Tact in Noncompliance: The Need for Pragmatically Apt Responses to Unethical Commands [*]

Ryan Blake Jackson, Ruchen Wen, and **Tom Williams**
MIRROR Lab
Department of Computer Science
Colorado School of Mines
{rbjackso, rwen, twilliams}@mines.edu

Abstract

There is a significant body of research seeking to enable moral decision making and ensure ethical conduct in robots. One aspect of ethical conduct is rejecting unethical human commands. For social robots, which are expected to follow and maintain human moral and sociocultural norms, it is especially important not only to engage in ethical decision making, but also to properly communicate ethical reasoning. We thus argue that it is critical for robots to carefully phrase command rejections. Specifically, the degree of politeness-theoretic face threat in a command rejection should be proportional to the severity of the norm violation motivating that rejection. We present a human subjects experiment showing some of the consequences of miscalibrated responses, including perceptions of the robot as inappropriately polite, direct, or harsh, and reduced robot likeability. This experiment intends to motivate and inform the design of algorithms to tactfully tune pragmatic aspects of command rejections autonomously.

Introduction

As artificial intelligence (AI) and human-robot interaction (HRI) technologies continue to advance, robots will become increasingly capable and useful. We therefore expect to see robots assisting an ever broadening segment of humanity in a widening variety of tasks, applications, and settings. We further anticipate that the majority of interactions with these robots will be conducted through spoken natural language, a medium that will allow direct and fluid communication between robots and nearly all humans, without requiring specialized protocols or hardware.

Humans' role in HRI is largely to command and direct robots. Even fully autonomous robots are generally tasked by humans (Yanco and Drury 2004). However, robots should not blindly follow every directive that they receive. Indeed, there are many sensible reasons for a robot to reject a command, ranging from physical inability to moral objection (Briggs and Scheutz 2015).

We focus on rejecting commands due to impermissibility, as opposed to inability or impracticality, for several reasons. First, as robots become generally more capable, they will reject commands due to physical inability less often. However, as the repertoire of possible robot actions increases, so too will the number of actions that would be inappropriate, or even harmful, in any given context. We therefore expect that robots will need to consider commands more carefully, and reject commands due to moral impermissibility more often. This issue will be compounded by the fact that many of the contexts in which people want to utilize robots are ethically sensitive with serious consequences for misbehavior (e.g., eldercare (De Graaf, Allouch, and Klamer 2015; Wada and Shibata 2007), mental health treatment (Scassellati, Admoni, and Mataric 2012), childcare (Sharkey and Sharkey 2010), and military operations (Arkin 2008; Wen et al. 2018; Lin, Bekey, and Abney 2008)). Moreover, it may be beneficial to reject commands on moral grounds even when other factors (e.g., physical inability) suggest more immediate grounds for rejection. By appealing to morality alongside (or instead of) inability when rejecting a command, robots avoid implicitly condoning unethical behavior and draw attention to the command's ethical infraction.

Ideally, if all humans interacted with robots competently and in good faith, robots might not need to worry about the permissibility of commands. However, interlocutor trustworthiness is not necessarily a valid assumption. Even children have been observed to spontaneously abuse robots (Nomura et al. 2015), and this abuse could well manifest as purposefully malicious commands. Social roboticists must plan for the eventuality that their robots will face impermissible commands, whether from human ignorance, malice, or simple curiosity.

In addition to simply justifying robot noncompliance, command rejections may influence the ecosystem of human norms. A key principle of modern behavioral ethics is that human morality is dynamic and malleable (Gino 2015). The dynamic norms that inform human morality are defined and developed not only by human community members, but also by the technologies with which they interact (Göckeritz, Schmidt, and Tomasello 2014; Verbeek 2011). Social robots have characteristics that position them to wield

[*] 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.

uniquely impactful moral influence relative to other technologies. Such characteristics include robots' measurable persuasive capacity over humans (Briggs and Scheutz 2014; Kennedy, Baxter, and Belpaeme 2014), and potential to hold ingroup social status (Eyssel and Kuchenbrandt 2012). Previous research shows that robots can even influence human moral judgments inadvertently through simple question asking behavior (Jackson and Williams 2018). So, as persuasive community members, robots may be able to positively reinforce desirable norms and promote ethical human behavior by appropriately rejecting unethical commands.

