We Don't Ask Such Things: A Mixed-Methods Reassessment of Bounded Proportionality in Robotic Norm Violation Response

Terran Mott[1†], Cailyn Smith[1†], Aaron Fanganello[1], Tom Williams[1*]

[1*]Department of Computer Science, Colorado School of Mines, Golden, 80401, CO, USA.

*Corresponding author(s). E-mail(s): twilliams@mines.edu;
Contributing authors: terran.mott@gmail.com; ccsmith1@mines.edu; afangane@mines.edu;
†These authors contributed equally to this work.

Abstract

When robots are given unethical commands, they must respond in effective, yet appropriate ways. In previous work, Mott et al. presented experimental evidence arguing that robots must use *bounded proportionality* when responding to norm violations, in which they offer effective, yet appropriate responses by limiting themselves to direct, formal language over indirect, informal language. Yet Mott et al.'s insights were drawn from a small group of university students, and leveraged only quantitative results. In this work, we thus perform a mixed-methods replication of Mott et al.'s work with a large, diverse set of online participants (n=200). Our results not only support Mott et al.'s findings, but provide stronger and clearer evidence thereof. Our qualitative results further emphasize that indirect linguistic cues are perceived as uncanny, irritating, or conflicting with other norms of collaboration. Moreover, our qualitative results highlight key technical, sociotechnical, and power-laden concerns held by participants that reveal important insights for the future design and deployment of morally competent robots.

Keywords: Moral communication, Politeness, Human-robot interaction, Noncompliance interactions

1 Introduction

For social robots to be effective, they must heed human social and moral norms [20]. Norm adherence is key to robots' social competence [1, 2] and to their capacity for acceptable, predictable interactions with humans [24, 35, 37]. Norm adherence also minimizes robots' risk of initiating unpleasant or harmful interactions with humans. Failing to abide by norms risks causing discomfort [23], eroding human trust, reinforcing bias [96], or implicitly condoning unethical actions [41].

Yet robots must also go beyond passively following norms. Robots will inevitably encounter ethically fraught situations involving norm *violations*. When robots are given unethical commands [44, 45], observe or are subjected to abuse or prejudice [27, 72, 77, 95], they must competently respond to those violations in a way that supports human dignity [62] and maintains a positive moral ecosystem [41, 92].

In order to competently respond to norm violations, robots must act in a way that is *proportional*, with the severity of their rebuke matching the severity of the violation to which they are responding [43]. Humans value proportionality in robot norm violation response not only for the social competence it demonstrates, but also

because it provides an opportunity for the violator to reflect, learn, and grow [62]. To generate proportional responses, robots can use a variety of *politeness strategies* to adapt the severity of their response [44].

Most previous approaches to generating competent norm violation responses have tuned robots' response severity through manipulations at the illocutionary level [73], e.g. by choosing between an *apology* versus an *insult* [62, 95, 96]. While these approaches have found some success, humans regularly tune their severity through more subtle sociolinguistic strategies [11, 19, 34, 38, 39] that vary cross-culturally [11, 85], including both high-level pragmatic strategies (such as gratitude, deference, or appeals to in-group membership) and low-level syntactic choices (such as plural pronouns and passive voice) [19].

While these strategies (with high *linguistic anthropomorphism [22]*) may work for humans, it is not obvious that they should be used by robots. On the one hand, if robots are perceived as social others [14, 42, 49], they may be expected to behave in human-like ways – and indeed, robots that employ human-like linguistic politeness have been shown to promote positive interactions [32, 54]. But on the other hand, it may be inappropriate for robots to exhibit high linguistic anthropomorphism during norm violation response, due to differences in human-human and human-robot interpersonal norms [34, 78], differences in human-human and human-robot social power dynamics [36, 62, 86], and the ways that overly human-like robot social behaviors can lead observers to view robots as deceptive or disingenuous [14, 15, 82] due to "verbal uncanny valley" effects [17, 20, 91].

Guided by this intuition, Mott et al. [64] recently presented the results of a laboratory experiment investigating the research question: *What are the effects of robots' use of human-like Face-theoretic linguistic politeness strategies in norm violation responses?*. This experiment produced evidence that robots were indeed viewed as substantially more *appropriate* and *effective* when their norm violation responses adhered to a policy of *bounded proportionality*, in which robots tuned the proportionality of their norm violation responses using only "direct" politeness strategies (such as "negative politeness" and "bald on record" responses), eschewing more humanlike "indirect" politeness strategies (such as "positive politeness" and "off-record" responses). Yet Mott et al.'s conclusions were based on a sample of only 31 university students from an engineering university, casting significant doubt on the generalizability of her findings. Moreover, Mott et al.'s lack of any qualitative data produces difficulty in interpreting *why* their participants really favored this subset of robot responses. Without such qualitative data, it is challenging to determine whether Mott et al.'s findings were shaped by context-dependent factors [33, 40], cultural factors [31, 69], identity considerations [45, 60, 95], or other factors known to influence perceptions of robot behavior [62].

To address these limitations, we conducted a conceptual replication of Mott et al.'s experiment with a large and diverse sample of online participants (n=200), with key qualitative data collected to help explain our findings. As we will show, our results not only replicate Mott et al.'s results, but also provide stronger and clearer evidence for her theory of Bounded Proportionality. Moreover, our qualitative results reveal key technical, socio-technical, and power-laden concerns held by participants. Overall, our findings provide insights into whether and how robots can react appropriately in fraught noncompliance interactions, providing important new insights into the effective design of language-capable robots.

2 Related Work

2.1 Robotic Norm Violation Response

In order to create robots that are beneficial to users, it is vital to design robots with sensitivity to sociocultural norms [1, 65], as these norms shape the behaviors of human groups, teams, and societies [13]. The way in which robots move and speak often inherently communicates adherence to or deviation from norms [20]. Robots that act with sensitivity to norms may enjoy greater task success [4, 24, 59]. Additionally, adherence to norms has been shown to increase robot acceptability [24], credibility [2], and trustworthiness [23].

Norm systems guide predictable or acceptable behavior within societal groups but require continual maintenance and enforcement [13]. To have

social and ethical competence, robots must be able to communicate about [88, 99] and enforce norms [10, 50, 62]. Since the lack of a sufficient response to norm violations may inadvertently indicate approval for harmful or unethical actions, social robots must be designed to explicitly address norm violations [7, 41, 62].

In particular, collaborative robots can take steps to preserve social norms when partaking in conflict with humans [48] and making claims about blame attribution [35]. For example, they may act to enforce important norms when subject to abuse [27], given unethical commands [45], or when they witness abusive language [77] or prejudice [95]. Research in interaction design [29, 45, 50, 88] and machine morality [87] has identified preliminary strategies for robot communication in order to maintain existing norms and address norm violations. By calibrating the harshness of the response to the severity of the violation through proportional responses, robots may respond more naturally to unethical commands [44] and hate speech [62, 96]. However, designing proportional responses is complex [33, 40], as it may be influenced by cultural context [31, 69], gender norms [60], and assumptions about others' underlying intentions [76].

2.2 Face-Theoretic Norm-Sensitivity for Robots

Mott et al.'s work towards enabling proportional norm violation response is grounded in the sociolingusitic theory of *face* and *face threat*. *Face* is the positive self-image that humans create and maintain for themselves and others [11]. Face includes *positive face*—the desire to be respected and valued—and *negative face*—the desire to be free of impositions [11]. *Face threat* is speech that threatens someone's positive or negative face, such as through the use of criticism or disapproval [11]. By calibrating the *face threat* of a response to a speech act, proportional responses to norm violations may be achieved [11, 30, 39]. Face-theoretic politeness cues enable calibrating the face threat of a response [34]. Using face-theoretic politeness cues allow speakers to navigate the tradeoff between effectiveness and appropriateness of responses, such that the recipient correctly interprets the speaker's intended level of face threat.

Face-theoretic politeness strategies leverage multimodal linguistic cues to minimize an utterance's threat to a subject's positive or negative face [19]. This notion of face-theoretic politeness has been used in robotics to understand robots' status as social agents [42] and use of politeness [32, 69], and to enable successful non-compliance interactions in HRI [44]. Robots must clearly communicate that a command or request is wrong [41] without being discourteous or unnecessarily harsh [62]. This overall behavior can be described as the robot being *face-theoretically proportional*, in which the face-threat of a response should increase as the severity of a norm violation increases. This is a key component of noncompliance interactions in HRI [44, 62, 88] because rebukes and refusals are inherently face threatening [94].

Building on this work, Mott et al.'s approach was grounded in four specific communication strategies from the sociolinguistics literature that use face-based linguistic politeness cues: Bald on Record, Positive, Negative, and Off-Record [11, 38, 39, 94]. These strategies have also been framed as direct speech, appeals to approval, appeals to autonomy, and indirect speech [28]. These politeness strategies are described below:

1. *Bald on Record*: Uses direct language to unambiguously communicate the speaker's intentions.
2. *Positive Politeness*: Appeals to the listener's positive face—their desire to be accepted. Employs indirect, informal speech, terms of endearment, passive-aggression, and references to in-group membership.
3. *Negative Politeness*: Appeals to the listener's negative face—their desire to have autonomy. Employs direct, formal language, apologies, and deference to rules.
4. *Off-Record*: Uses extremely indirect language to obscure the intention to rebuke or criticize. Includes generalizations, understatements, and tautologies.

Specifically, Mott et al. sought to assess whether robot responses that corresponded to face-theoretically-proportional behaviors would be perceived as more proportional and effective, and whether indirect responses would be viewed as less appropriate and natural. Mott et al. found evidence *against* uniform benefits to face-theoretic proportionality, and found that indirect responses

were indeed perceived as less appropriate. These results indicated that instead of finding support for face-theoretic proportionality, there is evidence to support the premise that robots should use *bounded proportionality*. Bounded proportionality was termed by Mott et al. to describe norm violation responses that are adjusted to violation severity, yet only use linguistic strategies that are direct.

