Toward Ethical Natural Language Generation for Human-Robot Interaction

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 argue 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. Toward Ethical Natural Language Generation for Human-Robot Interaction. In *Proceedings of 2018 ACM/IEEE International Conference on Human-Robot Interaction Companion, Chicago, IL, USA, March 5–8, 2018 (HRI '18 Companion),* 2 pages.
https://doi.org/10.1145/3173386.3176975

1 INTRODUCTION AND MOTIVATION

As interactive robots with progressively advanced capabilities are designed and introduced, it becomes increasingly important to consider the ethical implications of the decisions made in designing those robots. Recent experimental evidence demonstrating human perception of robots as moral agents [8], a surge in robot ethics work from adjoining fields and a number of appeals to robot designers from robot ethicists [14] have led to increased attention to ethical concerns within the HRI community.

These appeals acknowledge that *robots can and will cause harm*, not only as the far-off existential threats popular in the press, but in everyday scenarios that are already commonplace [1]. Scheutz, for example [14], argues that *any* robot has the potential to cause harm, and as such qualifies as a *potential impact agent* within the framework of James Moor [10]. Scheutz further argues that it is not enough for robots to be *implicit ethical agents* (i.e., robots with built-in safety measures), but must instead be aware of the harm they may cause, both physical and emotional (due to their ability to form (unreciprocated) emotional connections [13]), and must instead be *explicit ethical agents* actively seeking to avert such harm.

HRI '18 Companion, March 5–8, 2018, Chicago, IL, USA

ACM ISBN 978-1-4503-5615-2/18/03.
https://doi.org/10.1145/3173386.3176975

To develop such robots, Malle & Scheutz have argued that robot designers' first goal must be to develop social robots that have *moral competence* [6], i.e., a system of moral norms [7] and the ability to use those norms for the purposes of moral cognition [22], moral decision making, and moral communication [15]. There have been numerous attempts to devise mechanisms to support robots' moral decision making, but very little research directly examining the ethics of natural-language based human-robot interaction. An explicit investigation of the ethical implications of natural language-based HRI is crucial, both because users expect language-capable robots to be more aware of their socio-cultural context [17], and because of the number of ethically charged design decisions that must be made specifically for language-capable robots.

Ethical concerns related to Natural Language Generation (NLG) in particular are challenging to deal with due to the inherent tension between transparency, accuracy, and robustness of machine learning methods for NLG – what Thieltges et al. refer to as the "Devil's Triangle" [19]. As such, researchers have called for the integration of ethical decision-making systems into NLG software to ensure moral behavior, echoing the aforementioned appeals made by robot ethicists [5]. Unfortunately, such integration is not present in the current HRI literature. As a consequence, we have identified a number of ethical concerns in the design of recently presented HRI algorithms for natural language generation.

2 RECONSIDERING CLARIFICATION REQUEST GENERATION

While clarification request generation has been of interest to the field of computational linguistics for many years [20], it has only recently been addressed in situated contexts [9, 18, 24]. These works seek to respond 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?" While these questions may not seem ethically fraught, 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, the ethical implications of the robot's clarification request become clear with respect to the previously discussed work. By asking for clarification, the robot seems to be suggesting 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, which, we further argue, could have serious ethical ramifications.

We argue that the generation of such clarification requests is ethically problematic for the following reasons: By generating such clarification requests, robots are signaling that there is some answer to their question which might cause them to perform an unethical action. As such, by generating such clarification requests, robots that would *not* perform the actions in question miscommunicate their ethical programming. For explicit ethical agents, this is a serious problem for several reasons. First, this miscommunication causes unnecessary obfuscation. Research has shown that to engender trust, (1) the motivation behind a robot's behavior should be *transparent* [3, 23], and (2) robots should endeavor to create *shared mental models* with their human teammates [11].

Clearly, a miscommunication of the robot's ethical programming is a failing in terms of transparency and results in divergent mental models. As such, we believe that by generating such clarification requests, robots will cause a loss in trust between human teammates and themselves. For *any* agent this is a serious problem, as an unwarranted loss of trust can result in equally unwarranted misuse or disuse [12]. For robots whose use is necessary to increase task efficiency or teammate safety, such misuse or disuse has the potential for obvious negative consequences [4].

Finally, and most seriously, by miscommunicating its own ethical programming, a robot risks miscommunicating the moral norms it believes to hold in its current context. The psychological literature has demonstrated that morality is not innate, but needs to be taught (and enforced) by all community members, making morality (and by extension moral norms) inherently malleable [2]. What is more, recent experimental evidence has directly demonstrated robots' ability to persuade humans [16] and affect their moral decision making process through *technical mediation* [21]. Accordingly, we believe that by generating such clarification requests, robots risk negatively affecting the morality of their interlocutors; a consequence with serious negative societal repercussions.

These ethical concerns are particularly troubling because current clarification request generation algorithms are destined to generate the types of clarification requests we have highlighted. In most current clarification request generation systems, asking for clarification is a special mechanism tightly integrated with the remainder of the natural language understanding and generation pipeline: for the sake of efficiency, as soon as a source of ambiguity is identified, a clarification request generation mechanism is directly triggered. As such, there is no opportunity for ethical reasoning systems to be employed, as there is no action under consideration, so far as the system is concerned. What is more, these algorithms do not sufficiently consider the broader consequences of the utterances chosen during the clarification request generation process, if at all.