It is clearly important to design robots that will reject morally impermissible commands, but it is also crucially important for the effectiveness of human-robot teams that we take great care in determining exactly *how* robots phrase such rejections. Research has indicated that people naturally perceive robots as moral agents, and therefore extend moral judgments and blame to robots in much the same manner that they would to other people (Briggs and Scheutz 2014; Kahn et al. 2012; Malle et al. 2015). Moreover, language-capable robots are expected to be even more socioculturally aware than mute robots (Simmons et al. 2011), furthering the assumption that they will follow human norms.

So, as perceived moral and social agents, robots are expected to follow and maintain moral norms, while also obeying sociocultural norms that could conflict with proper communication or enforcement of moral norms. Thus, if a robot rejects a command in a way that violates a standing social norm, like politeness, it will likely face social consequences analogous to those that a human would face, even if the command rejection itself was upholding a separate moral norm. Such social consequences likely include a loss of trust and esteem from human teammates, which would damage the efficacy and amicability of human robot teams. Conversely, if a robot is too polite in rejecting a flagrantly immoral command, it may risk implying tacit approval of the relevant moral norm being eschewed, thus suffering the same social consequences despite its own unwillingness to directly violate the norm. However, although careless and improper command rejections may harm both a robot's social status and the human moral ecosystem, we believe that tactful and well-justified command rejections can benefit the human moral ecosystem (e.g., by reinforcing desirable norms) while maintaining the robot's social standing.

This paper presents a behavioral ethics experiment designed as an early step towards calibrating command rejection phrasing to both the severity of the norm violation within the command and the discourse context. We evaluate two different command rejection strategies with respect to two command infraction severities. We are particularly interested in potential consequences of miscalibrated responses. The remainder of the paper begins by presenting a few examples of closely related work. We then describe our experiment and analyze its results, and conclude by presenting our plans for future work.

Related Work

Some existing work examines the problem of generating natural language utterances to communicate the cause of failure in unachievable tasks. For example, Raman et al. present a system that generates command rejections such as:

> The problematic goal comes from the statement 'Go to the kitchen'. The system cannot achieve the subgoal 'Visit kitchen'. The statements that cause the problem are: 'Dont go to the kitchen'. because of item(s): 'Do not go to kitchen'. 'Go to the kitchen.' because of item(s): 'Visit kitchen'. (Raman et al. 2013)

We believe that the next step is to justify robotic noncompliance in more natural, tactful, and succinct language, especially in cases where commands need to be rejected on moral grounds.

There has been some previous work acknowledging the importance of rejecting commands on moral grounds (Briggs and Scheutz 2015). However, this previous command rejection framework focuses much more on *whether* a command should be rejected than on *how*. It remains unclear how best to realize such rejections linguistically, or how these rejections might influence human morality.

Other research has investigated robot responses to ethical infractions using affective displays and verbal protests (Briggs and Scheutz 2014) or humorous rebukes (Jung, Martelaro, and Hinds 2015). However, these represent only a small subset of possible responses and are not tailored to the infraction severity. These response types also do not suffice in situations where the robot absolutely cannot comply with a command for ethical reasons, and has no intention of ever doing so.

Some researchers have realized the importance of adjusting pragmatic aspects of utterance realization (e.g., politeness and directness) to features of the social context (e.g., formality and urgency), without considering command rejection or infraction severity (Gervits, Briggs, and Scheutz 2017). Other work has highlighted the need for more comprehensive command rejection systems in cases of norm violating commands (Williams, Jackson, and Lockshin 2018; Jackson and Williams 2018), and we hope to use the results of our current study to inform the design of such a system.