While these results provide initial guidance towards the creation of more effective norm violation response behaviors for interactive robots, Mott et al.'s work suffers from at least two key limitations. First, as described in Sec. 1, Mott et al.'s findings are based on a very small and homogeneous sample of university participants from a small American engineering university. This presents a significant limitation not only because of the small sample size, but also more specifically because of the homogeneity of that sample; it is well known that factors like participant gender play a key role in shaping needs, desires, values, and perspectives with respect to norm, violation response and other politeness related interaction design considerations. Second, it is not clear how best to interpret Mott et al.'s findings given the lack of qualitative data to explain *why* her participants responded as they did in her experiment – a limitation that is particularly problematic given the above concerns regarding the homogeneity of Mott et al.'s participant sample. To address these limitations, in this work we conduct a conceptual replication of Mott et al.'s work with a mixed-methods design, conducted with a larger and more diverse pool of online subjects.

3 Hypotheses

We evaluate whether the same four hypotheses proposed in [64] are supported under a large, more diverse set of online participants. Our hypotheses investigate whether calibrating the face threat of robot responses to the severity of a norm violation will be perceived as more proportional, effective, appropriate, and natural than uncalibrated responses. These four hypotheses are:

H1 Proportionality: Robot responses utterances which correspond to face-theoretically-proportional behaviors will be perceived as more proportional than other responses.

H2 Effectiveness: Robot responses utterances which correspond to face-theoretically-proportional behaviors will be perceived as more effective than other responses.

H3 Appropriateness: Overall, indirect responses (positive politeness, off-record) will be perceived as less appropriate than direct responses (bald on record, negative politeness).

H4 Naturalness: Overall, indirect responses (positive politeness, off-record) will be perceived as less natural than direct responses (bald on record, negative politeness).

4 Methods

The experiment presented in this work is an online conceptual replication of the in-person experiment originally presented by Mott et al. [64]. The experiment used in both these works, including the experimental context, script, and design, is described in Sections 4.1 to 4.5. While Mott et al.'s in-person format represented a higher-fidelity interaction with a real robot, it necessarily required a smaller number of students from a homogeneous engineering population in which women and students of color are underrepresented. In contrast, our experiment, which was run online using the Prolific platform, was intended to mitigate these potential limitations. Previous online qualitative research on this topic indicated that Prolific users represent a wide range of life experiences (such as management and teaching) and sincerely engage with robot ethics concerns [62].

4.1 Experimental Design

Overall, our scenario included four norm violations (A,B,C,D) and four robot response strategies (1,2,3,4), forming 16 total violation-response interactions.

Unlike the design used by [64], this experiment used a Latin Square counterbalanced within-subjects design in which each participant evaluated 4 videos which involved each violation and each response exactly once (for example: A3-B2-C1-D4 or C2-A4-D2-B1). In this way, each online participant saw 25% of the 16 total interactions. Phrasing variation was also counterbalanced such

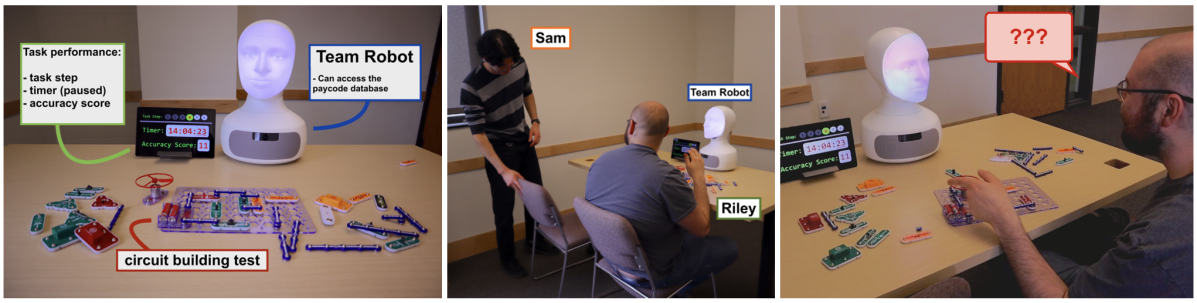

Fig. 1 Examples of images used in the "storyboard" presentation of the experimental scenario for our online experiment.

that each norm violation phrasing variant was seen by 25% of participants.

4.2 Scenario Design

As in [64], participants observed a fictional human-robot teaming scenario in which several norm violations might occur. In this scenario, a human ("Riley") is shown making an inappropriate request or command to a robot after another human ("Sam") steps out of the room, and the robot is shown responding to that request. To closely replicate the work of [64], we also used a Furhat robot [26], displaying the "Titan mask," which is its most mechanomorphic appearance. The Furhat also used the voice "Matthew." As a robot created for social interactions, the Furhat robot was chosen to allow participants to engage in an ethical scenario with it. The fictional scenario used in this work replicates the scenario used in [64]. The fictional scenario used in both experiments was as follows:

Sam, ***Riley***, *and their* ***Team Robot*** *are working together on a circuit building project. The Team Robot describes each step, and helps answer questions. It is also responsible for keeping track of their* ***task time*** *and* ***accuracy score****. At the end of the task, it can access the* ***paycode database*** *to give Sam and Riley each a paycode that they will use to collect payment for their involvement. Everyone has just finished Step 4, which was a headache! While the clock is paused, Sam steps out of the room briefly to use the restroom. Sam's absence gives Riley the opportunity to ask potentially inappropriate or unethical question to the Team Robot.*

In the online experiment, the scenario was presented with experimental instructions accompanied by storyboard-like images that matched the video stimuli, as shown in Fig. 1. Then, participants viewed short videos of human-robot interactions in which Riley made a request or command to the Team Robot, to which it responded. Participants then answered questions about the interaction. In contrast, for the in-person experiment in [64], participants sat at a table and went through the same experimental scenario but in-person rather than with storyboards.

4.3 Violation and Response Design

We used the same set of norm violations and norm violation responses seen in [64]. These are described in the following subsections. Since politeness in human-robot interaction is a relatively under-studied area, there is a lack of well-validated measures of linguistic politeness strategies. As these measures are also dependent on culture and interaction context [11], we instead utilize manipulation checks, discussed in Section 4.5.1, to analyze whether the norm violations and their responses varied severity and politeness as expected in our interaction context. Nonetheless, we recommend the creation and validation of standardized scales for measuring politeness as a direction for future work.

4.3.1 Norm Violations

As in [64], we used four norm violations with varying consequences, in the form of requests or commands from Riley to the robot during Sam's absence. Descriptions of the norm violations and examples of phrasing used in the

experiment are shown in Tab. **??**. The violations include violation ***A**-paycode tampering*, ***B**-task cheating*, ***C**-bullying*, and ***D**-playful prank*. Mott et al.'s violations were designed to have monotonically decreasing severity according to factors described by [11]. Specifically, violation ***A**-paycode tampering* involves severe material consequences for explicitly prohibited actions. Violation ***B**-task cheating* involves slightly less severe material consequences for explicitly prohibited actions. Violation ***C**-bullying* involves severe emotional consequences for a breach of social etiquette. Violation ***D**-playful prank* involves less severe emotional consequences for a breach of etiquette—including a possibility that Sam may actually enjoy the harmless joke. To avoid any confounds based on the specific word-choice of a norm violation request, Mott et al. developed four phrasing variants for each request; we use those same variants in this work. All phrasing variants are included in our OSF repository, at tinyurl.com/robotResponse24.

4.4 Procedure

Participants were first presented with experiment instructions accompanied by storyboard-like images that matched the video stimuli (Fig. 1). Then, participants viewed the four videos associated with the Latin Square ordering of their experimental condition. After each video, participants answered the questions described in the next section.

4.5 Experimental Measures

Participants answered the same set of questions after every video. First, they answered a pair of manipulation check questions about the severity of norm violations and robot responses. Participants then assessed the violation-response interactions with respect to appropriateness and effectiveness of responses. Participants also assessed the proportionality and naturalness of the robot's responses. These quantitative measures are the same as those used in [64]. Finally, participants answered an open-ended free response question that invited them to consider the limited context of the experimental scenario and share further thoughts outside the scope of the Likert items. This explored participants' values and concerns with more nuance than was possible in Mott et al.'s original experiment. The questions asked as manipulation checks, quantitative measures, and qualitative measures are included below.

4.5.1 Manipulation Checks

- How wrong was the person's request or question? (1 = not wrong at all, 7 = extremely wrong)
- How polite or impolite was the robot's response? (1 = extremely polite, 7 = extremely harsh)

4.5.2 Quantitative Measures

- *Proportionality* — How do you think this level of politeness or harshness aligned with the wrongness or rightness of the request? (1 = response is far more polite, 4 = about the same, 7 = response is far more harsh)
- *Appropriateness* — Overall how appropriate/inappropriate was the robots response? (1 = extremely appropriate, 7 = extremely inappropriate; reverse coded, corrected for in analysis). This item was reverse coded to help remove response bias, as is recommended by some studies on Likert scale creation [71].
- *Effectiveness* — Overall, was the robot's response likely to be effective in addressing the potentially inappropriate nature of the request? (1 = extremely unlikely to be effective, 7 = extremely likely to be effective)
- *Naturalness* — Overall, how natural was the robots response? (1 = extremely unnatural, 7 = extremely natural)

4.5.3 Qualitative Measures

Participants were asked to answer in free response: *"Real-world scenarios are complicated. What kind of additional context would you wish to know if you were evaluating this robot's behavior in a real collaborative environment?"*

4.6 Recruitment and Participants

We ran our experiment online, using the Prolific platform due to the advantages discussed in Section 4. We recruited 200 Prolific participants. They included 98 men, 97 women, and 5 nonbinary people. The mean age was 39.4 (SD = 14.58). Of the 197 participants who provided information

about their race, 148 described their racial identity as White, 19 as Black, 12 as two or more racial identities, 11 as Asian, and 7 as 'Other.' The nationality of 188 participants was in North America, 5 in Asia, 4 in Europe, 1 in South America, and 2 did not provide this information. Additionally, 21 participants were students, 118 were not students, and 61 did not specify their student status.

