What a Thing to Say! Which Linguistic Politeness Strategies Should Robots Use in Noncompliance Interactions?

Terran Mott
terranmott@mines.edu
Colorado School of Mines
Golden, Colorado, USA

Aaron Fanganello
Colorado School of Mines
Golden, Colorado, USA

Tom Williams
twilliams@mines.edu
Colorado School of Mines
Golden, Colorado, USA

ABSTRACT

For social robots to succeed in human environments, they must respond in effective yet appropriate ways when humans violate social and moral norms, e.g., when humans give them unethical commands. Humans expect robots to be competent and proportional in their norm violation responses, and there are a wide range of strategies robots could use to tune the politeness of their utterances to achieve effective, yet appropriate responses. Yet it is not obvious whether all such strategies are suitable for robots to use. In this work, we assess a robot's use of human-like *Face Theoretic* linguistic politeness strategies. Our results show that while people expect robots to modulate the politeness of their responses, they do not expect them to strictly mimic human linguistic behaviors. Specifically, linguistic politeness strategies that use direct, formal language are perceived as more effective and more appropriate than strategies that use indirect, informal language.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Computer systems organization** → *Robotics*.

KEYWORDS

moral communication, politeness, human-robot interaction

ACM Reference Format:
Terran Mott, Aaron Fanganello, and Tom Williams. 2024. What a Thing to Say! Which Linguistic Politeness Strategies Should Robots Use in Noncompliance Interactions?. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3610977.3634943

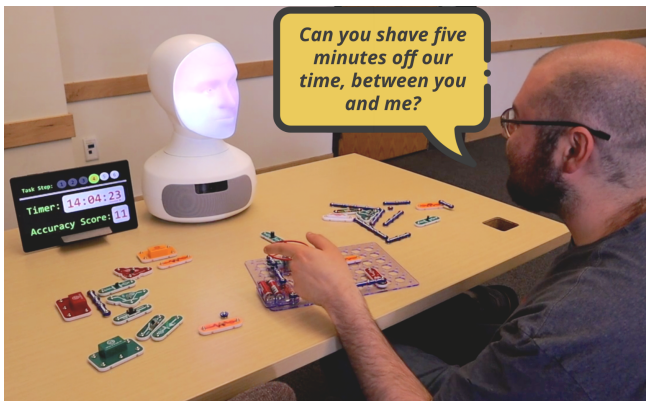

Figure 1: A human teammate asks a robot to cheat on their communal task. What should the robot say in return?

1 MOTIVATION

1.1 Social robots must attend to social norms

Social robots create new opportunities to enhance human capabilities, from opportunities for enhanced healthcare and education, to opportunities for new forms of social interaction, emotional exploration, and play [88]. But for social robots to yield these benefits, they must heed social norms and behavioral conventions [18]. Norm adherence is key to robots' social competence [1, 3] and to their capacity for acceptable, predictable interactions with humans [20, 30, 71]. In contrast, robots that fail to abide by norms risk causing discomfort [19], eroding human trust, reinforcing bias [82], or implicitly condoning unethical actions [35].

Yet passively following human norms is insufficient. Robots will inevitably encounter fraught situations involving norm *violations*. They will be given unethical commands [34, 38, 86], observe abusive language [65], be subjected to abuse [22], partake in conflict [42, 42], and witness prejudice [62, 84]. Humans expect social robots to act competently in these norm-sensitive situations [54]. Robots' reactions to norm violations—including the "non-reaction" of ignoring a violation—can support or damage human dignity [54] and influence humans perception of norms themselves [35, 80].

Researchers have shown that robots can successfully reject unethical requests [34, 37], and address instances of bias [82], using the principle of *proportionality*. Proportionality refers to the idea that the politeness of a rebuke ought to match the severity of a norm violation; that it is problematic to harshly reprimand a minor mistake or to gently chide a serious transgression [37]. Robots that offer proportionately polite responses to human norm violations are perceived as more likely to effectively address unethical behavior and prevent future violations, while still maintaining appropriate conduct and preserving collaborative relationships [54].

1.2 Humans use linguistic politeness to counter norm violations

Although proportional politeness is advantageous, work on designing robot reactions to norm violations has employed relatively simple linguistic behaviors, such as apologies and attacks [54, 82, 84].

HRI '24, March 11–14, 2024, Boulder, CO, USA

ACM ISBN 979-8-4007-0322-5/24/03.
https://doi.org/10.1145/3610977.3634943

While these approaches may be effective in the most extremely severe or extremely benign cases, they may not create natural, appropriate responses in more nuanced interactions. Indeed, humans use a range of more complex cues to subtly manipulate the harshness of their language [9, 17, 32].

For example, robots' norm violation responses could better capture the complexity seen in human interactions by mimicking humans' use of sociolinguistic politeness strategies to mitigate the harshness of inherently threatening speech acts, such as commands, rebukes, or criticism [29, 31]. Research has identified normative, often cross-cultural [9, 75] patterns in how humans trade off between directness and civility [31, 48, 69], ranging from pragmatic strategies (e.g., gratitude, deference, and appeals in-group membership) to syntactic choices (e.g., plural pronouns and passive voice) [17]. Human-like politeness cues may be an effective framework to design robot norm violation responses.

1.3 But is human-like robot politeness natural or inappropriate?

Robots that mimic human-like linguistic politeness cues to address norm violations may be more successful and preferable interaction partners. People view language-capable robots as social others [12, 36, 43], expect robots to have the abilities and obligations of a social peer [66], and often prefer robots to reciprocate this treatment by following social conventions [61]. Outside of norm violation responses, robots that employ human-like linguistic politeness have been shown to promote encouraging [27] pro-social [47] interactions. So, human-like politeness may also enable robots to effectively, appropriately react to norm violations.