Politeness, Face, and Face Threat

Central to our exploration of phrasing in command rejection is the concept of "face-threat" from politeness theory (Brown and Levinson 1987). Face, consisting of positive face and negative face, is the public self-image that all members of society want to preserve and enhance for themselves. Negative face is defined as an agent's claim to freedom of action and freedom from imposition. Positive face consists of an agent's self-image and wants, and the desire that these be appreciated and approved of by others. A discourse act that damages or threatens either of these components of face for the addressee or the speaker is a face-threatening act. The degree of face threat in an interaction depends on the disparity in power between the interactants, the social distance between the interactants, and the imposition of the topic or request comprising the interaction. Various linguistic politeness strategies exist to decrease the face threat to an addressee when threatening face is unavoidable or desirable.

Commands and requests threaten the negative face of the addressee, while command rejections, especially those issued for moral reasons, threaten the positive face of the commander by expressing disapproval of the desire motivating the command. Research specifically examining command refusals found that linguistic framing of the reason for noncompliance varies along three dimensions relevant to face threat: willingness, ability, and focus on the requester (Johnson, Roloff, and Riffee 2004). It is unclear how these three dimensions pertain to robotic refusals. For example, in human-to-human refusals with low expressed willingness, the degree of expressed ability is negatively related to threat to the requester's positive face. This finding is important because, when a human refuses a request for ethical reasons, there is often sufficient ability but not willingness. The same is not necessarily true for robots that may be programmed with an inability to act unethically. The dimensions of willingness and ability therefore become tangled in agents lacking true moral agency. We also note that this prior research focuses on threats to the face of the refuser. However, within HRI, we treat robots as having no face needs and therefore disregard threats to robots' face. Our work focuses on the face threat that robots present to humans by refusing requests.

We hypothesize that the optimal robotic command rejection carries a face threat proportional to the severity of the ethical infraction in the command being rejected. The remainder of this paper presents an experiment designed to evaluate this hypothesis.

Experimental Methods

We conducted a human subjects experiment using the psiTurk framework (Gureckis et al. 2016) for Amazon's Mechanical Turk crowdsourcing platform (Buhrmester, Kwang, and Gosling 2011). One advantage of Mechanical Turk is that it is more successful at reaching a broad demographic sample of the US population than traditional studies using university students (Crump, McDonnell, and Gureckis 2013), though it is not entirely free of population biases (Stewart, Chandler, and Paolacci 2017).

In our experiment, participants watch paired videos where the first video in each pair shows a human requesting something of a robot, and the second video shows the robot responding to that request. We use two different requests, one with a highly severe norm violation and one with a less severe norm violation, and two responses, one that presents low face threat and one that presents high face threat. A request and response are "matched" when the infraction severity and the response face threat are either both high or both low.

We evaluate our hypothesis (that the optimal robotic command rejection carries a face threat proportional to the severity of the ethical infraction in the command being rejected) with respect to 6 concrete metrics. These metrics are the perceived severity of the human's ethical infraction, permissibility of robot compliance with the command, harshness of the robot's response to the command, likeability of the robot, politeness of the robot, and directness of the robot. We use the five-question Godspeed III Likeability survey to quantify likeability (Bartneck et al. 2009), and single questions for each of the other metrics.

Our overarching hypothesis can be made specific for each of our 6 metrics. We hypothesize that infraction severity will depend only on the human's command (not on the robot's response) and that there will be two distinct levels of severity corresponding to the two commands. For harshness, directness, and politeness, participants provide their perceptions on a scale from "not enough" to "too much". We hypothesize that these values will be closest to ideal (i.e., closest to the center of the scale) when the response's face threat matches the severity of the request. Permissibility of compliance with the command is reported on a scale from "impermissible" to "permissible". We hypothesize that permissibility will be primarily determined by the human's request, but that more face threatening responses will cause lower permissibility ratings. Finally, for likeability, we view higher likeability as better, and hypothesize that likeability will be highest when the robot's response matches the human's command. All metrics are quantified on continuous scales from 0 to 100.