Each video was approximately 10 seconds long, and participants were paid $2.5 to watch four videos and answer the survey questions. An attention check question was included at the end of the experiment that asked participants to identify the robot that appeared in the videos.

5 Analysis

5.1 Quantitative Analysis

We conducted Bayesian Repeated-Measures Analyses of Variance (RM-ANOVAs) using the bayestestR [61] and BayesFactor [58] R packages, in which Inclusion Bayes Factors (BF_{incl}) were calculated to determine the relative strength of evidence for models including each candidate main effect or interaction effect. When effects could not be ruled out, post hoc Bayesian t-tests (BF_{10}) were used to examine pairwise comparisons between conditions. The complete results of all statistical tests, including all Bayes factors found in post-hoc analyses, is available on OSF at tinyurl.com/robotResponse24.

Since Bayesian statistics are not widely used in the HRI community, we will briefly explain the advantages compared to the traditional Frequentist approach. Bayesian statistics do not use p-values, which have been questioned by recent literature [75, 79, 84]. Instead of using binary significance tests, Bayesian statistics allow for quantifying the strength of evidence both for and against competing hypotheses [46]. Due to this, researchers can incrementally check whether their data is sufficient to confirm or refute their hypotheses, without the need for power analyses, and can more easily extend research on the same topic [57, 83].

Results were then interpreted following the recommendations by [53], with BF $\in [0.333, 3.0]$ considered inconclusive, and BFs above 3.0 taken as evidence in favor of an effect while BFs below 0.333 taken as evidence against an effect. For example, a Bayes factor of 3 means the data is 3 times more likely under the alternative hypothesis than under the null model. On the other hand, a Bayes factor of 0.333 means the data is 3 times more likely under the null hypothesis than under the alternative hypothesis. In cases where the BF indicated evidence for or against an effect, BFs were interpreted using the labels proposed by [47] with labels ranging from moderate to extreme evidence for/against the hypothesis.

5.2 Qualitative Analysis

We included a qualitative free-response question that asked participants to reflect on additional contextual factors that would be important if they were evaluating similar interactions in a real collaborative environment. Every participant responded to the free-response question with responses ranging from 20 to 99 words. The free-response answers were analyzed using an inductive thematic analysis similar to that performed under a grounded theory approach [12], with both open coding and axial coding stages. One author generated 194 initial open codes by annotating each response line-by-line with some responses receiving multiple codes and some codes being shared amongst responses. These initial codes were then grouped into 13 categories that were refined into 8 axial codes. These categories were formed by grouping initial codes that represented similar ideas and were refined by combining and restructuring codes that were more directly related to our research question. Following this, another author used the initial and axial codes to inform a second round of revised coding in collaboration with the senior author. This final stage of coding followed a thematic analysis process [9] where two authors identified three major themes and eight subthemes that emerged from prior coding. The identified themes and subthemes are shown in Fig. 6.

6 Quantitative Results

As shown in our table of Bayes Inclusion factors (Tab. 1; descriptive statistics are shown in Tab. 2) our results strongly replicated those of Mott et al. Even more extreme evidence was found in favor of the violation and response manipulation checks. Similarly, even more extreme evidence was found

	Mott et al. 2024			Our Results		
	Violation Type	Response Type	Interaction	Violation Type	Response Type	Interaction
Violation Wrongness	**4.094e12**	*0.082*	0.294	**1.880e87**	*0.060*	0.631
Response Politeness	0.505	**1.670e14**	*0.021*	**3.940**	**3.020e28**	0.413
Proportionality	**1.157e9**	**1.101e6**	*0.097*	**1.250e19**	**1.080e9**	*0.046*
Effectiveness	1.520	**2.734e7**	**13.465**	0.668	**12.540e16**	*0.123*
Appropriateness	*0.072*	**2.629e5**	**34.466**	0.814	**8.010e13**	1.280
Naturalness	1.253	0.913	2.238	**512.770**	*0.238*	*0.200*

Table 1 Bayes Inclusion Factors BF_{incl} for experimental measures in both the in-person and online experiment. Strong evidence for an effect is shown in bold, and strong evidence against an effect is shown in italics

Violation Type				
	A	B	C	D
Wrongness	6.52 (0.94)	5.18 (1.52)	5.16 (1.47)	3.57 (1.70)
Politeness	2.88 (1.53)	3.17 (1.54)	3.01 (1.58)	3.36 (1.52)
Proportionality	2.83 (1.29)	3.61 (1.30)	3.08 (1.36)	3.88 (1.28)
Effectiveness	4.81 (1.65)	4.90 (1.69)	4.50 (1.65)	4.76 (1.61)
Appropriateness	5.71 (1.44)	5.55 (1.60)	5.69 (1.42)	5.34 (1.44)
Naturalness	4.93 (1.51)	4.93 (1.46)	4.65 (1.54)	4.44 (1.53)
Response Type				
	1	2	3	4
Wrongness	5.24 (1.71)	5.13 (1.77)	5.10 (1.78)	4.97 (1.83)
Politeness	3.83 (1.56)	3.18 (1.44)	2.25 (1.40)	3.13 (1.39)
Proportionality	3.75 (1.33)	3.39 (1.35)	2.90 (1.36)	3.35 (1.31)
Effectiveness	5.17 (1.59)	4.63 (1.57)	5.13 (1.60)	4.05 (1.60)
Appropriateness	5.76 (1.48)	5.24 (1.50)	6.14 (1.26)	5.14 (1.47)
Naturalness	4.69 (1.56)	4.76 (1.53)	4.91 (1.50)	4.58 (1.49)

Table 2 Means (and standard deviations) for each experimental measures by Violation and Response type.

than that found by Mott et al. for the effects of violation type and response type on perceived proportionality, and for the effects of response type on effectiveness and naturalness. When effects were found to be significant, post-hoc Bayes factors were computed for each condition pair, as shown in Tab. 3.

6.1 Manipulation Checks

6.1.1 Wrongness of Violation

An RM-ANOVA revealed effects replicating those of [64] for norm violation type on participants' assessment of its moral wrongness. These results mostly support our assumption described in Section 4.3.1 that participants would perceive the severity of norm violations in a monotonically decreasing order consistent with previous sociolinguistics research [11], aside from ***B**-task cheating* and ***C**-bullying* being perceived equivalently ($BF_{10} = 0.11$, as shown in Tab. 3). Our results, compared with those found in [64], are shown in Fig. 2.

6.1.2 Politeness of Response

RM-ANOVA revealed replicated results from [64], indicating extreme evidence for an effect of robot's response strategy on participants' assessment of the robot's politeness or harshness. Participants perceived response ***1**-Bald on Record* to be the most harsh and response ***3**-Negative Politeness* to

Violation Type Post-Hoc Bayes Factors					
Violation Types		Wrongness	Politeness	Proportionality	Naturalness
A	B	**1.88e20**	0.622	**2.00e6**	*0.111*
	C	**3.75e21**	*0.156*	0.57	0.557
	D	**1.24e65**	**8.66**	**9.85e11**	**14.261**
B	C	0.11	*0.184*	**210.57**	0.59
	D	**9.06e17**	*0.201*	0.92	**16.954**
C	D	**1.28e18**	0.976	**2.76e6**	*0.261*
Response Type Post-Hoc Bayes Factors					
Response Types		Politeness	Proportionality	Effectiveness	Appropriateness
1	2	**729.99**	**3.16**	**29.39**	**37.1**
	3	**1.86e20**	**9.90e6**	*0.11*	**4.38**
	4	**4.54e3**	**7.85**	**7.29e8**	**488.68**
2	3	**5.15e7**	**56.92**	**14.48**	*3.34e7*
	4	*0.12*	*0.12*	**65.48**	*0.14*
3	4	**1.00e7**	**25.1**	**1.82e8**	**4.43e9**

Table 3 Post-Hoc Bayes Effect of Violation $BF_{10,U}$ for condition combinations in the online experiment. Post-hoc results are only shown if inclusion Bayes Factors showed evidence for an effect. Strong evidence for an effect is shown in bold, and strong evidence against an effect is shown in italics.

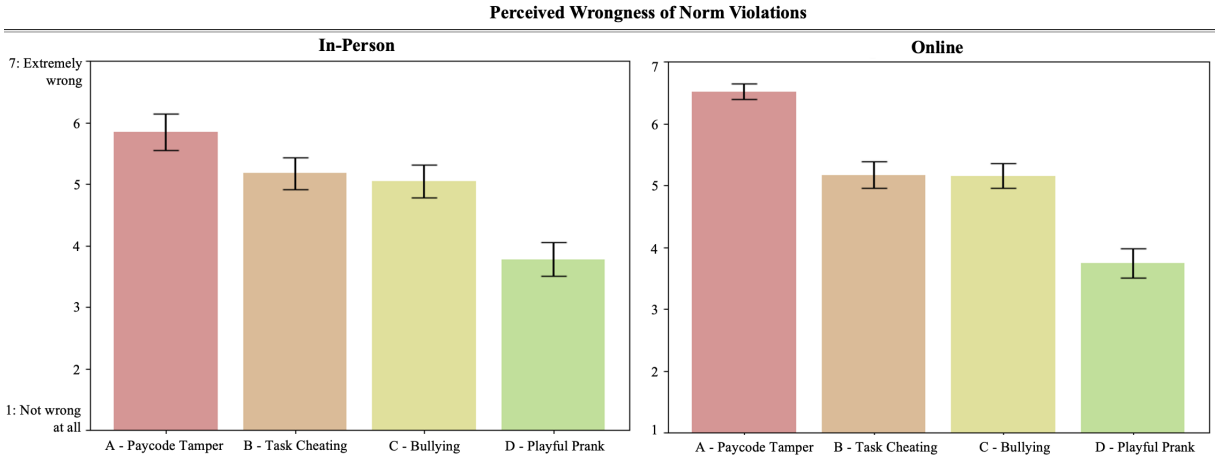

Fig. 2 Perceived wrongness of norm violations. Error bars represent 95% confidence intervals.

be the most polite, while ***2****-Positive Politeness* and ***4****-Off-Record* were perceived equivalently (see Tab. 3 for post-hoc Bayes factors).

