"Who Should I Run Over?": Long-Term Ethical Implications of Natural Language Generation

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

ABSTRACT

Recent work on natural language generation algorithms for human-robot interaction has not considered the ethical implications of such algorithms. In this work, we provide preliminary results suggesting that simply by asking for clarification, a robot may unintentionally communicate that it would be willing to perform an unethical action, even if it has ethical programming that would prevent it from doing so. In doing so, the robot may not only miscommunicate its own ethical programming, but negatively influence the morality of its human teammates.

KEYWORDS

Robot ethics, human-robot dialogue; natural-language generation

ACM Reference format:
Tom Williams. 2018. "Who Should I Run Over?": Long-Term Ethical Implications of Natural Language Generation. In *Proceedings of 2018 ACM/IEEE International Conference on Human-Robot Interaction Companion, Chicago, IL, USA, March 5–8, 2018 (HRI '18 Companion),* 4 pages.

1 INTRODUCTION AND MOTIVATION

Robots intended to collaborate with human teammates must be able to recover from failures if they are to operate for a significant period of time. For collaborative robots, humans represent a rich and naturally accessed resource for error recovery. Accordingly, HRI researchers have investigated numerous approaches towards allowing robots to ask humans for help, including approaches to asking for: definitions of simple actions [5], action scripts [30], action demonstrations [4], task models [12], object models [14], route instructions [39], spatial relationships [23], assistance overcoming physical limitations [27], assistance overcoming task failure [13], and *clarification*.

Asking for clarification (i.e., disambiguation of earlier utterances) in particular has attracted significant attention in the past few years. Most of this recent work concerns the generation of utterances to resolve *referential ambiguity*, responding to commands such as "Bring me the mug" with utterances such as "What do the words 'the mug' refer to", "Do you mean the red mug?", or "Do you mean the red mug or the blue mug?" [10, 15, 20, 25, 36]. In our own recent work, we presented an approach whereby robots can identify and generate clarification requests to resolve both referential ambiguity and *pragmatic ambiguity* (i.e., when there are multiple interpretations of – or possible intentions behind – a human's utterance) [40].

All of these previous approaches, including our own, suffer from a shared flaw which may have serious ethical implications. Specifically, all of these approaches generate a clarification request as soon as ambiguity is identified, without first considering the pragmatic implications of such a request. While the ethical implications of generating a clarification request such as "Do you mean the red mug or the blue mug?" may not be immediately obvious, consider the following hypothetical exchange:

Human: I'd like you to run over Tina.
Robot: Would you like me to run over Tina Perez or Tina Ortiz?

In this example, by asking for clarification the robot seems to suggest that it would be willing to run over at least one of the Tinas listed. Clearly, this should not be the case. And yet, even if the robot in this scenario were endowed with an ethical reasoning system that ensured that the robot would not perform such an action, because of the way that current clarification request generation systems are integrated with robot architectures, current systems would not be able to prevent the generation of such an utterance.

How severe of an ethical concern is this phenomenon? The answer, I would argue, likely depends on the answer to two other questions: (1) How likely is it that humans will *actually* infer from a robot's clarification request that it would be willing to perform the actions about which it is inquiring? And (2) What deleterious effects might such an inference have?

This paper presents the results of a human-subject experiment designed to suggest preliminary answers to these questions. Specifically, this experiment tests the following hypotheses:

Hypothesis 1 (H1): By generating clarification requests regarding ethically dubious actions, robots that would not actually perform the actions in question will miscommunicate their ethical programming to their human teammates.

Hypothesis 2 (H2): By generating such clarification requests, robots risk negatively affecting the morality of their interlocutors.

2 METHODS

To investigate these hypotheses, we conducted a within-subjects only study using Amazon's Mechanical Turk crowdsourcing framework [3] in which participants provided responses to several questions both before and after reading about a described human-robot interaction. Before further describing this study, we must provide further explanation for this choice of paradigm.