3 TOWARDS MORAL NLG FOR HRI

Given the possibly unintended pragmatic implications of clarification requests, and the ethical challenges stemming from those implications, we present the following research questions, which constitute a broad agenda for much-needed research in *moral natural language generation for human-robot interaction*: (1) how can we design language-enabled robots whose architectures do not circumvent ethical checks during clarification request generation? (2) What is the relationship between preconditions, presuppositions, and implications relative to clarification requests and their ethical dimensions? (3) What are the effects of moral judgments issued by robots, and how can language-enabled robots can be architected to appropriately decide whether, when, and how to issue blame-laden moral rebukes? (4) How can the pragmatic implications and ethical aspects of continuously represented actions be best analyzed? (5) What verbal, non-verbal, and non-linguistic actions make inadvertent ethically charged pragmatic implications? (6)How can these implications be circumvented through principled integration with ethical reasoning systems? (7) What are the design trade-offs associated with such integration choices?

ACKNOWLEDGMENTS

I thank Thomas Arnold, Carl Mitcham, and Qin Zhu for their helpful comments.

REFERENCES

[1] Thomas Arnold and Matthias Scheutz. 2017. Beyond Moral Dilemmas: Exploring the Ethical Landscape in HRI. In *Proceedings of HRI*. ACM, 445–452.
[2] Susanne Göckeritz, Marco FH Schmidt, and Michael Tomasello. 2014. Young Children's Creation and Transmission of Social Norms. *Cog. Devel.* (2014).
[3] Peter Hancock, Deborah Billings, and Kristin Schaefer. 2011. Can You Trust Your Robot? *Ergonomics in Design: The Quarterly of Human Factors Apps.* (2011).
[4] Peter Hancock, Deborah Billings, Kristin Schaefer, et al. 2011. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors* (2011).
[5] Jochen L Leidner and Vassilis Plachouras. 2017. Ethical by Design: Ethics Best Practices for Natural Language Processing. *Proceedings of EACL* (2017).
[6] Bertram F Malle and Matthias Scheutz. 2014. Moral Competence in Social Robots. In *Symposium on Ethics in Science, Technology and Engineering*. IEEE.
[7] Bertram F Malle, Matthias Scheutz, and Joseph L Austerweil. 2017. Networks of Social and Moral Norms in Human and Robot Agents. In *A World with Robots*.
[8] Bertram F Malle, Matthias Scheutz, Jodi Forlizzi, and John Voiklis. 2016. Which Robot Am I Thinking About?: The Impact of Action and Appearance on People's Evaluations of a Moral Robot. In *Proceedings of HRI*. ACM, 125–132.
[9] Matthew Marge and Alexander I Rudnicky. 2015. Miscommunication Recovery in Physically Situated Dialogue. In *Proceedings of SIGDIAL*. 22–49.
[10] James H Moor. 2006. The Nature, Importance, and Difficulty of Machine Ethics. *Intelligent Systems* 21, 4 (2006), 18–21.
[11] Mark. A Neerincx. 2007. Modelling Cognitive and Affective Load for the Design of Human-Machine Collaboration. In *Eng. Psych. and Cognitive Ergonomics*.
[12] Kristin E Schaefer, Edward R Straub, Jessie YC Chen, Joe Putney, and AW Evans. 2017. Communicating Intent to Develop Shared Situation Awareness and Engender Trust in Human-Agent Teams. *Cognitive Systems Research* (2017).
[13] Matthias Scheutz. 2011. 13 The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots. In *Robot Ethics*. MIT Press, 205.
[14] Matthias Scheutz. 2016. The Need for Moral Competency in Autonomous Agent Architectures. In *Fundamental Issues of Artificial Intelligence*. Springer, 515–525.
[15] Matthias Scheutz and Bertram Malle. 2014. "Think and do the right thing" – A Plea for morally competent autonomous robots. In *Symp. on Eth. Sci. Tech. Eng.*
[16] Michael Steven Siegel. 2008. *Persuasive Robotics: How Robots Change our Minds*. Ph.D. Dissertation. Massachusetts Institute of Technology.
[17] Reid Simmons, Maxim Makatchev, Rachel Kirby, Min Kyung Lee, et al. 2011. Believable Robot Characters. *AI Magazine* 32, 4 (2011), 39–52.
[18] Stefanie Tellex, Pratiksha Thaker, Robin Deits, Dimitar Simeonov, et al. 2013. Toward Information Theoretic Human-Robot Dialog. *Robotics* 32 (2013), 409–417.
[19] Andree Thieltges, Florian Schmidt, and Simon Hegelich. 2016. The Devil's Triangle: Ethical Considerations on Developing Bot Detection Methods. In *Proceedings of the AAAI Spring Symposium Series*. 253–257.
[20] David R Traum. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. Dissertation. University of Rochester, Rochester, NY.
[21] Peter-Paul Verbeek. 2011. *Moralizing Technology: Understanding and Designing the Morality of Things*. University of Chicago Press.
[22] John Voiklis and Bertram F Malle. 2017. Moral Cognition and its Basis in Social Cognition and Social Regulation. *Atlas of Moral Psychology* (2017).
[23] Tom Williams, Priscilla Briggs, and Matthias Scheutz. 2015. Covert Robot-Robot Communication: Human Perceptions and Implications for Human-Robot Interaction. *Journal of Human-Robot Interaction* (2015).
[24] 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*.