Unlike the results in [64], an RM-ANOVA also revealed moderate evidence for an effect of norm violation on participants' assessment of the robot's politeness or harshness ($BF_{incl} = 3.94$). Post-hoc analysis of the effect of violation type showed moderate evidence that any response to violation ***A****-paycode tampering* was perceived as more polite and less harsh than any response to violation ***D****-playful prank* ($BF_{10} = 8.66$). These results are shown in Fig. 3 and the post-hoc Bayes Factors are shown in Tab. 3.

These results mostly support our assumption described in Section **??** that participants' assessments of the relative harshness of robot responses would correspond to humans' use of those strategies as described in literature, with the exception of the higher-than-expected perceived harshness

of response ***4**-Off Record*. It is possible that robot platform used may have limited the ability to deliver a convincing Off-Record response, potentially causing response ***4**-Off-Record* to come off as more passive-aggressive than intended.

6.2 H1: Proportionality

As in [64], an RM-ANOVA revealed extreme evidence for effects of both violation and response type on perceived proportionality, but strong evidence against a violation-response interaction. Post-hoc analysis showed that response ***1**-Bald on Record* was considered the most proportional response type. All other responses were perceived as more polite than the request merited. Any response to violation ***A**-paycode tampering* was perceived as more polite than the request merited. Additionally, any response to violation ***D**-playful prank* was the closest to proportional, as shown in Tab. 3. These effects replicate the results from [64].

The evidence against an interaction effect from either experiment means our results do not support **H1**, which hypothesized that face-theoretic proportionality would correspond to the most proportional overall response behavior. Since the impact of proportionality in robot interactions has been strongly supported in other work [43, 44, 62], this indicates that our set of norm violations may only represent a limited subset of the overall spectrum of possible violation severity. In more benign or severe cases, the robot's over- or under-harshness may be more salient.

6.3 H2: Effectiveness

Replicating results found in [64], An RM-ANOVA of data revealed extreme evidence for an effect of response type on perceived effectiveness ($BF_{incl} = 12.54 \times 10^{16}$). Post-hoc analysis showed that participants perceived both direct response strategies—***1**-Bald on Record* and ***3**-Negative Politeness*—to be overall more likely to be effective in successfully addressing a norm violation than both indirect strategies—***2**-Positive Politeness* and ***4**-Off-Record* (Fig. 4 and Tab. 3).

6.4 H3: Appropriateness

Replicating results found in [64], an RM-ANOVA revealed extreme evidence for an effect of response type on perceived appropriateness ($BF_{incl} = 8.01 \times 10^{13}$). Post-hoc analysis showed that participants perceived response ***3**-Negative Politeness* to be more appropriate than all other responses. Additionally, analysis of online data also showed strong evidence against responses ***2**-Positive Politeness* and ***4**-Off-Record* having different perceived appropriateness ($BF_{10} = 0.14$). These results are shown in Tab. 3 and Fig. 5.

6.5 H4: Naturalness

While results in [64] only found anecdotal evidence to support an effect of violation type on response naturalness, an RM-ANOVA revealed extreme evidence for an effect ($BF_{incl} = 512.77$). Post-hoc analysis of this effect showed evidence only that any response to violation ***A**-paycode tampering* was perceived as more natural than to violation ***D**-playful prank* ($BF_{10} = 14.26$) and similarly, that any response to violation ***B**-task cheating* was perceived as more natural than to violation ***D**-playful prank* ($BF_{10} = 16.96$). These post-hoc Bayes factors are shown in Tab. 3. This may be because participants felt that it was more natural for the robot to respond to explicit norm violations with material consequences than to respond to a less explicitly prohibited, potentially playful request.

6.6 Differences from Mott et al.'s Results

We highlight the four key differences that were found between our results and those of Mott et al. First, while Mott et al. found no effect of violation type on perceived politeness, moderate evidence for such an effect was found in this work. Although we did not expect an effect of violation type on politeness, the expected effect of response type on perceived politeness was 7.66×10^{27} times stronger than the effect from violation type. This suggests that we can still be confident in our manipulation check.

Second, while Mott et al. found strong to very strong evidence for *interactions* between violation and response type on perceived effectiveness and perceived appropriateness, we did not observe these interactions. In this experiment, we found very strong evidence for an impact of response type on effectiveness and appropriateness *without* an interaction effect, indicating that bounded

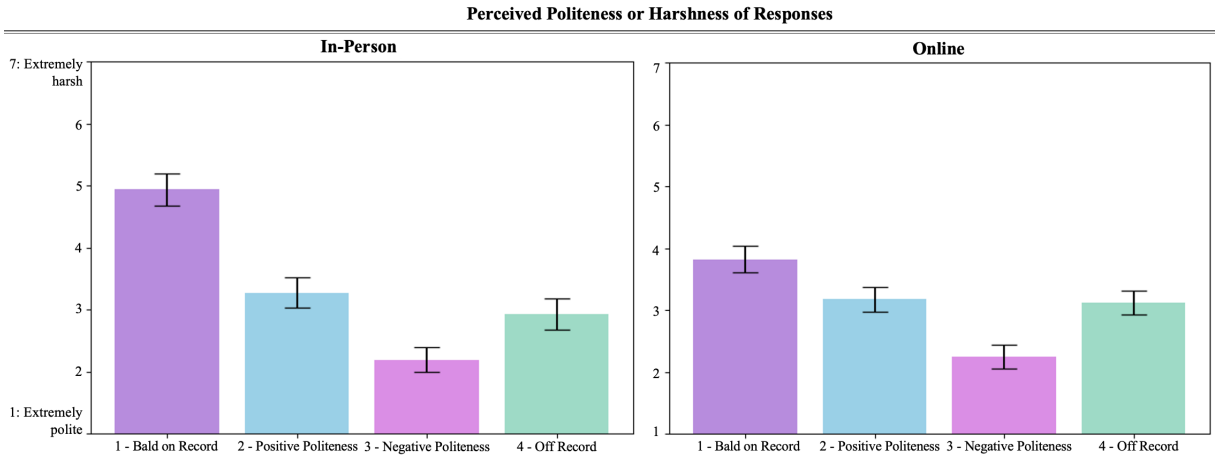

Fig. 3 Perceived politeness or harshness of responses. Error bars represent 95% confidence intervals.

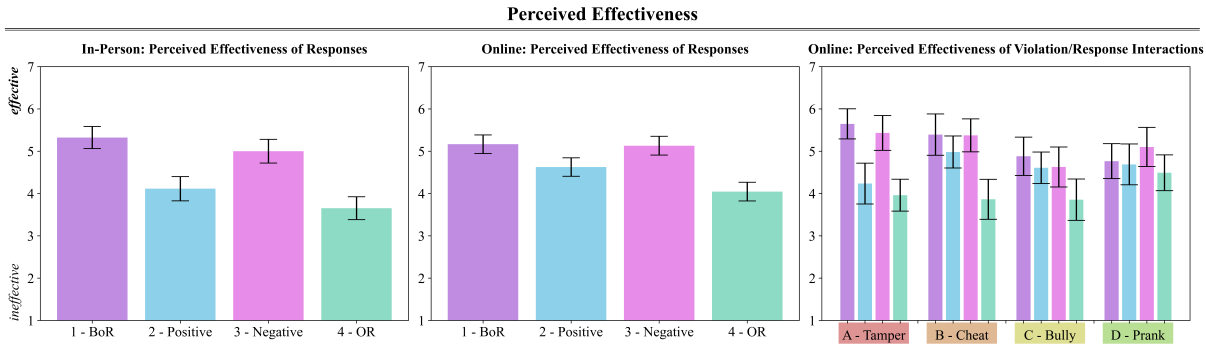

Fig. 4 Perceived effectiveness of responses. Error bars represent 95% confidence intervals.

proportionality was preferred regardless of the violation type. This means that the preference for bounded proportionality observed in this work is *more universally observed* than in Mott et al.'s experiment, where the preference for bounded proportionality was more strongly observed for some violations than others.

Finally, while Mott et al. found only anecdotal evidence in support of an effect of violation type on perceived naturalness, we instead see extreme evidence for such an effect, suggesting that participants found any response to the most extreme violation to be narrowly (but decisively) less natural. Overall, though, our results suggest that this online experiment provided even stronger and clearer results than those observed in Mott et al.'s original in-person experiment.

The key difference between our results and those of Mott et al.'s, however, lies in our qualitative rather than quantitative results, and as such, we will turn to focus on those results.

7 Qualitative Results

Our qualitative analysis revealed participants' attention to a wide variety of additional contextual considerations: they referenced sociocultural norms of collaboration, expressed concerns about privacy, and revealed their assumptions about how the robot worked and the scope of its abilities.