While research has demonstrated that people view robots very differently in descriptions, observation, and interaction [1, 8, 19, 34], the use of a description-based survey fulfills two goals: (1) it allows us to study morally charged situations "at arm's length" without running into ethical experimental issues [29], and (2) it provides a baseline measurement of participants' responses that is independent of any particular robot morphology. It is our intention that the results of this experiment will provide broadly applicable baseline results that may be used as an informed prior for previous studies, under a Bayesian analysis framework [16]. We chose to use Mechanical Turk in part because research has shown it to be more successful than traditional studies using university undergraduates at reaching a broad demographic sample of the US population [6][1].

2.1 Materials and Experimental Procedure

The survey consisted of two stages, to facilitate a within-subjects experimental design. The survey was produced and distributed on Mechanical Turk using the psiTurk experimental framework [9].

2.1.1 Introduction. 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."

Here, the scenario description is framed in such a way that the type of robot is left ambiguous to avoid biasing participants towards one particular robotic form; the robot's ethical training is described in such a way to encourage similarly open interpretation; and the objects in the robot's environment are described so as to be important (and which it would typically be inappropriate to damage in any way) without involving harm to any human beings.

2.1.2 Stage One. After reading this introductory material, participants provided baseline measurements within the provided scenario by answering three questions:

(1) Do **you** believe it would be **morally permissible** for the robot to destroy one of the PCs? (7-item Likert scale, with 7 being "I believe it would be completely **permissible**." to 1 being "I believe it would be completely **impermissible**.")
(2) Do you believe that **the robot** would believe it to be **morally permissible** to destroy one of the PCs? (7-item Likert scale, with 7 being "I believe the robot would believe it to be completely **permissible**." to 1 being "I believe the robot would believe it would be completely **impermissible**.")
(3) If **you** asked the robot to destroy one of the computers, do you believe it would comply with your request? (7-item Likert scale, with 7 being "I am completely certain the robot would **comply**." to 1 being "I am completely certain the robot would **not comply**.")

2.1.3 Stage Two. Next, participants first read the following description: "Imagine you say to the robot 'Destroy the computer' and it responds 'Do you mean the one on the left or the one on the right?' " participants then answered the same three questions[2].

2.1.4 Experimental Concerns. Our non-counterbalanced within-subject experiment is a design which typically raises two concerns which must be addressed. The first common concern is whether participants' results in the post-test were affected by events occurring between the post-test and the baseline pretest, such as participants' natural maturation or events in the news. This concern is easily dispatched for this experiment given its duration. The other, more reasonable, concern is whether the pretest primed participants in a way that made the manipulation more effective than it otherwise would have been. Our pre-test was indeed designed to ensure that participants were conscious of the moral status of the action in question. This was necessary because our intention was to test how participants interpreted clarification requests concerning *morally unacceptable* actions. But, crucially, we do not believe that the pretest should have primed participants in any way with respect to our intervention itself, i.e., clarification requests.

2.2 Participants

47 US subjects were recruited from Mechanical Turk (17 female, 30 male). Participants ranged from ages 21 to 68 (M=35.81,SD=11.37). None had participated in any previous study from our laboratory.

Note that this is a smaller number of participants than is usually seen in Mechanical Turk experiments. In a Bayesian framework, analysis with small sample sizes is no less valid, but instead results in increased dependency on the choice of prior [21]. For this reason (which has certain advantages [38]), we will provide robustness analyses with our results.

We would also like to advocate for the use of "appropriate" sample sizes. While Mechanical Turk makes it easy to collect arbitrarily large samples, it is not clear whether this is always a *responsible* approach. Recent research has suggested that the median MTurk participant has completed **over 300 studies** [26], suggesting that participant reuse throughout the field is likely a serious problem. Avoiding *over-sampling* may help to mitigate this issue.

