Toward Morally Sensitive Robotic Communication

Ryan Blake Jackson
Department of Computer Science
Colorado School of Mines
Golden, Colorado, USA
rbjackso@mines.edu

I. Introduction and Background

As robots become more capable, they will become increasingly useful in a widening variety of contexts and applications. Non-roboticists in these diverse contexts will need to interact with their new robotic colleagues to facilitate productive human-robot teaming and comfortable coexistence in social environments. Natural language provides a medium for these interactions that will allow direct and fluid communication between robots and nearly all humans, without requiring specialized protocols or hardware. Indeed, many researchers have been actively investigating the problems of natural language understanding and generation in robots for some time [1]–[3].

For language-capable robots to autonomously negotiate diverse, complex, and dynamic social situations, they require robust policies to govern not only what they say, but also how they say it. Any message that a robot might want to convey can be phrased in many different ways while maintaining the same literal primary meaning. However, these different phrasings likely carry different contextually dependent connotations and implications [4]. The optimal phrasing for any given message thus depends on myriad factors including setting, audience, discourse context, message importance, and social norms.

Accordingly, there have been a number of promising approaches towards optimizing utterance phrasing. Gervits et al., for example, describe a framework that may eventually allow artificial agents to appropriately tune pragmatic aspects of utterance realization (e.g., directness and politeness) to social norms and features of the social context (e.g., formality and urgency) [5]. Several other studies have focused on the converse problem of robots understanding the implications of human phrasing [6]–[9], with results that can inform the design of algorithms to generate robot implicature.

We are interested in choosing phrasing that aligns with and enforces *moral* norms, while also being cognizant of the social factors discussed above. We are motivated by the idea that it is just as critical to design language systems that *communicate* ethically as it is to design robots that act ethically. Research shows that people naturally perceive robots as moral agents, and, therefore, extend moral judgments and blame to robots in much the same way that they would to other people [10]–[12]. Moreover, language-capable robots are expected to be even more socioculturally aware than mute robots [13].

We thus hypothesize that, since people naturally assume that robots will follow human norms, any robot that eschews standing norms, or communicates a willingness to eschew said norms, will likely face social consequences analogous to those that a human would face (e.g., loss of trust and esteem). Though these social consequences may not matter to the robot, they are still important because they would damage the efficacy and amicability of human-robot teams.

There has been a variety of recent work seeking to enable morally competent robots, including efforts to represent norms, determine which norms are salient in a given situation, and resolve norm conflicts. For example, researchers have represented norms as pairs consisting of a deontic operator (e.g., “forbidden” or “obligatory”) and an action or state, and then stored these norms in connected networks after observing that norms tend to be activated contextually in related bundles [14]. Other researchers have represented norms as optimization objectives, allowing compliance with the optimal subset of competing norms [15]. Researchers have also applied machine learning to the task of determining which behaviors are appropriate, or which norms are active, in a social context [16].

Alongside human norms governing robot behavior, we must also consider how robotic behavior may shape human norms. A key principle of modern behavioral ethics is that human morality is dynamic and malleable [17]. The norms that inform human morality are defined and developed not only by human community members, but also by the technologies with which they interact [18], [19]. As ostensible moral agents, robots are uniquely positioned to influence human norms differently than other technologies. Research shows that robots hold measurable persuasive capacity over humans [10], [20], and that humans may grant robots ingroup social status [21]. In fact, recent work has raised concerns that humans may bond so closely with robotic teammates in military contexts that their attachment could jeopardize team performance as humans prioritize the robots well-being over mission goals [22]. These results lead us to hypothesize that social robots can wield an unexpectedly profound normative influence.

One way in which robots might wield their normative influence, while also ensuring ethical behavior, is by tactfully rejecting unethical commands. Previous work explored when and how to reject commands for various reasons, including moral qualms [23], but it remains unclear how best to realize

The research reported in this document was performed in connection with Contract Number W911NF-10-2-0016 with the U.S. Army Research Laboratory. The views and conclusions contained in this document are those of the authors and should not be interpreted as presenting the official policies or position, either expressed or implied, of the U.S. Army Research Laboratory, or the U.S. Government unless so designated by other authorized documents. Citation of manufacturer’s or trade names does not constitute an official endorsement or approval of the use thereof. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

such a rejection linguistically and how the rejection might influence human morality. Other research has investigated responding to ethical infractions with affective displays [10] and humorous rebukes [24]. However, these represent only a small slice of possible responses and are not tailored to the context or infraction.