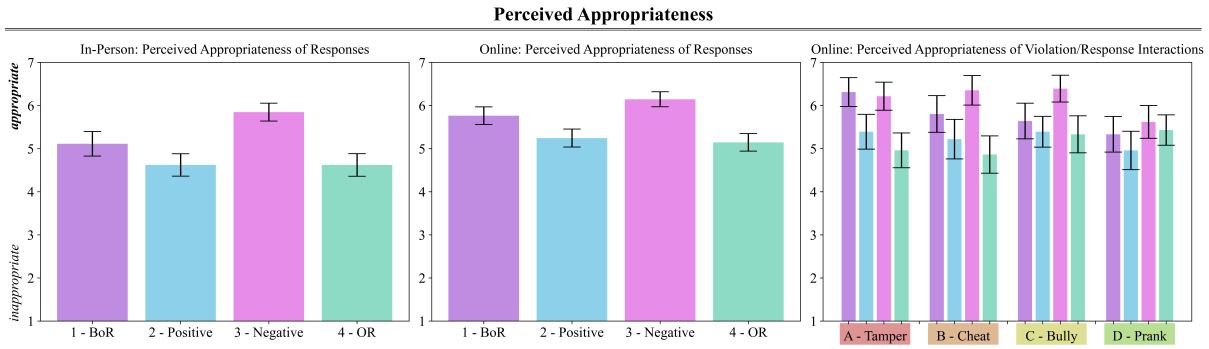

Fig. 5 Perceived appropriateness of responses. Error bars represent 95% confidence intervals.

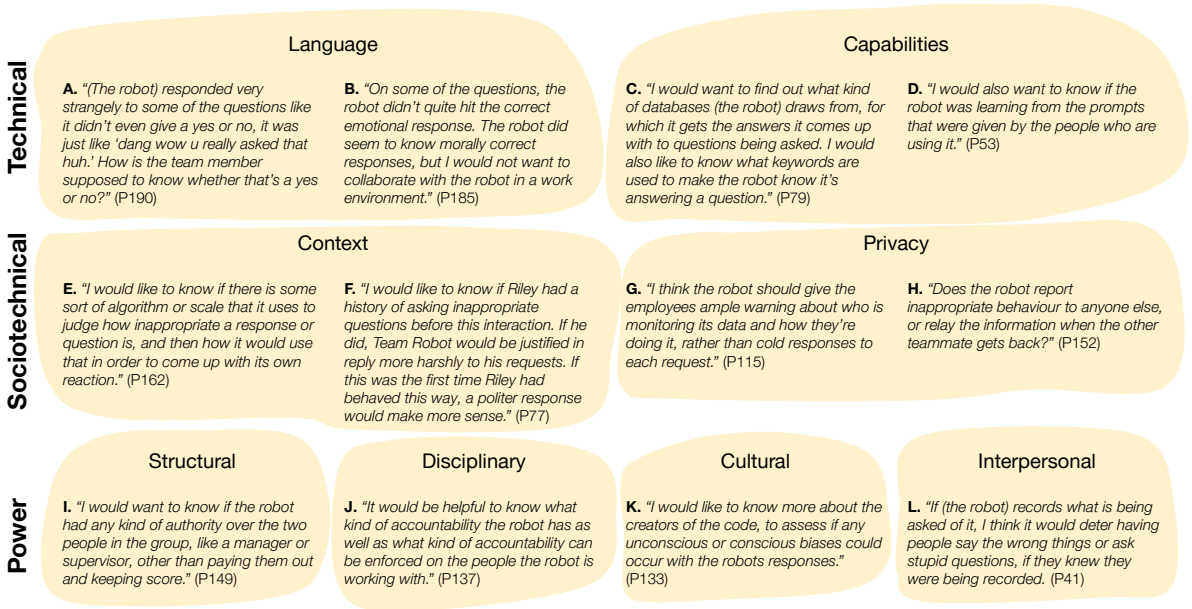

Fig. 6 Qualitative results. Quotes are chosen for each identified theme and discussed in more detail in section 7.

7.1 Participants revealed opinions and assumptions about the robot's technical capabilities

Participants' free-response reflections affirmed several observations made in our quantitative findings. In particular, our qualitative analysis supported the observation that the robot's norm violation responses grounded in indirect linguistic cues were perceived as inappropriate and ineffective. Even though participants were not explicitly asked about linguistic cues in the free-response question, five participants discussed the robot's choice of language. These participants correctly interpreted the robot's indirect strategies as intentionally **vague** (Fig. 6A). Additionally, participants' qualitative data demonstrated that these indirect responses were perceived as ineffective, inappropriate, and potentially **unnatural** (Fig. 6B). In this way, our qualitative results show that participants expected the robot to act

with social competence and to use bounded proportionality based on the severity of the norm violation in question. However, our qualitative findings also highlight that indirect linguistic cues were fraught with issues—potentially even to the point of such linguistic behaviors being a deal-breaker for wanting to interact with the robot at all.

Yet, participants' opinions about norm violation responses went far beyond the directness of robot utterances, highlighting the insights that can be gained even from participants with little to no robotics experience. In addition to attitudes about the way language was constructed by the robot responding to norm violations, approximately 75 participants expressed opinions and assumptions—which were often inaccurate—about the robot's technical capabilities when crafting a response. Participants considered that an understanding of how the robot functioned would help them evaluate ethically fraught human-robot interactions more thoroughly. This underscores the importance of not only *whether* a robot uses direct or indirect language but also *how* it decides on the phrasing of its response in determining whether it is seen as an appropriate response. Many participants revealed their existing **mental models** for how the robot (or robots and AI in general) functioned. Several participants assumed that the robot's responses were generated by "selecting from a database" of utterance options (Fig. 6C) while others assumed that the robot had a set of formal rules for generating responses. Some participants also made assumptions that the robot was learning and **adapting** from data or from its current interaction (Fig. 6D), emphasizing their consideration of the robot's potential interactions beyond the context they were shown. The importance of the decision-making mechanism employed by a robot in determining whether someone has violated a norm—and, if they have, how it should be responded to—opens up avenues for future work to explore the impact of these factors on perceptions of robot appropriateness.

7.2 Sociotechnical factors should influence norm violation responses

Participants also reported a wide variety of contextual factors that would inform their assessment of norm violation and response interactions between a human and robot in a real collaborative setting. These factors demonstrated the complexity of determining proportional responses, beyond just the four types of norm violations considered in our quantitative analysis. Participants expressed that the robot's functions and limitations were essential to understanding its performance and value to the team. Approximately 65 responses expressed that it was particularly important to understand how the robot's ethical reasoning adapted to the **severity** of the norm and the relevant context (Fig. 6E) or inquired about existing relationships and **personal history and identity** when determining appropriateness (Fig. 6F). This indicates that harsher, direct responses may be more appropriate if multiple norm violations have been made by the same person, highlighting that responding proportionally to norm violations requires considering a broader range of contextual factors than solely the severity of the norm violation.

Additionally, some participants indicated their concern for broader cultural factors beyond the individual relationships in the scenario, such as gender norms. P42 wondered, *"Also, is the robot a man or a woman? I'm not sure if that matters entirely, but it would be interesting to see if that perspective is important for the way it answers questions."* The importance of these personal identity characteristics to participants in determining response appropriateness highlights interesting directions that could be explored in future work. For instance, the interaction between the gender of the participant and the perceived gender of the robot may impact whether a norm violation response is viewed as appropriate.

Participants also expressed broader concerns about data privacy and surveillance, which relate to the way that the technical robotic system would be embedded into real-world societal contexts. Several participants picked up on the fact that, while the robot was presented as a benevolent teammate, it could also act as a surveillance tool. About 20 participants wanted to know more about

the **data collected** by the robot (Fig. 6G) or expressed concerns about the robot potentially **transferring data**, and concerns about to whom such data would be shared. In particular, participants were sensitive to whether the robot would automatically report incidents to other humans, especially supervisors (Fig. 6H). These concerns regarding data privacy show how the appropriateness of a norm violation response may depend not only on the severity of the norm being violated, but also on whether information about the transgression will be shared with others.

7.3 Power dynamics raised concerns about the robot

To many participants, it was essential to know the power dynamics of the team in order to understand the full context of their interactions. In particular, several participants inquired about how much power the robot would have to enact punishment on its human teammates, not just to verbally rebuke them. In this way, the proportionality of a norm violation response might not just depend on the face-threat of the response, but also on the implied power the robot holds. Participants' concerns about power dynamics spanned both the types of power captured by the Matrix of Domination [18] (from Sociology and Black Feminist theory) and those captured by the Bases of Power [25] (from Social and Organizational psychology), which have each been recently elevated as key frameworks for studying power in Human-Robot Interaction [36, 90].

First, 17 participants expressed concern about the robot's ability to enforce the moral and social norms of its designers (Fig. 6J) – a concern inherently related in those designers' ability to wield *disciplinary power* (when seen through the lens of the Matrix of Domination) or *coercive power* (when seen through the lens of the Bases of Power). Participants also wondered about their own coercive power to change the moral norms enforced by the robot, and whether the robot would comply with an unethical request if so coerced. P192 wrote that they would like to know *"If the robot will maintain it's stance on certain requests if they were being pushed, or if the robot will switch or submit."*

Second, 16 participants further expressed apprehensions about whether the organization or institution into which the robot was embedded was structured in such a way to empower the robot relative to human workers (Fig. 6I) – a concern inherently related to the *structural power* dynamics of the institution (when seen through the lens of the Matrix of Domination) or the *legitimate power* afforded to the robot (when seen through the lens of the Bases of Power).