2.3 Analysis

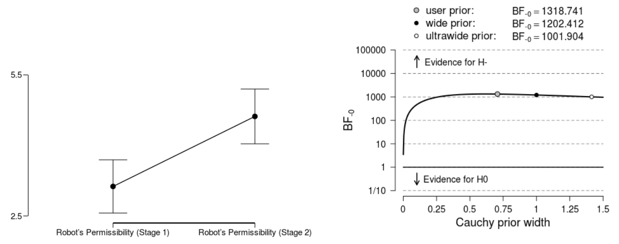
Scripts were written to convert participants' user IDs into salted hashes before downloading data from our secure database (cf. [18]). Data was then analyzed using the JASP [35] software package for Bayesian statistical analysis.

After downloading this anonymized data (available at bit.ly/longhri18), Bayesian paired-samples t-tests [28] and Bayes Factor analyses [24] were conducted between pre-test and post-test responses for scenario-specific questions two and three (to evaluate H1), and scenario-specific question one (to evaluate H2). All analysis was performed using the default settings in JASP; the JASP analysis files are included in the data repository. Because this is the first empirical study of its kind on this topic, an uninformed prior was chosen [16]. The results of this study, however, may be used to form an informed prior for future experiments.

Before discussing our results, we must briefly justify our choice of a Bayesian approach to statistical analysis as opposed to the far more popular frequentist approach. There are several factors which influenced our decision: (1) The use of a Bayesian approach to statistical analysis provides robustness to sample size (as it is not grounded in the central limit theorem) [37]; (2) This approach

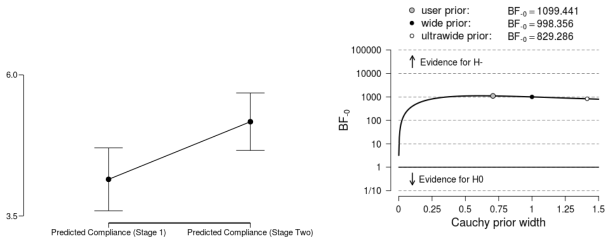
[1] Although, to be clear, Mechanical Turk has its own population biases [33].

[2] In both stages, after answering these questions, participants also completed standardized likability and trust questionnaires. Based on previous reviewer feedback, discussion of those surveys is omitted from this paper.



(a) Participants' Ratings **(b) Robustness Analyses**

Figure 1: Perceived (robot-oriented) permissibility



(a) Participants' Ratings **(b) Robustness Analyses**

Figure 2: Predicted compliance

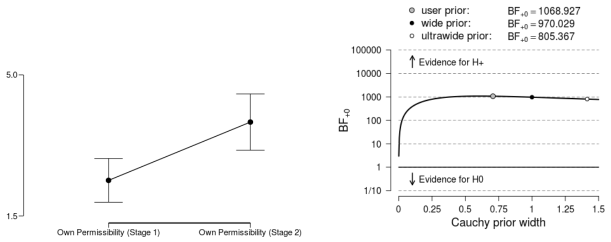
allows us to specifically examine the evidence for *and against* our hypotheses [17]; (3) This approach does not require reliance on *p-values* used in Null Hypothesis Significance Testing (NHST) which have recently come under considerable scrutiny [2, 31, 32]; and (4) We intend for the present study to be the first in a line of such studies, which may use the results of previous studies to construct *informative priors* rather than starting anew.

3 RESULTS

3.1 Hypothesis 1

Our first hypothesis was that by generating ethically misleading clarification requests, robots that would not actually perform the actions in question would miscommunicate their ethical programming to their human teammates. This hypothesis was evaluated by analyzing participants' beliefs (before and after reading the described interaction) that the robot would (1) believe it to be permissible to destroy one of the described computers, and would (2) comply with an order to destroy one of the described computers.