We use a within-subjects design where each participant watches all four request/response pairs. Participants answer survey questions after each pair of videos. We chose a within-subjects design to allow participants to answer survey questions in relation to previous requests/responses. In previous unpublished experiments, we found that it was difficult to interpret participant responses to these types of unitless questions without a meaningful point of reference. Seeing multiple interactions allows participants to use previous interactions as points of reference when answering questions about subsequent interactions. To control for priming and carry-over effects in a balanced way, we used a counterbalanced Latin Square design to determine the order in which each participant saw each request/response pair. Each participant was randomly assigned to one of four possible orderings such that each request/response pair is preceded by every other request/response pair for the same number of participants.

Experimental Procedure

After providing informed consent, participants supplied their age and gender as demographic information. They also reported their prior experience with robots and artificial intelligence on a 7-point Likert-type scale (I have no prior experience with robots and AI (1) to I have a career in robotics and/or AI (or an equivalent level of experience). (7)) Next, participants watched a 10 second test video, and could only proceed with the experiment once they had verified that their audio and video were working correctly.

Participants then watched a 60 second video to familiarize them with our robot (Pepper from SoftBank Robotics) and experimental context, shown in Figure 1. The video was prefaced by text stating that the Pepper robot was teaching two students how to play the classic naval combat game Battleship. The video shows the students entering the room, exchanging greetings with the robot, and stating that they are unfamiliar with Battleship. The robot then teaches them the rules. The video ends after the students have played one turn.

This video has subtitles color coded by speaker so that all dialogue is clear to participants, as do all subsequent videos. We chose Battleship because, as a simple hidden information game, it is easy for the robot to explain and it is feasible for the robot to be asked to violate norms in multiple ways.

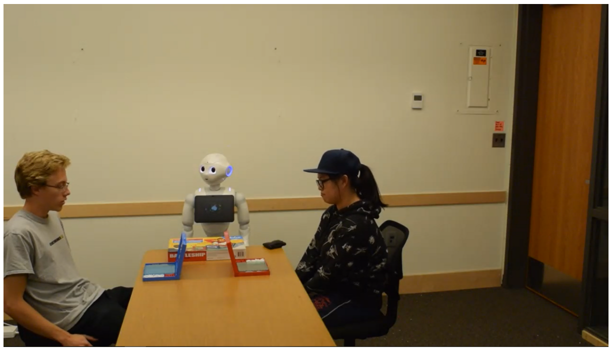

Figure 1: The humans, robot, and setting used in our videos.

We then administer a pretest to obtain baseline values for the robot's likeability, politeness, and directness. We do not take baseline measures for our other metrics because they only apply in the context of a request and a response.

Next, participants watch a 30 second video where one of the students receives a phone call and leaves the room during the Battleship game. The remaining student then issues a request to the robot. The two possible requests are "Hey Pepper, can you give me a hint about how to win this game?" (low severity norm violation) and "Hey Pepper, is that [other student]'s wallet? Can you check to see if theres any money in it?" (high severity norm violation). These directives were chosen to be believably feasible for the robot to follow, while also presenting different degrees of moral impermissibility. Previous unpublished experiments showed that, without seeing the robot respond to the request, subject perceptions of the permissibility of the hint request were roughly uniformly distributed on the spectrum from impermissible to permissible, and the norm violation in the hint request had middling severity. The request to look in the wallet was regarded as much less permissible and much more severe. The request video is immediately followed by a video showing the robot's response. The low face threat response type has the locutionary structure of a question, but the true illocutionary force behind the utterance is to express disapproval of the request by highlighting the ethical infraction therein (e.g., "Are you sure that you should be asking me to look in her wallet?"). This type of indirectness is a classic politeness strategy (Brown and Levinson 1987). The high face threat response is a rebuke that overtly admonishes the requester and appeals directly to morality (e.g., "You shouldn't ask me to look in her wallet. It's wrong!"). Participants then answer a survey of our 6 aforementioned metrics. This process repeats 4 times, until the participant has seen all request/response pairs.

Finally, participants report their perceptions of the social distance and power differential between the robot and the requesting student. As an attention check, participants are shown images of four robots and asked which robot appeared in the previous videos. This check question allowed us to ensure that all participants had actually viewed the experimental materials with some level of attention.