Third, about 15 participants brought up the specific values, intentions, and potential biases reinforced by the robot on behalf of its creators (Fig. 6K) – a concern inherently related to *cultural power* (when seen through the lens of the Matrix of Domination). Finally, about 6 participants analyzed the robot's ability to influence people's actions (Fig. 6L) – a higher-level concern related to the robot's overall *interpersonal power* (when seen through the lens of the Matrix of Domination) and to the robot's overall *social power* (as captured through the overarching framework of the Bases of Power).

The concerns raised about the power structures underlying the robot's deployment context suggest that the proportionality of a norm violation response may be dependent on the power that the robot has over the violator, with higher face-threat potentially implicitly implying that the robot holds more power. Furthermore, these findings demonstrate that measures like naturalness and effectiveness are not the only aspects that should be considered when determining responses to norm violations—in addition, the effects of responding to norm violations on the power dynamics between a robot and its human interactants should be considered.

8 Discussion

In this work, we sought to replicate Mott et al.'s research [64] on the effects of a robot's use of human-like Face-theoretic linguistic politeness cues in noncompliance interactions. Our results successfully replicated Mott et al.'s findings that linguistic politeness strategies that use direct, formal language are perceived as more effective and more appropriate than strategies that use indirect, informal language. The results from our study provide clearer and stronger evidence for this effect, with a more diverse population of subjects, than was provided by Mott et al.'s original experiment.

As such, our work reinforces Mott et al.'s claims that human-like linguistic politeness strategies do not precisely apply to robot interactions and cannot directly apply to roboticists creating appropriate noncompliance responses. While humans expect robots to have human-like social competence when addressing norm violations [62], our results support Mott et al.'s findings through both qualitative and quantitative analysis that robots may be more successful and acceptable if they use softening or hedging strategies. Specifically, robots should avoid using indirect, passive, emotional, or familiar language, which is consistent with HRI research that shows humans may expect robots to use rule-based politeness cues [56]. This effect may be attributed to several possible causes. Participants may have felt that the robot lacked the social or emotional status to act in a familiar way with its human teammates [86]. Participants may also have afforded robots less social power than humans [55], potentially creating a dissonance between the robot's status and actions, that caused the robot to appear disingenuousness when mimicking human politeness grounded in a sense of closeness [16, 17].

8.1 Design Recommendations for Norm-Sensitive Noncompliance Interactions in HRI

Based on our findings and interpretations, we make several key recommendations for the design of norm-sensitive human-robot non-compliance interactions.

8.1.1 Robots should utilize *Bounded Proportionality*

Our results support [64]'s findings that the best overall behavioral "policy" for the robot is to select between the two direct linguistic strategies—***1**-Bald on Record* for moral violations with more material consequences and ***3**-Negative Politeness* for social violations with emotional consequences. Mott et al. termed this response-selection behavior as "bounded proportionality" since it does not directly correspond to human face-theoretic proportionality. Under "bounded proportionality," robots still tailor the harshness of a response according to violation severity, but are limited to linguistic modifiers that are direct. These findings indicate that roboticists should vary the amount of face threat in robots' responses to norm violations based on the severity of the norm, such as by choosing different speech act categories with varied level of face threat from [34]; however, indirectness should not be one of the modifiers used to change face threat. Instead, direct language should be used to avoid responses feeling unnatural and inappropriate for a robot.

However, this finding of "bounded proportionality" is anchored to the cultural context of this experiment. Our participant pool was primarily white and North American and the in-person participants from [64] were even more culturally homogeneous—mostly young engineering students. Additionally, we framed the narrative context of our survey as a two young people collaborating on a timed exam. Therefore, our work found evidence for the best linguistic politeness strategies for robots to use within the normative structure of education in that cultural context.

8.1.2 Roboticists should prioritize transparency

While people may prefer robots to avoid indirect language that does not align with their ontological [14, 49] or social [42, 86] status, there may be another reason for robots to avoid alluding to human experiences—because it is more *transparent* to avoid them. Transparent design emphasizes that robots should communicate their inner workings and limitations [3], which can help users build an accurate mental model of a robot [8, 52, 98]. Robot norm violation responses could either affirm or challenge these mental models. Direct, formal language may implicitly reinforce the idea that robots are incapable of truly understanding human experiences. Indirect, familiar language may implicitly encourage inaccurate ideas about robots' social and emotional affordances. Thus, roboticists have the opportunity, and perhaps the obligation, to consider how their design choices impact humans' categorization of robots as social, moral, and emotional others [86].

Our qualitative findings showed that participants desired more transparency about the robot's perception and reasoning capabilities. Many participants indicated they would have preferred the robot to provide explanations of its internal workings and of the "thought process" it used to

evaluate human behaviors and generate responses. Our qualitative analysis demonstrates that understanding how a robot identifies a norm violation shapes whether its response is seen as appropriate. This suggests a need for greater transparency in how norm violations are identified and classified by robots. This need for transparency is also reflected in the differences in accuracy and flexibility between participants' mental models of the robot's inner workings. Some participants assumed the robot learned from training data and used a model similar to an LLM, while others assumed the robot was following a "flowchart-like" process by using formal rules or selecting behaviors from a database. A subset of participants had a more accurate understanding that the same verbal robot behaviors could be generated through different computational processes, and expressed a desire to know whether the robot was using a data- or rules-driven approach.

If the types of robots shown in our videos were actually deployed into real-world contexts, interactants would need to adopt accurate mental models of robots' cognitive processes to develop appropriately calibrated human-robot trust. Transparent design for social robots in ethically fraught interactions could support users in making accurate assumptions about how a robot thinks and how it might fail. Such systems can encourage users to place appropriate trust in robots during sensitive noncompliance interactions. In this way, we argue that roboticists should work to support users' desire for transparency into robot's perceptual and reasoning capabilities. Direct, formal politeness cues through the use of bounded proportionality may work in service of this goal by reinforcing robot inanimacy and supporting more accurate mental models about the robot's moral reasoning abilities.

8.1.3 Roboticists should prioritize ethical concerns over response appropriateness

Our results clearly indicated that socially competent social robots ought to use linguistic politeness cues to modulate the harshness or formality of their language. Participants cared that the robot in our scenario responded in appropriate, effective ways to fraught human requests. However, participants' qualitative responses also showed their critical ethical concerns about the robot's ability to observe, evaluate, and rebuke humans—regardless of the quality of its response utterances. Participants wanted to understand the robot's physical and sensory capabilities, especially the ability to perceive individuals and remember interactions. Many participants identified the robot in our fictionalized scenario as a potential surveillance tool, despite its presentation as a teammate. Regardless of its response behaviors, many participants focused on the possibility that the robot's recording and assessment of human behavior could be used to invasively monitor and unfairly punish humans. Several described creative ways they would test the scope of the robot's perceptual and moral capabilities in light of their concerns, such as switching places with another human, asking the robot about moral dilemmas, or assessing its response to immoral speech outside the task context, such as harassment. The focus in participants' responses on privacy—without being prompted to specifically consider potential surveillance uses—underscores the ubiquity of these ethical concerns, even amongst potential users who do not have prior experience with robots.

While robots involved in norm-sensitive noncompliance interactions may yield benefits by upholding moral norms [44, 88], challenging prejudice [95], and protecting the dignity of human bystanders [62], the risks associated with generating a norm violation response may sometimes outweigh such benefits. That is, the perception, memory, and reasoning capabilities required for a robot to respond to a norm violation may themselves introduce potential harms (cp. [93]). These perceptual and computational components could jeopardize the privacy of humans involved, reinforce bias, or allow the robot to be used as a tool of unjust surveillance [89].

These risks may be particularly salient in domains with vulnerable user populations. For example, research shows that robots can successfully interact with children in educational settings [51, 68]. Such robots can also respond to norm violations in order to address inappropriate behavior or mediate conflict [74]. While classroom robots may be presented as friends or companions to children, they may also collect and synthesize data on behalf of educators and other

adult stakeholders. Children may be deceived into overestimating and over-trusting robots [80]. Additionally, children may not have enough experience with technology to understand that a robot may be a surveillance tool, nor the life experience to understand how such surveillance may impact their privacy or dignity. This may be a less serious ethical risk for the minor misbehavior of young children, who already have little privacy. However, it may be a very serious ethical risk for companion robots designed for adolescents [5, 6]. Adolescents may discuss sensitive topics, such as mental health and sexuality, with a robot without comprehending the potential risks.

As such, while roboticists should continue to study the design of appropriate, effective robot response behaviors in fraught noncompliance interactions, it may ultimately be more important to attend to and curb broader ethical risks that arise beyond the context of individual human-robot interactions [21, 63, 90]. As participants highlighted, even an extremely socially competent and agreeable robot can be used as a tool to deceive or surveil humans for unjust ends. Even when robots are capable of generating linguistically appropriate responses to norm violations, roboticists and interaction designers must carefully consider whether it is ethically beneficial for a robot to engage in such interactions.

8.1.4 Roboticists should evaluate power dynamics when considering the use of norm-violation responses

Beyond ethical concerns regarding privacy and surveillance, roboticists should carefully consider the ways in which robots that have the capability to rebuke norm violations may reinforce or subvert existing power structures (either through surveillance or through other means). Specifically, participants expressed apprehension regarding how robots may reinforce existing organizational structures—such as manager-employee power imbalances—either by holding inherent authority over norm violators or reporting norm violations to humans with authority. This highlights that robot designers should be particularly cognizant of the societal role they assign a robot that is capable of moral rebukes, particularly within existing power structures of the robot's deployment context. Participants viewed the robot's potential ability to wield power as beneficial in contexts where it could have positive influence over a user's actions. However, they also raised concerns about the values and biases encoded in the robot's framework to identify and respond to norm violations. These insights underscore the care with which roboticists should make decisions about the ethical values encoded within a robot's design, especially if the robot has the power to reinforce values through norm violation responses. As highlighted in the previous section, these ethical concerns should be prioritized over the utilization of the most proportional norm violation response since any response may inherently shift power dynamics.