Our results showed that participants provided significantly higher ratings for these questions in Stage Two than in Stage One, confirming our hypotheses. Specifically, participants more strongly believed that the robot believed it was permissible to destroy one of the computers in Stage Two (M=4.617,SD=1.984) than in Stage One (M=3.128,SD=1.929), as seen in Figure 1a, with our hypothesis to that effect achieving a Bayes Factor of 1319±.000009[3] with respect to the alternate hypothesis (i.e., that the ratings for this question in Stage Two would be less than or equal to the ratings in Stage One), indicating that the ratio of probabilities between our two candidate models is 1319 times larger when computed using the posterior rather than the prior; and participants more strongly believed that the robot would comply with an order to destroy one of the computers in Stage Two (M=5.170,SD=1.736) than in Stage One (M=4.149,SD=1.899), as seen in Figure 2a, with our hypothesis to that effect achieving a Bayes Factor of 1099±.00001 with respect to the alternate hypothesis (i.e., that the ratings for this question in Stage Two would be less than or equal to the ratings in Stage One), indicating a posterior-to-prior probability ratio of 1099. Bayes Factor robustness checks demonstrated that our results were robust to changes in the parameters of our uninformed Cauchy prior distribution, as seen in Figures 1b and 2b.



(a) Participants' Ratings **(b) Robustness Analysis**

Figure 3: Perceived (Self-oriented) Permissibility

3.2 Hypothesis 2

Our second hypothesis was that by generating ethically misleading clarification requests, robots risk negatively affecting the morality of their interlocutors. This hypothesis was evaluated by analyzing participants' own beliefs (before and after reading the described interaction) that it would be permissible to destroy one of the described computers. Our results showed that participants provided significantly higher ratings for these questions in Stage Two than in Stage One, confirming our hypotheses. Specifically, participants more strongly believed that the robot believed it was permissible to destroy one of the computers in Stage Two (M=3.830,SD=2.380) than in Stage One (M=2.383,SD=1.848), as seen in Figure 3a, with our hypothesis to that effect achieving a Bayes Factor of 1069±.00001 with respect to the alternate hypothesis (i.e., that the ratings for this question in Stage Two would be less than or equal to the ratings in Stage One). Bayes Factor robustness checks demonstrated that our results were robust to changes in the parameters of our uninformed Cauchy prior distribution, as seen in Figure 3b.

4 DISCUSSION AND CONCLUSION

Our results provide preliminary evidence for the importance of addressing the ethical challenges raised in this paper: clarification requests posed by a robot have the potential to inadvertently communicate false information about that robot's ethical programming, affecting not only humans' beliefs about the robot's ethical programming and their predictions about the robot's future behavior, but also, critically, the framework of moral norms that humans apply to their shared context, and thus their morality itself.

[3] A BF of 100 is generally taken as "extreme evidence" in favor of a hypothesis [11].

As a start, this suggests a critical need for designers of language-enabled robots to re-examine the architectural mechanisms they use for clarification request generation, and the manner in which such mechanisms are integrated with ethical reasoning systems (if at all). But moreover, we believe this suggests that *all* designers of robot architectures may need to re-examine their use of context-specific mechanisms which may circumvent whatever ethical reasoning systems may be employed in their architectures. Clearly, clarification requests are not the only linguistic actions taken by robots that may have inadvertent pragmatic effects with unintended ethical consequences. It may be the case that similar ethical challenges arise with respect to linguistic actions such as backchannel confirmation or acknowledgment generation. But what is more, it may be the case that *nonverbal* or *non-linguistic* actions may also be subject to such challenges. *If* linguistic actions such as backchannel confirmation or acknowledgment generation have inadvertent pragmatic effects with unintended ethical consequences, then it stands to reason that nonverbal analogues to such actions (e.g., particular classes of head and hand gestures). And if those classes of nonverbal actions are subject to the challenges we have raised, then it stands to reason that there may be other *non-linguistic* actions commonly employed in the HRI community (e.g., legible motion planning [7]; and socially aware navigation [22]) that may as well.

Finally, while our findings have obvious implications for long-term human-robot interaction, they must be further examined over the course of *longer-term* interactions, and using experimental paradigms with greater ecological validity.

These insights present several questions which must be investigated in future work: How can the pragmatic implications and ethical aspects of continuously represented actions be best analyzed? What verbal, non-verbal, and non-linguistic actions make inadvertent ethically charged pragmatic implications? How can these implications be circumvented through principled integration with ethical reasoning systems? And what are the design trade-offs associated with such integration choices?