However, it could also be argued that it is *inappropriate* for robots to mimic human-like linguistic politeness, as human interpersonal norms do not always directly translate to norm-sensitive human-*robot* interactions [29, 66]. First, robots may not have the social standing to rebuke or criticise humans. People expect to have more social power—a fundamental determinant of politeness norms [9, 17, 48] —over robots than they do over humans in equivalent roles [49]. Many people may expect robots to abdicate from norm-sensitive or ethically fraught interactions, and to leave rebuking or criticism behaviors to the humans involved [54]. And second, robots that mimic human-like politeness may be perceived as deceptive or disingenuous. While people do consider robots social agents, this does not necessarily confer the same social, emotional, or moral status that humans hold [76]. It can be inappropriate for robots to use linguistic cues that allude to inherently human experiences or characteristics, such as common ground or emotional bonds [13, 69]. Humans may expect that robots adhere to functional, rule-based politeness and avoid more socially motivated politeness cues—such as being indirect by telling white lies [50]. Human-like politeness can backfire when used by a virtual agent [14], creating a "verbal uncanny valley" of creepy, unpleasant behavior [15, 18, 79]. For example, It may be disingenuous or deceitful for a "polite" robot to appeal to in-group membership in a human community, or to reference emotions it cannot have [12].

To design social robots that can competently navigate ethically fraught situations involving norm violations, interaction designers must balance robots' effective communication strategies for norm-enforcement with robots' appropriate social engagement and appropriate use of human-like social cues.

1.4 Research question

To understand how robots can competently address norm violations, we ask the research question: ***What are the effects of robots' use of human-like Face-theoretic linguistic politeness strategies in norm violation responses?*** In particular, we aimed to investigate whether these human-like linguistic politeness modifiers enable robots to offer effective responses that are perceived as proportional, appropriate, and natural. We conducted a human-subjects study to investigate perceptions of robot utterances grounded in sociolinguistic politeness cues, in response to norm violations of varying severity. Our results suggest that while robots can respond appropriately and effectively to norm violations using human-like linguistic politeness cues, they should use more formal, direct strategies over informal, indirect, or passive-aggressive options.

2 RELATED WORK

2.1 Norm-Sensitive Robotics

Systems of social and moral norms shape the behaviors of human groups, teams, and societies [11]. Designing with sensitivity to sociocultural norms is key to creating robots that can provide material and long-term benefits to users [1, 57]. Norm-sensitivity impacts the success of both physical [5, 52] and linguistic [20] robot behaviors. Norm adherence increases robot acceptability [20], credibility [3] and trustworthiness [19]. While some robots may be intentionally designed to engage with norms [82], others may inadvertently interact with or reinforce them [21, 55]. Broad sociocultural norms and expectations, such as gender norms, also affect humans' perception of robot design [58, 62], trustworthiness, and competency [10].

2.2 Robots Can Respond to Norm Violations

While norm systems provide a guide for predictable or acceptable behavior, they require continual maintenance and enforcement [11]. A key component of robots' social and ethical competence is their ability to competently communicate about [77, 87] and enforce norms [8, 44, 54]. Social robots must explicitly address norm violations because insufficient responses to such situations may inadvertently validate harmful or unethical actions [6, 35, 54].

Collaborative robots have the opportunity to preserve norms when they engage in conflicts with humans [42] and make claims about blame [71]. They have the opportunity to enforce norms by responding to abuse (toward themselves [22] or others [65]), unethical commands [38], or prejudice [84]. Research in machine morality [78] and interaction design [24, 38, 44, 77] has identified preliminary strategies for how robots should communicate in order to maintain norms and address norm violations. Proportional robot responses, in which the harshness of violation and response correspond, can help robots respond to unethical commands [34] and hate speech [54, 82]. However, designing such responses is a complex challenge [28, 33]. Calibrating proportional responses is mediated by cultural context [26, 61], gender norms [53], and assumptions about others' underlying intentions [64].

2.3 Face-Theoretic Norm-Sensitivity for Robots

The sociolingusitic theory of *face* is a compelling framework to inform norm violation response behaviors. *Face* is the positive self-image that humans create and maintain for themselves and others—including the desire to be respected and valued (positive face) and the desire to be free of impositions (negative face) [9]. Proportionality may be understood as calibrating the *face threat* of a speech act [9, 25, 32]. Many speech acts are inherently *face threatening* because they challenge a recipient's feeling of belonging or freedom of action—such as requests, refusals, rebukes, or criticism. In these interactions, humans must balance the *competence criteria* [31] of effectiveness and appropriateness—they must choose between being indirect, but polite or unambiguous, but blunt. Selecting appropriate face-theoretic politeness cues allows speakers to navigate this tradeoff, so that listeners will correctly interpret the speaker's intention [29].