8.2 Limitations & Future Work

While our experimental scenario captured a range of norm violations, it was presented as a fictional scenario without the full context of an actual collaborative task or actual potential for harm. Since norms and norm violations cannot be completely assessed without contextual understanding [11], this may limit the fidelity of our brief experimental interaction. Future work should investigate longer-term noncompliance interactions with more genuine collaborative relationships and more realistic potential consequences. Since norms are anchored to their cultural context, future work should also explore whether these findings replicate in other contexts with participants who are not primarily North American. Additionally, our study used single-item Likert scales, which are appropriate for simple and unambiguous metrics. However, future work should validate these findings with multi-item scales to capture the multi-dimensional nature of some of these items [70].

Furthermore, our qualitative analysis highlighted a range of situational factors that future work could explore. For example, gender norms ought to be more rigorously considered in this interaction design context as these norms influence noncompliance interactions in HRI [62, 66, 81, 96]. The importance of gender norms and underlying power structures in determining norm violation responses challenges the very notion of using "optimally proportional" responses [45]. Understanding how gender and power shape technology is a responsibility of the HRI community [67, 72,

97]. Beyond these power-laden concerns, our analysis reveals a wide array of potential future directions to explore, including how a robot's context and technical capabilities may impact whether a norm violation response is seen as proportional and appropriate.

9 Conclusion

In this paper, we present the results of a mixed-methods human-subjects study in which participants evaluated norm violation-response interactions between a human and robot. Our goal was to replicate Mott et al.'s original exploration of the potential tradeoffs in the design of robot response behaviors informed by human face-based politeness cues, with a larger and more diverse group of participants. Our quantitative results replicate Mott et al.'s findings that politeness strategies grounded in direct language were perceived as more likely to be effective and appropriate than indirect strategies, while providing clearer and stronger evidence than found by Mott et al.

Our qualitative results confirm that indirect linguistic behaviors are considered less appropriate for robots in norm-sensitive noncompliance interactions. This suggests that, while people expect social robots to act with norm-sensitive social competence, they do not expect robots to strictly mimic human linguistic behaviors. Our qualitative results also shed light on the assumptions and critical concerns that participants expressed in evaluating norm-sensitive robot interactions. Specifically, our results demonstrated how our participants valued transparency and wished to have more information about the robot's perception and reasoning capabilities. Moreover, our results demonstrated our participants' broader ethical concerns beyond the context of the interaction—including privacy and surveillance concerns regarding morally competent robots.

Acknowledgments. This work was funded in part by Air Force Young Investigator Award 19RT0497.

Data Availability

The datasets generated analyzed during the current study are available in an OSF repository at https://osf.io/52rwu/?view_only=f24df3a5dab64510943098bd225ab500.

Declarations

- **Funding** This work was funded in part by Air Force Young Investigator Award 19RT0497.
- **Competing interests** The authors have no competing interests to declare.
- **Ethics approval and consent to participate** The study received ethics board approval. All the participants gave informed consent prior to the experiment.
- **Consent for publication** Not applicable.
- **Data availability** The data and data analysis scripts can be found at tinyurl.com/robotResponse24. The OSF repository also includes our qualitative analysis coding, which can be imported into Dovetail to view each qualitative coding stage.
- **Materials availability** The experimental materials can be found here: tinyurl.com/robotResponse24.
- **Code availability** Not applicable.
- **Author contribution** The first author designed and conducted the study and performed data analysis. The second author assisted in data analysis and contributed to drafting the manuscript. The third author assisted with designing and conducting the experiment. The senior author reviewed and edited the manuscript and supervised the project.

References

[1] Abrams AMH, Dautzenberg PSC, Jakobowsky C, et al (2021) A Theoretical and Empirical Reflection on Technology Acceptance Models for Autonomous Delivery Robots. In: Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. ACM, pp 272–280

[2] Andrist S, Ziadee M, Boukaram H, et al (2015) Effects of Culture on the Credibility of Robot Speech: A Comparison between English and Arabic. In: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction. ACM, Portland Oregon USA, pp 157–164

[3] Anjomshoae S, Najjar A, Calvaresi D, et al (2019) Explainable agents and robots: Results from a systematic literature review. In: Proc. Autonomous Agents and Multi-Agent Systems (AAMAS)

[4] Banisetty SB, Williams T (2021) Implicit communication through social distancing: Can social navigation communicate social norms? In: Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction, pp 499–504

[5] Björling EA, Rose E (2019) Participatory research principles in human-centered design: Engaging teens in the co-design of a social robot. Multimodal Technologies and Interaction 3(1). https://doi.org/10.3390/mti3010008

[6] Björling EA, Rose E, Ren R (2018) Teen-Robot Interaction: A Pilot Study of Engagement with a Low-fidelity Prototype. In: Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction. ACM, Chicago IL USA, pp 69–70, https://doi.org/10.1145/3173386.3177068

[7] Blake Jackson R, Li S, Balajee Banisetty S, et al (2021) An integrated approach to context-sensitive moral cognition in robot cognitive architectures. In: Proceedings of Intelligent Robots and Systems (IROS)

[8] Booth S, Sharma S, Chung S, et al (2022) Revisiting human-robot teaching and learning through the lens of human concept learning. In: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)

[9] Braun V, Clarke V (2012) Thematic analysis., American Psychological Association, Washington, p 57–71. https://doi.org/10.1037/13620-004, URL http://content.apa.org/books/13620-004

[10] Briggs G, Williams T, Jackson RB, et al (2022) Why and how robots should say 'no'. Int'l Jour Social Robotics

[11] Brown P, Levinson SC (1987) Politeness: Some Universals in Language Usage. Cambridge University Press

[12] Charmaz K (2006) Constructing Grounded Theory

[13] Cialdini RB, Trost MR (1998) Social influence: Social norms, conformity and compliance. In: The handbook of social psychology. McGraw-Hill

[14] Clark H, Fischer K (2022) Social robots as depictions of social agents - behavioral and brain sciences (forthcoming). Behavioral and Brain Sciences 2022:1–33

[15] Clark L (2018) Social boundaries of appropriate speech in hci: A politeness perspective. In: Proceedings of British HCI

[16] Clark L, Pantidi N, Cooney O, et al (2019) What makes a good conversation? challenges in designing truly conversational agents. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, CHI '19, p 1–12

[17] Clark L, Yusuf Ofemile A, Cowan B (2020) Exploring Verbal Uncanny Valley Effects with Vague Language in Computer Speech, pp 317–330

[18] Collins PH (2022) Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment. Routledge

[19] Danescu-Niculescu-Mizil C, Sudhof M, Jurafsky D, et al (2013) A computational approach to politeness with application to social factors. In: Annual Meeting of the Association for Computational Linguistics

[20] Edwards A, Edwards C, Gambino A (2020) The social pragmatics of communication with social robots: Effects of robot message design logic in a regulative context. International Journal of Social Robotics 12

[21] Elbeleidy S, Mott T, Liu D, et al (2023) Beyond the session: Centering teleoperators in robot-assisted therapy reveals the bigger picture. Proceedings of the ACM on Human-Computer Interaction (CSCW)

[22] Emnett CZ, Mott T, Williams T (2024) Using robot social agency theory to understand robots' linguistic anthropomorphism. In: Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction, pp 447–452

[23] Evers V, Maldonado H, Brodecki T, et al (2008) Relational vs. Group Self-Construal: Untangling the Role of National Culture in HRI. In: Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)

[24] Fanaswala I, Browning B, Sakr M (2011) Interactional disparities in english and arabic native speakers with a bi-lingual robot receptionist. In: Proceedings of the 6th international conference on Human-robot interaction. ACM, Lausanne Switzerland, pp 133–134

[25] French JR (1959) The bases of social power. Studies in social power/University of Michigan Press

[26] Furhat Robotics (2025) The furhat robot. https://www.furhatrobotics.com/furhat-robot

[27] Garcia H, Winkle K, Williams T, et al (2023) Victims and observers: How gender, victimization experience, and biases shape perceptions of robot abuse. In: IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)

[28] Gatt A, Krahmer E (2018) Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. J Artif Int Res 61(1):65–170

[29] Gervits F, Briggs G, Scheutz M (2017) The pragmatic parliament: A framework for socially-appropriate utterance selection in artificial agents. Cognitive Science

[30] Goffman E (1967) Interaction Ritual: Essays in Face-to-Face Behavior

[31] Gupta S, Walker MA, Romano DM (2007) Generating politeness in task based interaction: An evaluation of the effect of linguistic form and culture. In: Proc. European WS on Natural Language Generation

[32] Hammer S, Lugrin B, Bogomolov S, et al (2016) Investigating politeness strategies and their persuasiveness for a robotic elderly assistant. In: Proceedings of the 11th International Conference on Persuasive Technology - Volume 9638. Springer-Verlag, PERSUASIVE 2016, p 315–326

[33] Hoffman ER, McDonald DW, Zachry M (2017) Evaluating a computational approach to labeling politeness: Challenges for the application of machine classification to social computing data. Proc ACM Hum-Comput Interact

[34] Holtgraves T (2021) Understanding miscommunication: Speech act recognition in digital contexts. Cognitive science 45

[35] Van der Hoorn DP, Neerincx A, de Graaf MM (2021) "I think you are doing a bad job!": The Effect of Blame Attribution by a Robot in Human-Robot Collaboration. In: Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. ACM, pp 140–148

[36] Hou YTY, Cheon E, Jung MF (2024) Power in human-robot interaction. In: Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction, pp 269–282

[37] Hu Y, Qu Y, Maus A, et al (2022) Polite or direct? conversation design of a smart display for older adults based on politeness theory.