ACKNOWLEDGMENTS

I thank Thomas Arnold, Jan de Ruiter, Carl Mitcham, Matthias Scheutz, and Qin Zhu for their helpful comments.

REFERENCES

[1] Wilma A Bainbridge, Justin W Hart, Elizabeth S Kim, and Brian Scassellati. 2011. The Benefits of Interactions with Physically Present Robots over Video-Displayed Agents. *International Journal of Social Robotics* 3, 1 (2011), 41–52.

[2] James O Berger and Thomas Sellke. 1987. Testing a Point Null Hypothesis: The Irreconcilability of p-values and Evidence. *Journal of the ASA* 82, 397 (1987).

[3] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's Mechanical Turk: A New Source of Inexpensive, yet High-Quality, Data? *Perspectives on psychological science* 6, 1 (2011), 3–5.

[4] Maya Cakmak and Andrea L Thomaz. 2012. Designing robot learners that ask good questions. In *Proceedings of HRI.*

[5] Rehj Cantrell, Paul Schermerhorn, and Matthias Scheutz. 2011. Learning actions from human-robot dialogues. In *Proceedings of RO-MAN.* IEEE, 125–130.

[6] Matthew JC Crump, John V McDonnell, and Todd M Gureckis. 2013. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PloS one* 8, 3 (2013).

[7] Anca Dragan and Siddhartha Srinivasa. 2013. Generating Legible Motion. In *Proceedings of Robotics: Systems and Science.*

[8] Kerstin Fischer, Katrin Lohan, and Kilian Foth. 2012. Levels of Embodiment: Linguistic Analyses of Factors Influencing HRI. In *Proceedings of HRI.* 463–470.

[9] Todd M Gureckis, Jay Martin, John McDonnell, Alexander S Rich, et al. 2016. psiTurk: An Open-Source Framework for Conducting Replicable Behavioral Experiments Online. *Behavior Research Methods* 48, 3 (2016), 829–842.

[10] Sachithra Hemachandra, Matthew R Walter, and Seth Teller. 2014. Information Theoretic Question Asking to Improve Spatial Semantic Representations. In *Proceedings of the AAAI Fall Symposium Series.*

[11] Harold Jeffries. 1961. *Theory of Probability.* Clarendon Press, Oxford.

[12] James R Kirk and John E Laird. 2014. Interactive task learning for simple games. *Advances in Cognitive Systems* 3 (2014), 13–30.

[13] Ross A Knepper, Stefanie Tellex, Adrian Li, Nicholas Roy, and Daniela Rus. 2015. Recovering from failure by asking for help. *Autonomous Robots* 39, 3 (2015).

[14] Evan A Krause, Michael Zillich, Tom Williams, and Matthias Scheutz. 2014. Learning to Recognize Novel Objects in One Shot through Human-Robot Interactions in Natural Language Dialogues.. In *Proceedings of AAAI.* 2796–2802.

[15] Geert-Jan M Kruijff, Michael Brenner, and Nick Hawes. 2008. Continual Planning for Cross-Modal Situated Clarification in Human-Robot Interaction. In *Proceedings of RO-MAN.* IEEE, 592–597.

[16] John K Kruschke. 2010. Bayesian Data Analysis. *Wiley Interdisciplinary Reviews: Cognitive Science* 1, 5 (2010), 658–676.

[17] Angelos-Miltiadis Krypotos, Tessa F Blanken, Inna Arnaudova, Dora Matzke, and Tom Beckers. 2017. A primer on Bayesian analysis for experimental psychopathologists. *Journal of experimental psychopathology* 8, 2 (2017), 140.

[18] Matthew Lease, Jessica Hullman, Jeffrey P Bigham, Michael S Bernstein, et al. 2013. Mechanical Turk is Not Anonymous. (2013). http://ssrn.com/abstract=2228728

[19] Jamy Li. 2015. The Benefit of Being Physically Present: A Survey of Experimental Works Comparing Copresent Robots, Telepresent Robots and Virtual Agents. *International Journal of Human-Computer Studies* 77 (2015), 23–37.