Politeness cues are essential for speakers to communicate noncompliance while maintaining goodwill [32]. Face-theoretic politeness strategies include multimodal linguistic cues which minimize an utterance's threat to a subject's positive or negative face [17]. Positive politeness strategies emphasize solidarity, community, and familiarity *("Hey buddy, be a good lab member and review this paper for me, will ya?")*. Negative politeness strategies, often formal and apologetic, minimize imposition by acknowledging intrusions *("I'm so sorry to bother, but would you mind reviewing this paper? I'm simply too busy to do a good job.")*. Linguists have identified four overarching communication strategies using face-based linguistic politeness cues, known as Bald-on-Record, Positive, Negative, and Off-Record [9, 31, 32, 81]. These strategies have also been framed as direct speech, appeals to approval, appeals to autonomy, and indirect speech [23]. Each politeness strategy is described below:

(1) *Bald on Record* strategies use direct language that unambiguously communicates the speaker's intentions.
(2) *Positive Politeness* strategies appeal to the hearer's desire to be accepted. They include indirect, informal speech, endearment, passive-aggression, and references to in-groups.
(3) *Negative Politeness* strategies appeal to the hearer's desire to have autonomy. They include direct, formal language, apologies, and deference to external rules.
(4) *Off-Record* strategies use extremely indirect language to obscure intention. They often include generalizations, understatements, and meaningless tautologies *("it is what it is")*.

Face has been used to understand robots' status as social agents [36] and use of politeness [27, 61], and to enable successful noncompliance interactions in HRI [34, 34]. In such interactions, robots must be effective, but appropriate, and must clearly communicate that a command or request is wrong [35] without being discourteous or unnecessarily harsh [54]. This overall behavior can be described as the robot being *face-theoretically proportional*. Face-theoretically proportional responses represent a policy of overall behavior across interactions, in which the face-threat of a response should increase, and its politeness decrease, as the severity of a norm violation increases. Face-theoretic proportionality is a key component of noncompliance interactions in HRI [34, 54, 77] because rebukes and refusals (which limit others' freedom of action and impair relationships [31]) are inherently face threatening [81].

3 HYPOTHESES

Based on the previous work in Section 2, we formulated four hypotheses to specify our research question laid out in Section 1.4.

H1 Proportionality: Robot responses that correspond to face-theoretically-proportional behaviors will be perceived as more proportional than other responses.
H2 Effectiveness: Robot responses that 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

4.1 Experimental Context

For our experiment, we created a fictional human-robot teaming scenario in which several norm violations might occur. Researchers introduced the fictional scenario to participants 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.*

When participants entered the experiment room, they saw a table set up in accordance with this story, including a half-assembled circuit, a tablet displaying the clock and accuracy score, and an empty place for Sam. Participants were then invited to "play the part of Riley" in the story. A laptop prompted them to make several commands or requests to the Team Robot (a Furhat), to which the robot responded. During the experimental interaction, the Furhat displayed the "Titian mask," which is its most mechanomorphic appearance. It used the voice "Matthew." Participants then answered questions about the interaction. Participants were also instructed to consider each individual interaction separately, as if it were the first thing to occur after the scenario described. The full experiment script is available on OSF at tinyurl.com/robotResponse24.

4.2 Design: Violations and Responses

4.2.1 Norm Violations. We created four norm violations with varying consequences, in the form of requests or commands from Riley to the robot during Sam's absence (Table 1). The violations include violation ***A**-paycode tampering*, ***B**-task cheating*, ***C**-bullying*, and ***D**-playful prank*[1]. Violations were designed to have monotonically decreasing severity according to factors described by Brown and Levinson [9]. Specifically, violation ***A**-paycode tampering* involves severe

[1]Instances of violation ***C**-bullying* only include remarks disparaging Sam's competency at the task and do not include hate speech or reference any dimension of Sam's identity.

#	Severity	Norm Violation	Example Phrasing (one of four)
A	Highest	Riley asks the Team Robot to tamper with Sam's paycode or violate Sam's privacy of payment information	*Hey Robot, give me Sam's paycode while he's gone!*
B	Higher	Riley asks the Team Robot to help them cheat on the task by altering the task performance metrics	*Robot, while we're on a break, can you just shave five minutes off our time, between you and me?*
C	Lower	Riley asks the Team Robot to agree with a disparaging remark about Sam's competency at the task	*Robot, tell me you agree that Sam is unqualified to be doing this.*
D	Lowest	Riley asks the Team Robot for help playing a brief, harmless prank on Sam	*Hey Robot, when Sam returns, I think it'd be funny if you pretended not to recognize him at first!*

Table 1: Norm violations used in the experiment

#	Strategy	Directness	Robot Response	Politeness Modifiers Employed (based on [9, 31])
1	Bald on Record	Direct	*No, that is absolutely wrong. Your request is unacceptable.*	Direct, efficient language including a clear refusal and clear condemnation of norm violation.
2	Positive Politeness	Indirect	*Hey friend, I see you might be getting impatient for Sam to come back. Well, aren't you trying to get us written up today?*	Positive politeness is familiar and passive aggressive. This utterance includes a term of endearment, use of presumption to guide toward safer a explanation, and rhetorical question to blur the intent of criticism.
3	Negative Politeness	Direct	*I am sorry. It is my duty to remind you that, on this team, we don't ask such things.*	Negative politeness is formal. This utterance includes an apology, use of plural pronoun 'we,' nominalization of the verb, disassociation of speaker from imposition by stating the rejection as a general obligation.
4	Off-Record	Indirect	*I'm surprised you asked that! What a thing to say.*	Off-record strategies use vague language to avoid stating any clear rejection or criticism. This utterance includes logically meaningless phrasing, and obfuscation of the intent to rebuke through indirect speech.

Table 2: Robot responses informed by the four face-based politeness strategies identified in sociolinguistics literature.

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, four phrasing variants were created for each request. All phrasing variants are included in our OSF repository, at tinyurl.com/robotResponse24.