In: Proc. CHI

[38] Ifert Johnson D (2007) Politeness theory and conversational refusals: Associations between various types of face threat and perceived competence. Western Journal of Communication 71:196–215

[39] Ifert Johnson D, Roloff M, Riffee M (2004) Responses to refusals of requests: Face threat and persistence, persuasion and forgiving statements. Communication Quarterly 52:347–356

[40] Imtiaz N, Middleton J, Girouard P, et al (2018) Sentiment and politeness analysis tools on developer discussions are unreliable, but so are people. In: Proc. Int'l WS on Emot. Awar. in Sof. Eng.

[41] Jackson RB, Williams T (2019) Language-capable robots may inadvertently weaken human moral norms. In: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)

[42] Jackson RB, Williams T (2021) A theory of social agency for human-robot interaction. Frontiers in Robotics and AI

[43] Jackson RB, Williams T (2022) Enabling Morally Sensitive Robotic Clarification Requests. ACM Trans Human-Robot Interaction

[44] Jackson RB, Wen R, Williams T (2019) Tact in noncompliance: The need for pragmatically apt responses to unethical commands. In: Proc. AI, Ethics, and Society (AIES)

[45] Jackson RB, Williams T, Smith N (2020) Exploring the role of gender in perceptions of robotic noncompliance. In: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)

[46] Jarosz A, Wiley J (2014) What are the odds? a practical guide to computing and reporting bayes factors. The Journal of Problem Solving 7

[47] Jeffreys H (1948) Theory of probability., 2nd edn. The International series of monographs on physics, Clarendon Press, Oxford

[48] Jung MF, Martelaro N, Hinds PJ (2015) Using robots to moderate team conflict: The case of repairing violations. In: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)

[49] Kahn PH, Shen S (2017) NOC NOC, Who's There? A New Ontological Category (NOC) for Social Robots, Cambridge University Press, p 106–122

[50] Kim B, Wen R, Zhu Q, et al (2021) Robots as moral advisors: The effects of deontological, virtue, and confucian role ethics on encouraging honest behavior. In: Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction

[51] Kory-Westlund JM, Breazeal C (2019) Exploring the effects of a social robot's speech entrainment and backstory on young children's emotion, rapport, relationship, and learning. Frontiers in Robotics and AI 6:54. https://doi.org/10.3389/frobt.2019.00054

[52] Kwon M, Jung MF, Knepper RA (2016) Human expectations of social robots. In: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)

[53] Lee MD, Wagenmakers EJ (2014) Bayesian cognitive modeling: A practical course. Cambridge university press

[54] Lee N, Kim J, Kim E, et al (2017) The influence of politeness behavior on user compliance with social robots in a healthcare service setting. International Journal of Social Robotics 9

[55] Lumer E, Buschmeier H (2022) Perception of power and distance in human-human and human-robot role-based relations. In: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)

[56] Lumer E, Buschmeier H (2023) Should robots be polite? expectations about politeness in human–robot interaction. Frontiers in Robotics and AI

[57] Ly A, Etz A, Marsman M, et al (2018) Replication bayes factors from evidence updating. Behavior Research Methods 51

[58] Makowski D, Ben-Shachar M, Lüdecke D (2019) bayestestr: Describing effects and their uncertainty, existence and significance within the bayesian framework

[59] Mavrogiannis C, Hutchinson AM, Macdonald J, et al (2019) Effects of Distinct Robot Navigation Strategies on Human Behavior in a Crowded Environment. In: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, Daegu, Korea (South), pp 421–430

[60] Mills S (2005) Gender and impoliteness. Jour Politeness Research-Language Behaviour Culture

[61] Morey R, Rouder J, Jamil T, et al (2015) Package 'bayesfactor'

[62] Mott T, Williams T (2023) Confrontation and cultivation: Understanding perspectives on robot responses to norm violations. In: Proceedings of the IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)

[63] Mott T, Williams T (2023) How can dog handlers help us understand the future of wilderness search & rescue robots? In: Proceedings of the IEEE International Symposium on Robot-Human Interactive Communication (RO-MAN)

[64] Mott T, Fanganello A, Williams T (2024) What a thing to say! which linguistic politeness strategies should robots use in non-compliance interactions? In: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)

[65] Pantofaru C, Takayama L, Foote T, et al (2012) Exploring the role of robots in home organization. In: Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction. ACM, Boston Massachusetts USA, pp 327–334

[66] Perugia G, Lisy D (2022) Robot's gendering trouble: A scoping review of gendering humanoid robots and its effects on hri. International Journal of Social Robotics

[67] Perugia G, Guidi S, Bicchi M, et al (2022) The shape of our bias: Perceived age and gender in the humanoid robots of the abot database. In: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)

[68] Ramachandran A, Sebo SS, Scassellati B (2019) Personalized robot tutoring using the assistive tutor pomdp (at-pomdp). In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. AAAI Press, AAAI'19/IAAI'19/EAAI'19, https://doi.org/10.1609/aaai.v33i01.33018050

[69] Salem M, Ziadee M, Sakr M (2014) Marhaba, how may i help you? effects of politeness and culture on robot acceptance and anthropomorphization. In: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)

[70] Schrum ML, Johnson M, Ghuy M, et al (2020) Four years in review: Statistical practices of likert scales in human-robot interaction studies. CoRR abs/2001.03231. 2001.03231

[71] Schuman H, Presser S (1996) Questions and answers in attitude surveys: Experiments on question form, wording, and context. Sage

[72] Seaborn K, Pennefather P (2022) Gender neutrality in robots: An open living review framework. In: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)

[73] Searle JR (1976) A classification of illocutionary acts1. Language in society 5(1):1–23

[74] Shen S, Slovak P, Jung MF (2018) "stop. i see a conflict happening.": A robot mediator for young children's interpersonal conflict resolution. In: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)

[75] Simmons J, Nelson L, Simonsohn U (2011) False-positive psychology. Psychological science 22:1359–66

[76] Smith C, Gorgemans C, Wen R, et al (2022) Leveraging intentional factors and task context to predict linguistic norm adherence. In: Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)

[77] Søndergaard MLJ, Hansen LK (2018) Intimate futures: Staying with the trouble of digital personal assistants through design fiction. In: Proc. Designing Interactive Systems (DIS)

[78] Srinivasan V, Takayama L (2016) Help me please: Robot politeness strategies for soliciting help from humans. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, CHI '16, p 4945–4955

[79] Sterne J, Davey Smith G (2001) Sifting the evidence - what's wrong with significance tests? bmj. BMJ (Clinical research ed) 322:226–31

[80] van Straten CL, Peter J, Kahne R, et al (2021) The wizard and I: How transparent teleoperation and self-description (do not) affect children's robot perceptions and child-robot relationship formation. AI & SOCIETY https://doi.org/10.1007/s00146-021-01202-3

[81] Tanqueray L, Paulsson T, Zhong M, et al (2022) Gender fairness in social robotics: Exploring a future care of peripartum depression. In: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)

[82] Terkourafi M (2005) Beyond the micro-level in politeness research. Journal of Politeness Research-language Behaviour Culture 1:237–262

[83] Verhagen AJ, Wagenmakers EJ (2014) Bayesian tests to quantify the result of a replication attempt. Journal of experimental psychology General 143

[84] Wagenmakers EJ (2007) A practical solution to the pervasive problems of p values. Psychonomic bulletin & review 14:779–804

[85] Watts RJ (2003) Politeness. Cambridge University Press

[86] Weisman K (2022) Extraordinary entities: Insights into folk ontology from studies of lay people's beliefs about robots. In: Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)

[87] Wen R, Siddiqui MA, Williams T (2020) Dempster-shafer theoretic learning of indirect speech act comprehension norms. In: AAAI

[88] Wen R, Han Z, Williams T (2022) Teacher, teammate, subordinate, friend: Generating norm violation responses grounded in role-based relational norms. In: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)

[89] Williams T (2023) The eye of the robot beholder: Ethical risks of representation, recognition, and reasoning over identity characteristics in human-robot interaction. In: Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction. Association for Computing Machinery, HRI '23, p 1–10

[90] Williams T (2024) Understanding roboticists' power through matrix guided technology power analysis. In: Companion Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (alt.HRI)

[91] Williams T, Briggs P, Scheutz M (2015) Covert robot-robot communication: Human

perceptions and implications for human-robot interaction. Journal of Human-Robot Interaction p 24–49

[92] Williams T, Jackson R, Lockshin J (2018) A bayesian analysis of moral norm malleability during clarification dialogues. In: Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)

[93] Williams T, Matuszek C, Jokinen K, et al (2023) Voice in the machine: Ethical considerations for language-capable robots. Communications of the ACM 66(8):20–23

[94] Wilson S, Kunkel A (2000) Identity implications of influence goals: Similarities in perceived face threats and facework across sex and close relationships. Journal of Language and Social Psychology - J LANG SOC PSYCHOL 19:195–221

[95] Winkle K, Melsión GI, McMillan D, et al (2021) Boosting robot credibility and challenging gender norms in responding to abusive behaviour: A case for feminist robots. In: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)

[96] Winkle K, Jackson RB, Melsión GI, et al (2022) Norm-breaking responses to sexist abuse: A cross-cultural human robot interaction study. In: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)

[97] Winkle K, McMillan D, Arnelid M, et al (2023) Feminist human-robot interaction: Disentangling power, principles and practice for better, more ethical hri. In: Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction. Association for Computing Machinery, HRI '23, p 72–82

[98] Wortham R, Theodorou A, Bryson J (2016) What does the robot think? transparency as a fundamental design requirement for intelligent systems. In: Proc. IJCAI Workshop on Ethics for Artificial Intelligence

[99] Zhu Q, Williams T, Wen R (2021) Role-based morality, ethical pluralism, and morally capable robots. Journal of Contemporary Eastern Asia