Generating Appropriate Responses to Inappropriate Robot Commands

Ryan Blake Jackson
MIRROR Lab
Department of Computer Science
Colorado School of Mines

Introduction

For language-capable robots to negotiate complex 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 connotations and implications (Levinson 2000). The optimal phrasing for any message thus depends on myriad factors including context, audience, and social norms.

We are interested in choosing phrasing that aligns with and enforces (possibly context-sensitive) *moral* norms, while also being cognizant of social factors. 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 (Kahn et al. 2012). We thus hypothesize that any robot that eschews, or communicates a willingness to eschew, standing norms will likely face social consequences analogous to those that a human would face (e.g., loss of trust and esteem), which could damage the efficacy and amicability of human-robot teams.

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 (Gino 2015). Human moral norms are defined and developed not only by people, but also by the technologies with which they interact (Verbeek 2011). As ostensible moral agents with the capacity to persuade (Briggs and Scheutz 2014) and to hold ingroup social status (Eyssel and Kuchenbrandt 2012), robots are uniquely positioned to influence human norms differently than other technologies.

Research Questions

My work explores two complimentary overarching questions at the intersection of natural language generation and robot ethics: (1) how might current language generation algorithms create unintended implicatures that damage the ecosystem of human norms, and (2) how can we design language systems to phrase utterances such that they purposefully influence human norms as productively as possible (e.g., by implicitly reinforcing desirable norms). My work thus far explores these questions as they pertain to *clarification request generation* and *command rejection*.

Related Work

Optimizing utterance phrasing is an active research area. For example, Gervits et al. 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) (Gervits, Briggs, and Scheutz 2017).

Previous work has also explored when and how to reject commands for various reasons, including moral qualms (Briggs and Scheutz 2015), but it remains unclear how best to realize such rejections linguistically or how the rejection might influence human morality. Other research has investigated responding to ethical infractions with affective displays (Briggs and Scheutz 2014) and humorous rebukes (Jung, Martelaro, and Hinds 2015). However, these represent only a small slice of possible responses and are not tailored to the context or infraction.

Various studies have sought to computationalize norms. Researchers have represented norms as pairings of deontic operators (e.g., "forbidden" or "obligatory") with actions or states (Malle, Scheutz, and Austerweil 2017), treated norms as optimization objectives (Ghose and Savarimuthu 2012), and sought to learn norms (Barraquand and Crowley 2008).

Completed Studies

My initial work explored ethical concerns surrounding current algorithms for requesting clarification. 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 may inadvertently imply a willingness to act unethically by reflexively asking for clarification. For example, if a robot knows about two statues, and is asked to "break the statue", it may generate an utterance like "Should I break the one on the left or the one on the right?" 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 the experimental context (Williams, Jackson, and Lockshin 2018; Jackson and Williams 2018)).

In a paper currently under review, 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, we believe that the observation-based approach gives our results far greater external and ecological validity. Second, the experimental 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.

Current Work

Because my previous research showed the consequences of inappropriate responses to unethical commands, my most recent work, submitted to AIES'19, explores command rejection phrasing. 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 (Brown and Levinson 1987) to the human. We found that a command rejection with a face threat disproportional to the severity of the command's norm violation damages robot likeability, and is viewed as inappropriate. I am also collaborating on related work, motivated by eastern ethical traditions, examining appeals to robots' social roles in command rejection.

Finally, we are designing experiments investigating 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 situation awareness and trust, and whether the robot's phrasing impacts teammates' actual behavior.

Future Work

My work will continue to explore phrasing in clarification requests and command rejections. Having identified concerns 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 guidance for choice of communication strategy, the goal of my future work is to design algorithms to automatically adjust phrasing based on details of social context. I plan to explore both logical and data driven approaches.

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 effect on human norms last? Will these effects differ across cultures? We hope to answer these questions.

References

Barraquand, R., and Crowley, J. L. 2008. Learning polite behavior with situation models. In *Proceedings of HRI*.

Briggs, G., and Scheutz, M. 2014. How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress. *Int'l Journal of Social Robotics*.

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

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

Eyssel, F., and Kuchenbrandt, D. 2012. Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psych.*

Gervits, F.; Briggs, G.; and Scheutz, M. 2017. The pragmatic parliament: A framework for socially-appropriate utterance selection in artificial agents. In *COGSCI*.

Ghose, A., and Savarimuthu, T. B. R. 2012. Norms as objectives: Revisiting compliance management in multi-agent systems. In *Proc. COIN*, 105–122. Springer-Verlag.

Gino, F. 2015. Understanding ordinary unethical behavior: Why people who value morality act immorally. *Current opinion in behavioral sciences* 3:107–111.

Jackson, R. B., and Williams, T. 2018. Robot: Asker of questions and changer of norms? In *Proceedings of ICRES*.

Jung, M. F.; Martelaro, N.; and Hinds, P. J. 2015. Using robots to moderate team conflict: The case of repairing violations. In *Proceedings of HRI*, 229–236. ACM.

Kahn, P. H.; Kanda, T.; Ishiguro, H.; Gill, B. T.; Ruckert, J. H.; Shen, S.; Gary, H.; Reichert, A. L.; Freier, N. G.; and Severson, R. L. 2012. Do people hold a humanoid robot morally accountable for the harm it causes? In *HRI*, 33–40.

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

Malle, B. F.; Scheutz, M.; and Austerweil, J. L. 2017. Networks of social and moral norms in human and robot agents. In *A World with Robots*. Springer. 3–17.

Verbeek, P.-P. 2011. *Moralizing Technology: Understanding and Designing the Morality of Things*.

Williams, T.; Jackson, R. B.; and Lockshin, J. 2018. A bayesian analysis of moral norm malleability during clarification dialogues. In *Proceedings of COGSCI*.