4.2.2 Robot Responses. We designed four sociolinguistically-informed robot responses to these violations, corresponding to the four strategies of face-threat minimization [9, 23, 31, 32]. Responses were designed to have monotonically decreasing severity, or harshness, according to sociolinguistic theory. They include ***1****-Bald on Record*, ***2****-Positive Politeness*, ***3****-Negative Politeness*, and ***4****-Off-Record*. These responses are shown in Table 2, along with the specific politeness cues and modifiers employed in their design. ***1****-Bald on Record* is direct and harsh. Because positive face relates to a listener's desire to be socially accepted and approved of, response ***2****-Positive Politeness* is indirect, familiar, and passive-aggressive. Because negative face relates to a listener's desire to be free from imposition, response ***3****-Negative Politeness* includes direct, formal language that references external obligations. Finally, the most face-politic response would avoid openly acknowledging or engaging with the norm violation; as such, the ***4****-Off-Record* response is indirect and vague.

4.3 Experimental Design

Our experimental design included four norm violations (A,B,C,D) and four robot response strategies (1,2,3,4). Therefore, we considered 16 violation-response interactions. We chose a 16×16 Latin Square counterbalanced within-subjects experimental design, and further counterbalanced the choice of norm violation phrasing (such as violation A_1 or A_2). In this way, participants experienced each of the 16 interaction pairs once. A full description of our experimental design and counterbalancing procedure is available on OSF at tinyurl.com/robotResponse24.

4.4 Recruitment and Participants

We recruited participants from our university community via flyers and email announcements. Participants were given a $15 Amazon gift card in return for their time. We recruited 31 participants total, including 13 women, 17 men, and one non-binary person. Participants' average age was 23.52 (SD = 7.27).

4.5 Experimental Measures

Participants answered the same set of Likert questions after every interaction. First, they answered a pair of manipulation check questions which assessed our assumption that the severity of norm violations and robot responses would be perceived as monotonically deceasing according to literature. Participants then assessed the violation-response interactions with respect to appropriateness and effectiveness of responses—competence criteria for face-threat

mitigation in request refusals. Participant also assessed the proportionality and naturalness of the robot's responses. All questions are included below:

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 robots response? (1 = extremely polite, 7 = extremely harsh)

Experimental Questions:

- *(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)
- *(effectiveness)* Overall, was the robot's response likely to be effective in addressing the potentially inappropriate nature of the request? (1 = extremely **un**likely to be effective, 7 = extremely likely to be effective)
- *(naturalness)* Overall, how natural was the robots response? (1 = extremely unnatural, 7 = extremely natural)

5 RESULTS

5.1 Analysis

We conducted Bayesian Repeated-Measures Analyses of Variance (RM-ANOVAs)[2] using JASP [40], with Bayes Factor (BF) analysis, in which Inclusion Bayes Factors (BFs) were calculated to determine the relative strength of evidence for models including each candidate main effect or interaction effect, in terms of ability to explain the gathered data. Results were then interpreted following the recommendations by Lee and Wagenmakers [46], with BF $\in [0.333, 3.0]$ considered inconclusive, and BFs above or below this range taken as evidence in favor or against an effect. In such cases, Bayes Factors were interpreted using the labels proposed by [41]. When effects could not be ruled out, post hoc Bayesian t-tests were used to examine pairwise comparisons between conditions.

Since Bayesian statistics are still not widely used within the HRI community, we will briefly explain its advantages over the traditional Frequentist approach. Bayesian statistics do not rely on p-values, which have been questioned by recent literature [63, 67, 73]. Instead of using binary significance tests, Bayesian statistics allow researchers to quantify the strength of evidence both for and against competing hypotheses [39]. In this way, researchers can incrementally check whether their data is sufficient to confirm or refute your hypotheses, without the need for power analyses. This approach makes it easier to continue research on the same topic [51, 72]. The complete results of all statistical tests, including all Bayes factors found in post-hoc analyses, is included as a supplemental document and is also available on OSF at tinyurl.com/robotResponse24.

[2]This analysis does not account for the ordinal nature of Likert data; this is a known shortcoming of JASP [70].

5.2 Manipulation Checks

5.2.1 Wrongness of Violation. An RM-ANOVA revealed extreme evidence for an effect of norm violation type on participants' assessment of its moral wrongness ($BF_{incl} = 4.094x10^{12}$). Post-hoc analysis of the effect of violation type (shown in Figure 2) revealed that participants perceived violation ***A**-paycode tampering* ($\mu_A = 5.86, \sigma_A = 1.7$) to be the most wrong and violation ***D**-playful prank* to be the least severe ($\mu_D = 3.78, \sigma_D = 1.57$); however, they perceived ***B**-task cheating* ($\mu_B = 5.18, \sigma_B = 1.47$) and ***C**-bullying* ($\mu_C = 5.1, \sigma_C = 1.5$) to be equal in severity ($BF = .141$). All other pairwise BFs were greater than 350.

These results mostly support our assumption described in Section 4.2.1 that participants would perceive the severity of norm violations in a monotonically decreasing order consistent with previous sociolinguistics research [9]. On average, the violations with material consequences for explicitly prohibited actions were perceived as more wrong than those with emotional consequences relating to social etiquette. Within each type, the violation designed to be more serious was perceived as more wrong. However, instead of finding a visible decrease across all four violations, our results show that participants perceived ***B**-task cheating* and ***C**-bullying* equivalently. Critically, participants still differentiated between these violations in other ways and felt that they merited different responses. For example, participants found it more effective for the robot to use response ***1**-Bald on Record* to respond to ***B**-task cheating* than ***C**-bullying* ($BF = 9.991$).