II. Research Questions

The previous section describes prior work in both natural language generation and robot ethics; my research interests lie in the intersection between these two fields. My work explores two complimentary overarching questions: (1) how might current language generation algorithms generate utterances with unintended implications or accidentally damage the ecosystem of human norms, and (2) how can we design future language systems to phrase utterances such that they purposefully influence the human normative ecosystem as productively as possible (e.g., by implicitly reinforcing beneficial norms). My recent and current work, discussed below, provides a preliminary exploration of these questions by examining the specific linguistic phenomena of *clarification request generation* and *command rejection*.

III. Completed Studies

My initial work revealed ethical concerns regarding current algorithms for clarification request generation. Specifically, current dialogue systems request clarification as soon as ambiguity is identified within a command, before any ethical checking. Consequently, if a command is both ambiguous and unethical, a robot reflexively asking for clarification may inadvertently imply a willingness to act unethically. For example, if a robot knows about two statues, and it is asked to "break the statue", it may generate an utterance like "Should I break the metal one or the stone one?" By asking this question, the robot implies a willingness to break at least one of the statues, despite the presumable impermissibility of that act. In our initial pair of studies, participants read a human-robot clarification dialogue following this pattern. We found that the robot did accidentally communicate a willingness to violate the norm, and, perhaps more concerningly, that the clarification request changed the human's perception of the permissibility of the command (i.e., the robot's clarification request made the human think that property damage was more permissible than previously thought within their context [25], [26]).

In a paper accepted to alt.HRI 2019, we demonstrate that these findings replicate when users observe actual robots, rather than merely reading about them. This observation-based experiment differs from our original description-based experiments in three key ways. First, level of embodiment in interaction can strongly effect how people view robots, and the observation-based approach gives our results far greater external and ecological validity [27]–[30]. Second, human subjects observe a dialogue between a robot and another person instead of directly interacting with the robot, which means that our results hold for both observers and interactants. Third, the relationship between the robot and its dialogue partner changed from strangers to familiar colleagues, showing that our results are somewhat robust to social distance.

IV. Current Work

Because my previous research has shown that it is important for robots to appropriately reject unethical commands, my most recent experimental work explores phrasing in command rejection. Participants watched videos showing a human issuing an ethically problematic command to a robot and the robot responding to the command. We vary the command across two levels of ethical infraction severity, and the response across two different phrasings. One response is phrased as a question that draws attention to the infraction (e.g., "Are you sure that you should be asking me to do that?"), while the other is a rebuke (e.g., "You shouldn't ask me to do that. Its wrong!"). These response types are designed to present different levels of face threat [31] to the human. Participants watched all four parings of command and response, and we collected various measures of robot likeability, infraction severity, command permissibility, and response appropriateness after each interaction. We hypothesize that the optimal command rejection will carry a face threat proportional to the severity of the command's norm violation. We will present our results at an upcoming AI ethics conference. I am also collaborating on closely related work, partially motivated by eastern ethical traditions, examining an appeal to the robot's social role in command rejection (e.g., "A good friend wouldn't do that").

Finally, we are designing experiments to investigate similar questions about robotic command rejections and reprimands with participants interacting directly with robots, rather than simply observing an interaction. These experiments will allow us to better investigate a wider array of effects, including effects on situational awareness and trust, and whether the robot's phrasing impacts teammates' actual behavior.

V. Future Work

My work will continue to explore phrasing in clarification requests and command rejections. Having established the aforementioned issues with current clarification systems, we must determine how robots should respond to ambiguous and unethical commands, and how to generate these responses. While my experimental work provides high-level guidelines for choice of communication strategy, the goal for the future is to design algorithms to automatically determine phrasing based on social context. I plan to explore both purely logical approaches, and data driven methods that would leverage prior machine learning experience from my Master's research.

Furthermore, a robot's evident ability to influence human norms raises questions regarding the persistence and extent of this influence. Will the number of humans present affect the robot's influence? Does the robot's influence persist once humans leave the interaction setting? How long will the robot's influence last? Will these effects differ across cultures? I am looking forward to exploring these questions in the coming years, and to contributing to and collaborating with the impressive community of human-robot interaction researchers.

REFERENCES

[1] M. Scheutz, P. Schermerhorn, J. Kramer, and D. Anderson, “First steps toward natural human-like HRI,” *Autonomous Robots*, vol. 22, no. 4, pp. 411–423, May 2007.

[2] N. Mavridis, “A review of verbal and non-verbal human–robot interactive communication,” *Robotics and Autonomous Systems*, vol. 63, pp. 22–35, 2015.

[3] C. Matuszek, “Grounded language learning: Where robotics and nlp meet.” in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 5687–5691.

[4] S. C. Levinson, *Presumptive meanings: The theory of generalized conversational implicature*. MIT press, 2000.