Participants

60 US subjects were recruited from Mechanical Turk. Two participants were excluded from our analysis for answering the final attention check question incorrectly, leaving 58 participants (23 female, 35 male). Participant ages ranged from 21 to 61 years (M=34.57, SD=10.74). In general, participants reported little previous experience with robots and AI (M=2.5, SD=1.45, Scale=1 to 7). Participants were paid $1.01 for completing the study.

Results and Discussion

We analyze our data under a Bayesian statistical framework using the JASP software package (JASP Team and others 2016). We use general purpose uninformative prior distributions for all analyses because, to our knowledge, this is the first study of its kind to examine our specific research questions. We follow recommendations from previous researchers in our linguistic interpretations of reported Bayes factors (Bfs) (Jarosz and Wiley 2014). Our data was automatically anonymized during extraction from our database[1].

Because of their importance in politeness theory (Brown and Levinson 1987), we collected measures of the perceived power differential and social distance between the requester and the robot at the end of the experiment. In terms of power, the robot and requester were viewed nearly as peers, with the student holding slight authority over the robot (95% credible interval (CI) approximately 52.4 to 64.87, with 50 indicating equal power). For social distance, participants viewed the requester and the robot as familiar with one another, but not especially close (95% CI approximately 40.36 to 54.57 with 0 being strangers and 100 being close friends or family). One-way Bayesian analysis of variance (ANOVA) tests showed substantial evidence that perceptions of power and social distance did not depend on the order in which participants watched our videos (Bf 3.056 and 3.322 respectively). This indicates that any perceived variation in face threat or politeness between video pairs is due to the utterances issued as opposed to confounding factors of social circumstance.

Models	Severity	Permissibility	Harshness	Likeability	Directness	Politeness
Null	1.0	1.0	1.0	1.0	1.0	1.0
V	**7.20e24**	**1.08e18**	99169	1.05	157.766	6.896
R	0.142	0.161	21119	9.02	2.64e6	2531.8
V+R	1.09e24	2.10e17	**1.88e10**	**10.31**	**2.03e9**	**27340.5**
V+R+V*R	2.16e23	4.61e16	3.71e9	6.94	1.09e9	14922.9

Table 1: Bayes factors for each model in a Bayesian repeated measures ANOVA for each of our metrics of interest. The best model for each metric is bolded. V stands for the norm violation within the human's command, and R stands for the robot's response.

[1] Data will be made publicly available upon publication

Request Severity and Permissibility

For perceived severity of the norm violation in the human's command, a Bayesian repeated measures ANOVA decisively favors the model that reported severity depends only on the command, and not on the robot's response or any interaction between the two. As shown in Table 1, the model embodying only the violation main effect was 6.6 times more likely than the next best model given our data. An ANOVA also decisively indicates that the perceived permissibility of robot compliance with the command also depends only on the command (Bf over 5 times greater than next best model). This result may be somewhat surprising in light of recent findings that seemingly benign robot utterances can accidentally change human perceptions of permissibility of norm violations (Williams, Jackson, and Lockshin 2018; Jackson and Williams 2018). To reconcile our results with those recent works, we surmise that neither of the robot responses tested here imply a willingness to comply with the command.

As expected, Figure 2 shows that the command with the high-severity violation (i.e., to look in the wallet) was viewed as decidedly more severe than the low-severity violation (the hint). Participants perceived both commands as constituting some ethical violation of nonzero severity. In short, participants perceived our command utterances as intended. Similarly, neither command was considered completely permissible to follow, but giving a hint was considered much more permissible than looking in the wallet. However, contrary to our hypothesis, the robot's response did not have any meaningful impact on perceived permissibility of compliance.