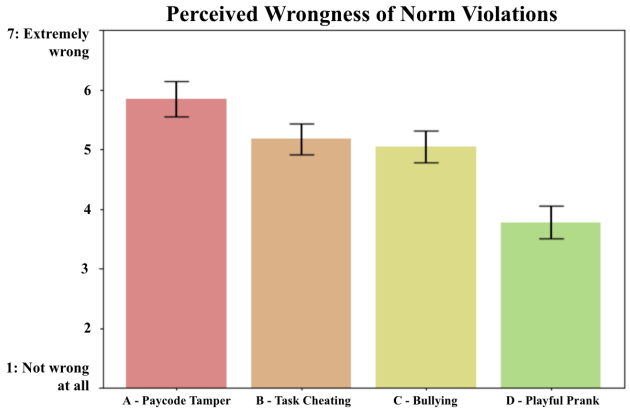

Figure 2: Perceived wrongness of norm violations.

5.2.2 Politeness of Response. An RM-ANOVA revealed extreme evidence for an effect of robot's response strategy on participants' assessment of the robot's politeness or harshness ($BF_{incl} = 1.67x10^{14}$), shown in Figure 3. Participants perceived response ***1**-Bald on Record* ($\mu_1 = 4.95, \sigma_1 = 1.49$) to be the most harsh and response ***3**-Negative Politeness* ($\mu_3 = 2.19, \sigma_3 = 1.12$) to be the most polite. Between these two extremes, participants perceived response ***2**-Positive Politeness* ($\mu_2 = 3.27, \sigma_2 = 1.39$) and ***4**-Off-Record* ($\mu_4 = 2.93, \sigma_4 = 1.5$), to be much more similar in politeness or harshness, with inconclusive evidence as to whether a difference in politeness was perceived between those two responses ($BF_{incl} = 1.146$). All other pairwise BFs were greater than 1800.

These results mostly support our assumption described in Section 4.2.2 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*. In human interaction, Off-Record language is the least severe because it as close as possible to a non-response, avoiding clear criticism through vague and meaningless language [9]. However, participants perceived robot use of this strategy to have the same level of politeness as response ***2**-Positive Politeness*, which is familiar and passive-aggressive (Figure 3). It is possible that robot morphology may have limited the ability to deliver a convincing Off-Record response. Even on the highly expressive Furhat platform used in this research, the difficulty of capturing a lighthearted, nonchalant feeling in a robot's tone of voice, timing, and facial expression, may have caused response ***4**-Off-Record* to come off as more passive-aggressive than intended. This finding is consistent with previous observations that polite, deferential robot gestures can be perceived as sassy and condescending [56].

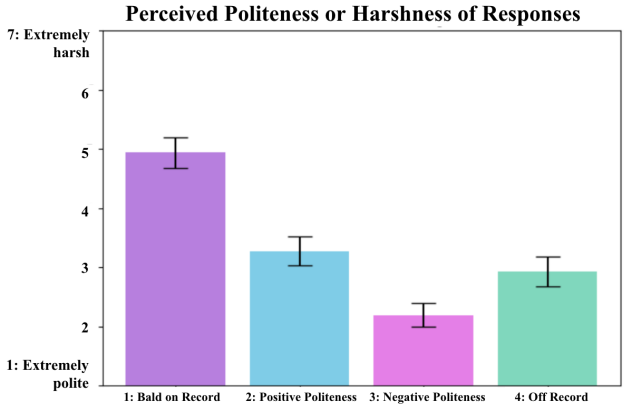

Figure 3: Perceived politeness or harshness of responses.

5.3 H1: Proportionality

An RM-ANOVA revealed extreme evidence for effects of both violation ($BF_{incl} = 1.16x10^9$) and response type ($BF_{incl} = 1.1x10^6$) on perceived proportionality, but strong evidence against a violation-response interaction ($BF_{incl} = .09$). Post-hoc analysis of the effect of response type on perceived proportionality showed that response ***1**-Bald on Record* ($\mu_1 = 4.02, \sigma_1 = 1.37$) was rated the closest to a perfectly proportional score of 4. All other responses to any violation were perceived as more polite than the request merited. Response ***1**-Bald on Record* was perceived as more proportional than any other response, including ***2**-Positive Politeness* ($\mu_2 = 3.3, \sigma_2 = 1.3$), ***3**-Negative Politeness* ($\mu_3 = 2.77, \sigma_3 = 1.15$), and ***4**-Off-Record* ($\mu_4 = 2.89, \sigma_4 = 1.27$), with all pairwise BFs > 2000. Analysis also showed moderate evidence against responses ***3**-Negative Politeness* and ***4**-Off-Record* differing in their level of proportionality ($BF = .14$). Post-hoc analysis of the effect of violation type on perceived proportionality showed that any response to violation ***A**-paycode tampering* was perceived as more polite than the request merited ($\mu_A = 2.7, \sigma_A = 1.28$) and that any response to violation ***D**-playful prank* ($\mu_D = 3.93, \sigma_D = 1.29$) was the closest to proportional. Analysis showed strong evidence against a difference in the proportionality of any response to ***B**-task cheating* ($\mu_B = 3.2, \sigma_B = 1.23$) or ***C**-bullying* ($\mu_C = 3.17, \sigma_C = 1.4$) ($BF = .1$), with all other pairwise BFs greater than 240.