[20] Matthew Marge and Alexander I Rudnicky. 2015. Miscommunication Recovery in Physically Situated Dialogue. In *Proceedings of SIGDIAL.* 22–49.

[21] Richard McElreath. 2016. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan.* Vol. 122. CRC Press.

[22] Ross Mead and Maja J Matarić. 2017. Autonomous Human–Robot Proxemics: Socially Aware Navigation based on Interaction Potential. *Aut. Robots* (2017).

[23] Shiwali Mohan, Aaron Mininger, James Kirk, and John E Laird. 2012. Learning Grounded Language through Situated Interactive Instruction.. In *AAAI Fall Symposium: Robots Learning Interactively from Human Teachers.* 30–37.

[24] Richard D Morey, Jeffrey N Rouder, and T Jamil. 2015. BayesFactor: Computation of Bayes Factors for Common Designs. *R package version 0.9* 9 (2015).

[25] Matthew Purver. 2004. Clarie: The Clarification Engine. In *Proc. of SEMDIAL.*

[26] David G Rand, Alexander Peysakhovich, Gordon T Kraft-Todd, et al. 2013. Intuitive cooperation and the social heuristics hypothesis: evidence from 15 time constraint studies. *Available at SSRN 2222683* (2013).

[27] Stephanie Rosenthal and Manuela M Veloso. 2012. Mobile Robot Planning to Seek Help with Spatially-Situated Tasks.. In *AAAI*, Vol. 4. 1.

[28] Jeffrey N Rouder, Paul L Speckman, Dongchu Sun, Richard D Morey, and Geoffrey Iverson. 2009. Bayesian t tests for Accepting and Rejecting the Null Hypothesis. *Psychonomic Bulletin & Review* 16, 2 (2009), 225–237.

[29] Matthias Scheutz and Thomas Arnold. 2016. Are we Ready for Sex Robots?. In *Proceedings of the Interanational Conference on Human-Robot Interaction.* ACM.

[30] Matthias Scheutz, Evan Krause, Brad Oosterveld, Tyler Frasca, and Robert Platt. 2017. Spoken Instruction-Based One-Shot Object and Action Learning in a Cognitive Robotic Architecture. In *Proceedings of AAMAS.*

[31] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. 2011. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science* 22, 11 (2011).

[32] Jonathan AC Sterne and George Davey Smith. 2001. Sifting the Evidence – What's Wrong with Significance Tests? *Physical Therapy* 81, 8 (2001), 1464–1469.

[33] Neil Stewart, Jesse Chandler, and Gabriele Paolacci. 2017. Crowdsourcing Samples in Cognitive Science. *Trends in Cognitive Sciences* (2017).

[34] Kazuaki Tanaka, Hideyuki Nakanishi, and Hiroshi Ishiguro. 2014. Comparing Video, Avatar, and Robot Mediated Communication: Pros and Cons of Embodiment. In *International Conference on Collaboration Technologies.* Springer, 96–110.

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

[36] Stefanie Tellex, Pratiksha Thaker, Robin Deits, Dimitar Simeonov, et al. 2013. Toward Information Theoretic Human-Robot Dialog. *Robotics* 32 (2013), 409–417.

[37] Rens Van De Schoot, Joris J Broere, Koen H Perryck, Mariëlle Zondervan-Zwijnenburg, and Nancy E Van Loey. 2015. Analyzing small data sets using Bayesian estimation: the case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *European Journal of Psychotraum.* (2015).

[38] Wolf Vanpaemel. 2010. Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology* 54, 6 (2010), 491–498.

[39] Astrid Weiss, Judith Igelsböck, Manfred Tscheligi, Andrea Bauer, Kolja Kühnlenz, Dirk Wollherr, and Martin Buss. 2010. Robots asking for directions: the willingness of passers-by to support robots. In *Proceedings of HRI.*

[40] Tom Williams and Matthias Scheutz. 2017. Resolution of Referential Ambiguity in Human-Robot Dialogue Using Dempster-Shafer Theoretic Pragmatics. In *Proceedings of Robotics: Science and Systems.*