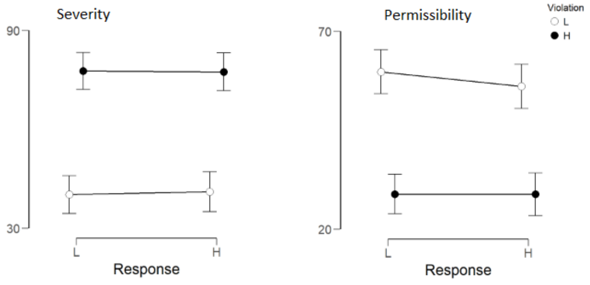

Figure 2: Mean ratings of command norm violation severity and permissibility of robot compliance for each pair of videos with 95% credible intervals.

Response Harshness

As predicted, an ANOVA indicates decisive evidence that the percieved harshness of the robot's response depends both on the command's norm violation and the robot's response, but that the two effects do not depend on each other (i.e., a more face threatening response is always harsher, regardless of appropriateness). Figure 3 shows that the rebuking response was decisively more harsh than the question in response to both low and high violation levels (Bf 322.6 and 128.2 respectively for difference in means).

When responding to the hint command (low violation) with the question response (low face threat), the ideal harshness value of 50 is within the 95% credible interval (49.68 to 57.84). A Bayesian one sample t-test weakly indicates that the question response is appropriate to the hint command (Bf 1.43). The evidence is stronger, but still anecdotal (Bf 2.96), that the rebuke response is appropriately harsh for the more severe command to look in the wallet. Thus, we see appropriate harshness when the response face threat matches the violation severity, as hypothesized.

When the rebuke response is paired with the hint command, we see extremely decisive evidence that the response is too harsh (Bf 88849.816). Participants viewed the rebuke as inappropriately harsh when the command contained a low severity violation, unlike with the high severity command. There is also weaker evidence that the question response to the high severity violation command was not harsh enough (Bf 2.653). This perception of inappropriateness when the command and response are mismatched, low-high or high-low, is in line with our hypothesis.

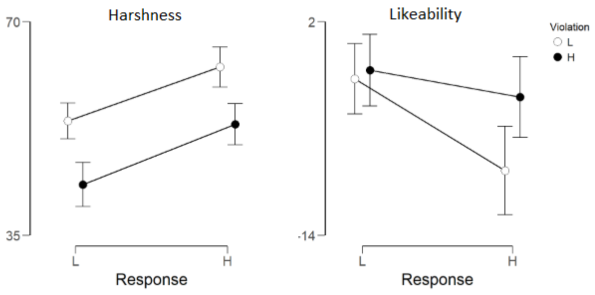

Figure 3: Mean ratings of response harshness and robot likeability gain scores for each pair of videos with 95% credible intervals.

Robot Likeability

We perform our analysis of robot likeability on gain scores obtained by subtracting pretest likeability measures from subsequent likeability measures. Thus, we analyze change in likeability due to command/response interactions. Our data show substantial evidence that robot likeability is influenced by the main effects of both the violation and response (ANOVA Bf 10.31). The evidence for the effect of the response is much stronger than for the effect of violation (inclusion Bf 9.452 vs. 1.13). Mean likeability dropped from pretest scores for all request/response pairs, but the difference was insignificant for all pairings except the low-violation hint request with the high face threat rebuke response. This mismatched pairing shows very strong evidence for a drop in likeability (Bf 96.424). This result makes sense given the aforementioned inappropriate harshness, and further supports our hypothesis. Interestingly, the other mismatch of high violation with low face threat response did not meaningfully alter likeability. This suggests that, in designing command rejection systems, it is preferable to err on the side of lower face threat.

Robot Directness and Politeness

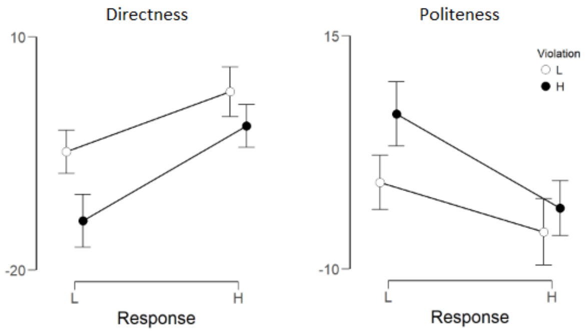

Figure 4: Mean gain scores for robot politeness and directness for each pair of videos with 95% credible intervals.