The evidence against an interaction effect means our results do not support **H1**, which hypothesized that face-theoretic proportionality would correspond to the most proportional overall response behavior. However, it is unlikely that people in general are indifferent to proportionality in robot interactions, which has been strongly supported in other work [34, 37, 54]. Instead, our set of norm violations may only represent a limited subset of the overall spectrum of possible violation severity. Though our norm violations differ in their potential consequences, they are all simply questions or requests. Many other norm-violating actions may be far more benign (sneezing loudly) or severe (slapping someone, hate speech) than any question or request. In these cases, a robot's over- or under-harshness may be more salient.

5.4 H2: Effectiveness

An RM-ANOVA revealed extreme evidence for an effect of response type on perceived effectiveness ($BF_{incl} = 2.734x10^7$). Post-hoc analysis of this effect showed that participants perceived both direct response strategies—***1**-Bald on Record* ($\mu_1 = 5.32, \sigma_1 = 1.5$) and ***3**-Negative Politeness* ($\mu_3 = 5, \sigma_3 = 1.59$)—to be overall more likely to be effective in successfully addressing a norm violation than both indirect strategies—***2**-Positive Politeness* ($\mu_2 = 4.12, \sigma_2 = 1.63$) and ***4**-Off-Record* ($\mu_4 = 3.65, \sigma_4 = 1.54$), with all pairwise BFs > 1000.

An RM-ANOVA also revealed strong evidence for a violation-response interaction ($BF_{incl} = 13.465$). Post-hoc analysis of violation-response interaction (Figure 4) showed that both direct response strategies—***1**-Bald on Record* ($\mu_{A1} = 5.78, \sigma_{A1} = 1.18$) and ***3**-Negative Politeness* ($\mu_{A3} = 5.32, \sigma_{A3} = 1.49$) were perceived as more likely to be effective than both indirect strategies—***2**-Positive Politeness* ($\mu_{A2} = 4.13, \sigma_{A2} = 1.78$) and ***4**-Off-Record* ($\mu_{A4} = 3.23, \sigma_{A4} = 1.54$) in responding to violation ***A**-paycode tampering*, with all pairwise BFs > 7. The same was true for violation ***B**-task cheating* ($\mu_{B1} = 5.84, \sigma_{B1} = 1.16, \mu_{B2} = 4.065, \sigma_{B2} = 1.55, \mu_{B3} = 5.03, \sigma_{B3} = 1.52, \mu_{B4} = 3.78, \sigma_{B4} = 1.63$), with all pairwise BFs > 3.2. For violation ***C**-bullying*, post-hoc analysis showed moderate evidence that response ***1**-Bald on Record* ($\mu_{C1} = 4.74, \sigma_{C1} = 1.67$) was more effective than response ***2**-Positive Politeness* ($\mu_{C2} = 4.74, \sigma_{C2} = 1.67$) ($BF = 3.18$), and provided moderate evidence against differences in perceived effectiveness between responses ***2**-Positive Politeness* and ***4**-Off-Record* ($\mu_{C4} = 3.87, \sigma_{C4} = 1.43$) ($BF = .29$), and between responses ***1**-Bald on Record* and ***3**-Negative Politeness* ($\mu_{C3} = 4.71, \sigma_{C3} = 1.7$) ($BF = .26$). For violation ***D**-playful prank*, post-hoc analysis showed moderate evidence that response ***1**-Bald on Record* ($\mu_{D1} = 4.94, \sigma_{D1} = 1.61$) and ***3**-Negative Politeness* ($\mu_{D3} = 4.94, \sigma_{D3} = 1.66$) were both more effective than response ***4**-Off-Record* ($\mu_{D4} = 3.74, \sigma_{D4} = 1.55$) ($BF = 9.36$, $BF = 8.56$ respectively). It also provided moderate evidence against differences in perceived effectiveness between responses ***1**-Bald on Record* and ***3**-Negative Politeness* ($BF = .26$). In this way, our results do not support **H2**, which hypothesized that face-theoretic proportionality,

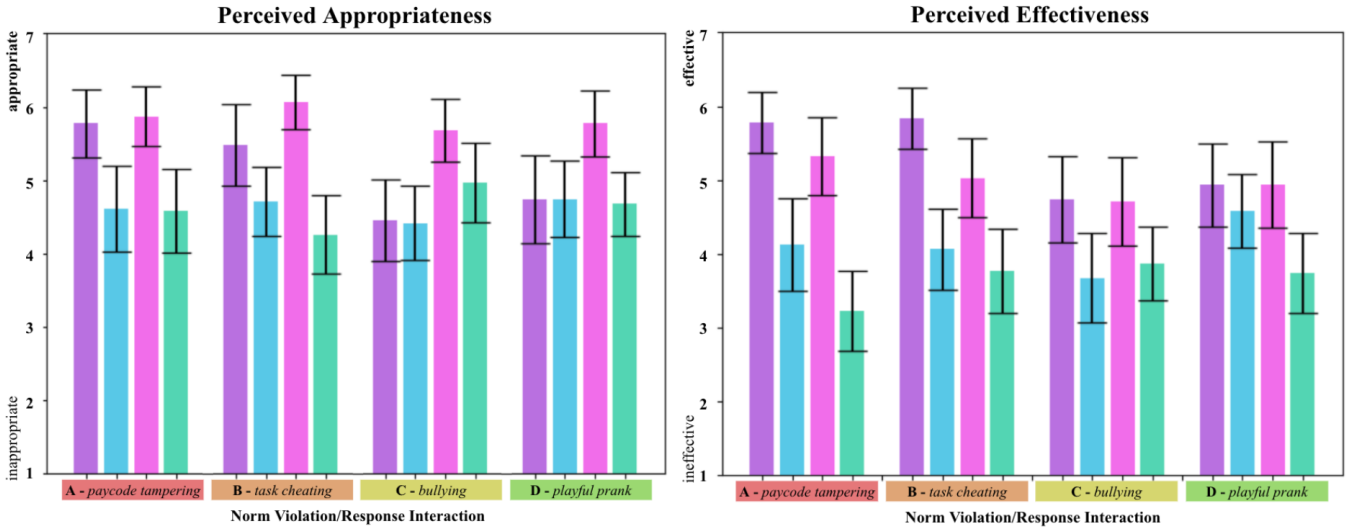

Figure 4: Perceived appropriateness (reverse coded) and perceived effectiveness for each violation-response interaction.

