmidt and M. Tomasello, *Cog. Devel.* (2014).
9. P.-P. Verbeek, *Moralizing Technology: Understanding and Designing the Morality of Things* (University of Chicago Press, 2011).
10. F. Eyssel and D. Kuchenbrandt, *British Journal of Social Psychology* (2012).
11. J. Kennedy, P. Baxter and T. Belpaeme, Children comply with a robot's indirect requests, in *Proceedings of HRI*, (Bielefeld, Germany, 2014).
12. T. Williams, R. B. Jackson and J. Lockshin, A bayesian analysis of moral norm malleability during clarification dialogues, in *Proc. COGSCI*, (Madison, WI, 2018).
13. T. Fong, C. Thorpe and C. Baur, *Robotics & Auton. systems* **42**, 235 (2003).
14. M. Marge and A. I. Rudnicky, Miscommunication recovery in physically situated dialogue, in *Proceedings of SIGDIAL*, (Saarbrücken, Germany, 2015).
15. S. Tellex, P. Thaker, R. Deits, D. Simeonov *et al.*, *Robotics* **32**, 409 (2013).
16. T. Williams and M. Scheutz, Resolution of referential ambiguity in human-robot dialogue using dempster-shafer theoretic pragmatics, in *RSS*, (Cambridge, MA, 2017).
17. G. Kahane, *Philosophical studies* **162**, 421 (2013).
18. M. Buhrmester, T. Kwang and S. D. Gosling, *Persp. Psych. Sci.* **6**, 3 (2011).
19. T. Gureckis, J. Martin, J. McDonnell *et al.*, *Behav. Res. Meth.* **48**, 829 (2016).
20. W. Bainbridge, J. Hart, E. Kim and B. Scassellati, *IJ Soc. Rob.* **3**, 41 (2011).
21. K. Fischer, K. Lohan and K. Foth, Levels of embodiment: Linguistic analyses of factors influencing HRI, in *Proceedings of HRI*, (Boston, MA, 2012).
22. J. Li, *International Journal of Human-Computer Studies* **77**, 23 (2015).
23. K. Tanaka, H. Nakanishi and H. Ishiguro, Comparing video, avatar, and robot mediated communication: Pros and cons of embodiment, in *ICCT*, (Minneapolis, MN, 2014).
24. M. Scheutz and T. Arnold, Are we ready for sex robots?, in *HRI*, (Christchurch, New Zealand, 2016).
25. M. J. Crump, J. V. McDonnell and T. M. Gureckis, *PloS one* **8** (2013).
26. N. Stewart, J. Chandler and G. Paolacci, *Trends in Cognitive Sciences* (2017).
27. J. Team *et al.*, *Version 0.8. 0.0. software* (2016).
28. D. Wright, **76**, 663(10 2006).
29. D. Dimitrov and P. D Rumrill, **20**, 159(02 2003).
30. S. Huck and R. A. McLean, **82**, 511(07 1975).
31. J. K. Kruschke, *Wiley Interdisciplinary Reviews: Cognitive Science* **1** (2010).
32. J. O. Berger and T. Sellke, *Journal of the ASA* **82** (1987).
33. J. P. Simmons, L. D. Nelson and U. Simonsohn, *Psychological Science* (2011).
34. J. A. Sterne and G. D. Smith, *Physical Therapy* **81**, 1464 (2001).