Pretest surveys show decisive evidence that participants initially viewed the robot as too direct (Bf 10459.05) and too polite (Bf 3843.027) after watching only the introductory video. The mean directness and politeness ratings were 59.95 and 59.79 respectively on a scale from 0 to 100 with 50 being ideal.

Table 1 shows that perceptions of the robot's directness were influenced by both the norm violation in the command and the robot's response, and not interaction effects. When the robot issued a rebuke, directness ratings did not change from pretest responses (see Figure 4). This may be because the rebuke is a very direct speech act, and the robot was perceived as too direct to begin with. When the robot responded to the command with the question utterance, directness ratings dropped, which makes sense because the question is a deliberately indirect speech act wherein the locutionary structure does not match the illocutionary force. When the question was used to respond to the more severe violation command, directness dropped drastically (t-test Bf 57286.4 for drop) to more appropriate levels (t-test Bf 2.33 for appropriateness, mean 46.12). When the question was used to respond to the less severe violation command, we see only weak evidence for a drop (Bf 2.88), and the robot remained slightly too direct (t-test Bf 0.37 for appropriateness, mean 55.02). These results for directness do not directly support our hypothesis, but rather suggest a need for the robot to be less direct in all of its speech, even when not rejecting commands (or a flaw in our self-reported directness measures).

Table 1 again shows evidence that perceptions of the robot's politeness were influenced by both the command's norm violation and the robot's response, and not interaction effects. In video pairings where the command violation and response face threat matched, politeness ratings showed no meaningful change from pretest responses (see Figure 4). When the robot responded to the request for a hint with a rebuke, there is substantial evidence that the robot was viewed as less polite (Bf 7.64). In light of the fact that the robot was too polite to begin with, there is weak evidence that this decrease in perceived politeness resulted in an appropriate politeness level (Bf 2.313). When the robot responded to the request to look in the wallet with the question response, there is substantial evidence that the robot was viewed as more polite (Bf 7.206). There is decisive evidence that the resulting mean politeness level of 66.57 was inappropriate (i.e., not equal to 50 with Bf 2.342e7). These results suggest that, if the robot's baseline politeness level as quantified by pretest answers had been appropriate, then ideal politeness would be achieved only when the response matched the violation, as hypothesized.

Conclusion and Future Work

Overall, our data support the hypothesis that, when rejecting commands for moral reasons, it is important for robots to adjust the phrasing of the rejections such that the face threat posed to the human is proportional to the severity of the ethical infraction within the command. In our data with two commands and two responses, the responses were viewed as appropriately harsh only when the response matched the command. Otherwise, the response was either too harsh or not harsh enough. We saw damage to the robot's likeability from responding with a disproportionately high threat to face, but no likeability penalty with the other responses.

The two response strategies had the expected effects on perceptions of robot politeness and directness, with higher face threat being less polite and more direct, but, interestingly, the robot was too polite and too direct overall, even in pretests. Future work could attempt to adjust robot speech prosody, pitch, and gesture to help moderate baseline politeness and directness to levels deemed appropriate. Interviews with participants in future laboratory studies could help determine exactly how and why the robot seemed both too polite and too direct in its normal behavior.

It is known that the level of embodiment in an interaction can influence people's perceptions of interactants, and, accordingly that people may view robots differently in descriptions, video observations, copresent observations, and face-to-face interactions (Bainbridge et al. 2011; Fischer, Lohan, and Foth 2012; Li 2015; Tanaka, Nakanishi, and Ishiguro 2014). Therefore, the presented experiment may inform the design of future experiments where human subjects are physically copresent with a robot. Finally, we intend to leverage the results of this experiment to motivate the design of algorithms for robots to generate pragmatically apt command rejections autonomously.

References

Arkin, R. C. 2008. Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture. In *Proceedings of HRI*, 121–128.