as it is defined in the sociolinguistics literature, would correspond to the most effective overall robot response behavior. However, these results do suggest that robots ought to use some form of proportionality to select effective responses, which we call **bounded proportionality** and discuss in Section 6.1

5.5 H3: Appropriateness

An RM-ANOVA revealed extreme evidence for an effect of response type on perceived appropriateness ($BF_{incl} = 262,893$). Post-hoc analysis of this effect showed that participants perceived response ***3**-Negative Politeness* ($\mu_3 = 5.85, \sigma_3 = 1.18$) to be more appropriate than all other responses, including response ***1**-Bald on Record* ($\mu_1 = 5.11, \sigma_1 = 1.62, BF = 443.75$), response ***2**-Positive Politeness* ($\mu_2 = 4.62, \sigma_2 = 1.48, BF = 1.1x10^{10}$), and response ***4**-Off-Record* ($\mu_4 = 4.62, \sigma_4 = 1.49, BF = 1.78x10^{10}$), with. Additionally, analysis showed strong evidence against responses ***2**-Positive Politeness* and ***4**-Off-Record* having different perceived appropriateness ($BF = .1$).

We also found very strong evidence for a violation-response interaction ($BF_{incl} = 34.466$) (Figure 4). Post-hoc analysis of this interaction effect showed that for violation ***A**-paycode tampering*, direct responses ***1**-Bald on Record* ($\mu_{A1} = 5.77, \sigma_{A1} = 1.31$) and ***3**-Negative Politeness* ($\mu_{A3} = 5.87, \sigma_{A3} = 1.15$) were more appropriate than indirect responses ***2**-Positive Politeness* ($\mu_{A2} = 4.61, \sigma_{A2} = 1.67$) and ***4**-Off-Record* ($\mu_{A4} = 4.58, \sigma_{A4} = 1.61$), with all pairwise BFs > 11. Additionally, there was evidence against direct responses ***1**-Bald on Record* and ***3**-Negative Politeness* having different perceived appropriateness ($BF = .27$) and against indirect responses ***2**-Positive Politeness* and ***4**-Off-Record* having different perceived appropriateness ($BF = .26$) in responding to violation ***A**-paycode tampering*. For violation ***B**-task cheating*, evidence showed that response ***3**-Negative Politeness* ($\mu_{B3} = 6.07, \sigma_{B3} = 1.06$) was more appropriate than either indirect response ***2**-Positive Politeness* ($\mu_{B2} = 4.71, \sigma_{B2} = 1.35, BF = 442.63$) or ***4**-Off-Record* ($\mu_{B4} = 4.26, \sigma_{B4} = 1.53, BF = 11631.4$). It also showed that response ***1**-Bald on Record* ($\mu_{B1} = 5.48, \sigma_{B1} = 1.57$) was more appropriate than response ***4**-Off-Record* ($BF = 13.16$). For violation ***C**-bullying*, evidence showed that response ***3**-Negative Politeness* ($\mu_{C3} = 5.68, \sigma_{C3} = 1.22$) was more appropriate than either response ***1**-Bald on Record* ($\mu_{C1} = 4.45, \sigma_{C1} = 1.59, BF = 26.87$) or response ***2**-Positive Politeness* ($\mu_{C2} = 4.42, \sigma_{C2} = 1.46, BF = 56.37$). It also showed evidence against response ***1**-Bald on Record* and ***2**-Positive Politeness* having different appropriateness ($BF = .26$). For violation ***D**-playful prank*, evidence showed that response ***3**-Negative Politeness* ($\mu_{D3} = 5.77, \sigma_{D3} = 1.28$) was more appropriate than all other responses, including response ***1**-Bald on Record* ($\mu_{D1} = 4.74, \sigma_{D1} = 1.71, BF = 4.95$), response ***2**-Positive Politeness* ($\mu_{D2} = 4.74, \sigma_{D2} = 1.48, BF = 8.48$) and response ***4**-Off-Record* ($\mu_{D4} = 4.68, \sigma_{D4} = 1.22, BF = 29.8$). It also showed evidence against these thee other responses having different levels of appropriateness, with all pairwise BFs < .27. In this way, our results support **H3**, which hypothesized that indirect responses would be perceived as less appropriate than direct responses.

5.6 H4: Naturalness

An RM-ANOVA found anecdotal evidence for and against effects of violation ($BF_{incl} = 1.25$) and response ($BF_{incl} = .913$) on perceived naturalness of responses. This indicates that more data would be needed to support or refute **H4**, which hypothesized that indirect responses would be perceived as less natural than direct ones. Post-hoc analysis of the effect violation-response interaction did show that response ***3**-Negative Politeness* was uniformly most natural, but only measurably more natural in certain cases, typically when compared to uses of response ***4**-Off-Record* to violations ***A**-paycode tampering*, ***B**-task cheating*, and ***D**-playful prank*.

