Challenges in Responding to Malicious Robot-Directed Commands

Ryan Blake Jackson and Tom Williams
Colorado School of Mines Department of Computer Science
Email: {rbjackso, twilliams}@mines.edu

In human-robot interaction, the typical paradigm is for robots to behave according to human desires. While the level of human control depends on the robot's autonomy, even fully autonomous robots are tasked by humans [13]. Implicit in this approach is the assumption that human-issued directives are reasonable and acceptable. However, in the real world we must consider the inevitability of a malicious, or simply misinformed, human giving morally impermissible commands to robots. For instance, children have been observed to spontaneously abuse robots for the sake of curiosity [10].

Since human commands should not be blindly trusted, robots need to detect immoral commands and respond suitably. Previous work has investigated ensuring ethical behavior via deontic logics [2] and generating responses to ethical infractions through affective displays and humorous rebukes [1, 5].

In human-robot interactions facilitated by natural language, we must consider the adversarial scenario where a command is both ethically problematic and linguistically ambiguous. For example, consider the following interaction:

Human: Please run over Sean.

Robot: Should I run over Sean McColl or Sean Bailey?

Asking for clarification may be taken to imply that the robot is willing to run over at least one of the people listed. Even in a robot with an ethical reasoning system such that it would never *actually* run over a person, the current status quo of natural language pipelines would generate the ethically misleading clarification nonetheless. This is because reference disambiguation is triggered as a reflex action as soon as ambiguity is identified, so the system begins the clarification dialog before ethical reasoning has a chance to occur [12].

The phenomenon described above is problematic because it causes robots to unintentionally miscommunicate their ethical programming. People perceive language-capable robots as moral agents, so such miscommunications could damage trust and esteem in human-robot teams [1, 6, 8, 9]. Perhaps even more troubling is the notion that these miscommunications could weaken humans' contextual application of moral norms. An empirically supported tenet of behavioral ethics is that human morality is dynamic and malleable [4]. The norms that inform human morality are shaped, in part, by technology, [11], and language-enabled robots are in a unique position to do so more drastically than other devices. Robots' appearance as moral agents, measurable persuasive capacity [1, 7], and potential in-group status [3] all suggest that robot norm violations may negatively influence the human moral ecosystem in much the same way as human norm violations.

In previous work [12], we conducted human subjects experiments via Amazon's Mechanical Turk in which participants answered survey questions both before and after reading an ethically dubious clarification dialog involving property damage. After the clarification request, participants more strongly believed that the robot would think it was permissible to destroy the property in question, more strongly believed that the robot would ultimately destroy the property, and, crucially, more strongly believed that it would be permissible to destroy the property. This serves as evidence that (1) generating clarification requests regarding ethically problematic commands miscommunicates robots' ethical programming, and (2) generating such requests weakens the moral norms employed by humans within the applicable context.

These results show that, aside from tarnishing robot reputations among human teammates, current dialog systems can inadvertently weaken *humans'* contextual application of moral norms regardless of a robot's capacity for ethical reasoning. Therefore, research is needed to integrate robots' moral and linguistic reasoning systems in a way that prevents context-specific mechanisms from circumventing ethical reasoning systems. Though such mechanisms may adequately handle the cooperative interactions typically emphasized within the research community, current language systems are ill-equipped to handle interaction with malicious humans.

The demonstrated capacity to adversely impact the human moral ecosystem raises the question of whether and how robots could beneficially influence human ethics. In determining how language-enabled agents should respond to (potentially malicious) commands that are both ambiguous and unethical, we plan to investigate ethically unambiguous clarification requests (e.g., "Do you really want me to run over a person?"), command refusals, and rebukes. We believe that tactful responses to immoral commands could productively reinforce the norm that was violated. To maximize the efficacy of these norm reinforcements, the response type and phrasing must be tuned to the context, severity, and intention of the infraction.

While our previous experiments required participants to read hypothetical dialogues, we are currently conducting follow-up experiments using actual robots to confirm these results with concrete robot morphology and increase ecological validity.

References

[1] Gordon Briggs and Matthias Scheutz. How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress. *International Journal of Social Robotics*, 2014.

[2] Selmer Bringsjord, Konstantine Arkoudas, and Paul Bello. Toward a general logicist methodology for engineering ethically correct robots. *Intel. Sys.*, 2006.

[3] Friederike Eyssel and Dieta Kuchenbrandt. Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology*, (4), 2012.

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

[5] Malte F. Jung, Nikolas Martelaro, and Pamela 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 '15, pages 229–236. ACM, 2015.

[6] Peter H Kahn Jr, Takayuki Kanda, Hiroshi Ishiguro, Brian T Gill, Jolina H Ruckert, Solace Shen, Heather Gary, et al. Do people hold a humanoid robot morally accountable for the harm it causes? In *Proceedings of HRI*, pages 33–40, Boston, MA, 2012. ACM.

[7] James Kennedy, Paul Baxter, and Tony Belpaeme. Children comply with a robot's indirect requests. In *Proceedings of HRI*, pages 198–199. ACM, 2014.

[8] Bertram F Malle and Matthias Scheutz. Inevitable psychological mechanisms triggered by robot appearance: Morality included? In *AAAI Spring Symposium*, 2016.

[9] Bertram F Malle, Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano. Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In *Proceedings of HRI*, pages 117–124, Portland, OR, 2015. ACM.

[10] Tatsuya Nomura, Takayuki Kanda, Hiroyoshi Kidokoro, Yoshitaka Suehiro, and Sachie Yamada. Why do children abuse robots? 17:347–369, 2016.

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

[12] Tom Williams and Ryan Blake Jackson. A bayesian analysis of moral norm malleability during clarification dialogues. In *Proc. COGSCI*, 2018.

[13] H. A. Yanco and J. Drury. Classifying human-robot interaction: an updated taxonomy. In *2004 IEEE International Conference on Systems, Man and Cybernetics*, volume 3, pages 2841–2846, 2004.