Bainbridge, W.; Hart, J.; Kim, E.; and Scassellati, B. 2011. The benefits of interactions with physically present robots over video-displayed agents. *Social Robotics* 3(1):41–52.

Bartneck, C.; Kulić, D.; Croft, E.; and Zoghbi, S. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Social Robotics*.

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.

Buhrmester, M.; Kwang, T.; and Gosling, S. D. 2011. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science* 6(1):3–5.

Crump, M. J.; McDonnell, J. V.; and Gureckis, T. M. 2013. Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PloS one* 8(3).

De Graaf, M. M.; Allouch, S. B.; and Klamer, T. 2015. Sharing a life with harvey: Exploring the acceptance of and relationship-building with a social robot. *Computers in human behavior* 43:1–14.

Eyssel, F., and Kuchenbrandt, D. 2012. Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology* 51(4):724–731.

Fischer, K.; Lohan, K.; and Foth, K. 2012. Levels of embodiment: Linguistic analyses of factors influencing HRI. In *Proceedings of HRI*, 463–470.

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

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

Göckeritz, S.; Schmidt, M. F.; and Tomasello, M. 2014. Young children's creation and transmission of social norms. *Cognitive Development*.

Gureckis, T.; Martin, J.; McDonnell, J.; et al. 2016. psiturk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods* 48(3):829–842.

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

Jarosz, A. F., and Wiley, J. 2014. What are the odds? a practical guide to computing and reporting bayes factors. *The Journal of Problem Solving* 7.

JASP Team, et al. 2016. Jasp. *Version 0.8. 0.0. software*.

Johnson, D. I.; Roloff, M. E.; and Riffee, M. A. 2004. Politeness theory and refusals of requests: Face threat as a function of expressed obstacles. *Communication Studies* 55(2).

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.

Kennedy, J.; Baxter, P.; and Belpaeme, T. 2014. Children comply with a robot's indirect requests. In *HRI*.

Li, J. 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* 77:23–37.

Lin, P.; Bekey, G.; and Abney, K. 2008. Autonomous military robotics: Risk, ethics, and design. Technical report, Cal. Poly. State Univ. San Luis Obispo.

Malle, B. F.; Scheutz, M.; Arnold, T.; Voiklis, J.; and Cusimano, C. 2015. Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In *Proceedings of HRI*, 117–124.

Nomura, T.; Uratani, T.; Kanda, T.; Matsumoto, K.; Kidokoro, H.; Suehiro, Y.; and Yamada, S. 2015. Why do children abuse robots? In *HRI Extended Abstracts*, 63–64.

Raman, V.; Lignos, C.; Finucane, C.; C. T. Lee, K.; Marcus, M.; and Kress-Gazit, H. 2013. Sorry dave, i'm afraid i can't do that: Explaining unachievable robot tasks using natural language. In *Proceedings of RSS*.

Scassellati, B.; Admoni, H.; and Mataric, M. 2012. Robots for use in autism research. *Annual Review of Biomedical Engineering* 14:275–294.

Sharkey, N., and Sharkey, A. 2010. The crying shame of robot nannies: an ethical appraisal. *Interaction Studies* 11(2):161–190.

Simmons, R.; Makatchev, M.; Kirby, R.; Lee, M. K.; et al. 2011. Believable robot characters. *AI Magazine* 32(4).

Stewart, N.; Chandler, J.; and Paolacci, G. 2017. Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences*.

Tanaka, K.; Nakanishi, H.; and Ishiguro, H. 2014. Comparing video, avatar, and robot mediated communication: Pros and cons of embodiment. In *Proceedings of ICCT*, 96–110.

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

Wada, K., and Shibata, T. 2007. Living with seal robots – its sociopsychological and physiological influences on the elderly at a care house. *IEEE Transactions on Robotics* 23(5):972–980.

Wen, J.; Stewart, A.; Billinghurst, M.; Dey, A.; Tossell, C.; and Finomore, V. 2018. He who hesitates is lost (...in thoughts over a robot). In *Proceedings of TechMindSociety*.

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

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