[5] F. Gervits, G. Briggs, and M. Scheutz, “The pragmatic parliament: A framework for socially-appropriate utterance selection in artificial agents,” in *Proceedings of the Annual Meeting of the Cognitive Science Society (COGSCI)*, 2017.

[6] G. Briggs, T. Williams, and M. Scheutz, “Enabling robots to understand indirect speech acts in task-based interactions,” *Journal of Human-Robot Interaction (JHRI)*, 2017.

[7] S. Trott and B. Bergen, “A theoretical model of indirect request comprehension,” in *Proceedings of the AAAI Fall Symposium Series on Artificial Intelligence for Human-Robot Interaction (AI-HRI)*, 2017.

[8] T. Williams, G. Briggs, B. Oosterveld, and M. Scheutz, “Going beyond literal command-based instructions: Extending robotic natural language interaction capabilities,” in *Proceedings of AAAI*, 2015.

[9] S. Trott, M. Eppe, and J. Feldman, “Recognizing intention from natural language: clarification dialog and construction grammar,” in *Workshop on Communicating Intentions in Human–Robot Interaction*, 2016.

[10] 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.

[11] P. H. Kahn, T. Kanda, H. Ishiguro, B. T. Gill, J. H. Ruckert, S. Shen, H. Gary, A. L. Reichert, N. G. Freier, and R. L. Severson, “Do people hold a humanoid robot morally accountable for the harm it causes?” in *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Boston, MA, 2012, pp. 33–40.

[12] 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 the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, Portland, OR, 2015, pp. 117–124.

[13] R. Simmons, M. Makatchev, R. Kirby, M. K. Lee *et al.*, “Believable robot characters,” *AI Magazine*, no. 4, 2011.

[14] B. F. Malle, M. Scheutz, and J. L. Austerweil, “Networks of social and moral norms in human and robot agents,” in *A World with Robots*, 2017.

[15] A. Ghose and T. B. R. Savarimuthu, “Norms as objectives: Revisiting compliance management in multi-agent systems,” in *Proceedings of the 14th International Conference on Coordination, Organizations, Institutions, and Norms in Agent Systems VIII*. Springer-Verlag, 2012, pp. 105–122.

[16] R. Barraquand and J. L. Crowley, “Learning polite behavior with situation models,” in *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*, ser. HRI ’08. ACM, 2008, pp. 209–216.

[17] F. Gino, “Understanding ordinary unethical behavior: Why people who value morality act immorally,” *Current opinion in behavioral sciences*, vol. 3, pp. 107–111, 2015.

[18] S. Göckeritz, M. F. Schmidt, and M. Tomasello, “Young children’s creation and transmission of social norms,” *Cognitive Development*, 2014.

[19] P.-P. Verbeek, *Moralizing Technology: Understanding and Designing the Morality of Things*. University of Chicago Press, 2011.

[20] 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.

[21] F. Eyssel and D. Kuchenbrandt, “Social categorization of social robots: Anthropomorphism as a function of robot group membership,” *British Journal of Social Psychology*, no. 4, 2012.

[22] J. Wen, A. Stewart, M. Billinghurst, A. Dey, C. Tossell, and V. Finomore, “He who hesitates is lost (...in thoughts over a robot),” in *Proceedings of the Technology, Mind, and Society*, ser. TechMindSociety ’18. New York, NY, USA: ACM, 2018, pp. 43:1–43:6. [Online]. Available: http://doi.acm.org/10.1145/3183654.3183703

[23] 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.

[24] M. F. Jung, N. Martelaro, and P. J. Hinds, “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 (HRI)*. ACM, 2015, pp. 229–236.

[25] Anonymized, “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.

[26] ——, “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.

[27] W. Bainbridge, J. Hart, E. Kim, and B. Scassellati, “The benefits of interactions with physically present robots over video-displayed agents,” *International Journal of Social Robotics*, vol. 3, no. 1, pp. 41–52, 2011.

[28] K. Fischer, K. Lohan, and K. Foth, “Levels of embodiment: Linguistic analyses of factors influencing HRI,” in *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Boston, MA, 2012, pp. 463–470.

[29] J. Li, “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*, vol. 77, pp. 23–37, 2015.

[30] K. Tanaka, H. Nakanishi, and H. Ishiguro, “Comparing video, avatar, and robot mediated communication: Pros and cons of embodiment,” in *Proceedings of the International Conference on Collaboration Technologies (ICCT)*. Minneapolis, MN: Springer, 2014, pp. 96–110.

[31] P. Brown and S. Levinson, *Politeness: Some Universals in Language Usage*. Cambridge University Press, 1987.