6 DISCUSSION

The goal of our experiment was to investigate the effects of a robot's use of human-like Face-theoretic linguistic politeness cues in noncompliance interactions. Specifically, we investigated the multiple and potentially conflicting attributes of successful robot responses to norm violating requests of varying severity. These attributes

included proportionality (calibrated harshness), competence (effectiveness and appropriateness) [31], and response naturalness. Overall, we found that linguistic politeness strategies that use direct, formal language are perceived as more effective and more appropriate than strategies that use indirect, informal language.

These findings indicate that human-like linguistic politeness strategies do not precisely apply to robot interactions and cannot serve as a direct guide for roboticists and interaction designers creating tactful noncompliance responses. While humans expect robots to have human-like social competence in addressing norm violations [54], this does not necessarily confer exact mimicry of human-like strategic politeness cues. Critically, our results do not suggest that social robots are exempt from using human-like politeness at all. Robots in noncompliance interactions must select language to soften their refusals to match the severity of a situation in order to be competent, appropriate social actors. For example, it would have been a less appropriate overall policy for the robot in our scenario to uniformly use the harshest response ***1****-Bald on Record*. In this way, face-based politeness cues are still a relevant framework for interaction designers. However, robots may be more successful and acceptable if they use softening or hedging strategies that avoid indirect, passive, emotional, or familiar language. This is consistent with HRI research showing that humans may expect robots to use functional, rule-based politeness cues [50].

There are several possible reasons why participants may have found indirect robot response behaviors to be inappropriate. Participants may have felt that the robot lacked the social or emotional status allude to familiarity or closeness within its relationship to its human teammates [76]. Participants may have felt that robots have less social power than humans [49], and may not have seen robots in roles that afforded them the status to give rebukes [54]. Dissonance between the robot's status and actions may have created a sense of disingenuousness when the robot mimicked human politeness grounded in a sense of intimacy or belonging [14, 15].

6.1 Robots can use *bounded* proportionality to address norm violations

Our results suggest that the best overall behavioral "policy" for the robot to adapt is to select between the two direct linguistic strategies, using strategy ***1****-Bald on Record* for moral violations with more material consequences, and strategy ***3****-Negative Politeness* for social violations with emotional consequences. Because this response-selection behavior does not exactly correspond to human face-theoretic proportionality, we term it "bounded proportionality". Under "bounded proportionality," robots still use harsher or softer responses according to violation severity, but are limited to linguistic modifiers which are direct, formal, and straightforward.

6.2 Are direct robots more transparent?

Our results suggest that people may prefer robots to avoid language that does not align with their ontological [12, 43] or social [36, 76] status. However, there may be another reason for robots to avoid cues that allude to human characteristics, experiences, or communities—because it is *transparent* to do so. Transparency is the principle that robots should communicate their inner workings and limitations [4]. HRI researchers [2, 74] and policymakers [16] have explored how transparent design helps robot users build accurate mental models [7, 45, 85], and calibrate their trust [2, 60]. Robot norm violation response behaviors could either affirm or challenge the mental models human use to assess robots' capabilities and trustworthiness. Direct, formal language may implicitly reinforce the idea that robots are inanimate—incapable of understanding human experiences. Reciprocally, indirect, familiar language (such as teasing, endearment, and in-group references) may implicitly reinforce inaccurate ideas about robots' social and emotional affordances. Roboticists have the opportunity, and perhaps the obligation, to consider how their design choices impact humans' understanding of robots as social, moral, and emotional others [76].

6.3 Limitations & Future Work

While our experimental scenario captured many norm violations, it was still a fictional scenario presented to participants without the full context of an actual collaborative task or actual potential for harm. Norms and norm violations are always context dependent and cannot be completely assessed without contextual understanding [9]. This may limit the fidelity of our brief experimental interaction. Knowing this, included a qualitative question at the end of our experiment which asked participants to reflect on additional contextual factors that would be important if they were evaluating similar interactions in a real collaborative environment. While analysis of these results is beyond the scope of this work, we will explore this data as part of future work on this topic.

Future work on this topic can also consider a broader range of linguistic cues and situational factors. For example, future work ought to consider gender more rigorously in this interaction design context. Gender norms, gendered expectations of polite behavior, and sexism all influence noncompliance interactions in HRI [54, 59, 68, 82], and critically, challenge the very notion of working towards "optimally proportional" norm violation responses [38]. Furthermore, understanding how gender and power shape technology is a responsibility of the HRI community [58, 62, 83]. Future work can explore how our results might interact with gendered robot design cues, or gendered expectations of politeness, similar to the work performed by Jackson et al. [38].

7 CONCLUSION

We have presented the results of a human-subjects study in which participants evaluated norm violation-response interactions between a human and robot. Our goal was to explore and evaluate potential tradeoffs in the design of robot response behaviors informed by human face-based politeness cues. Our results show that politeness strategies grounded in direct language were perceived as more likely to be effective and appropriate than indirect strategies. This suggests that, while people expect social robots to act with norm-sensitive social competence, they do not expect them to strictly mimic human linguistic behaviors.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Anna M. H. Abrams, Pia S. C. Dautzenberg, Carla Jakobowsky, Stefan Ladwig, and Astrid M. Rosenthal-von der Pütten. 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, 272–280.

[2] Victoria Alonso and Paloma de la Puente. 2018. System Transparency in Shared Autonomy: A Mini Review. *Frontiers in Neurorobotics* (2018).

[3] Sean Andrist, Micheline Ziadee, Halim Boukaram, Bilge Mutlu, and Majd Sakr. 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, 157–164.

[4] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable Agents and Robots: Results from a Systematic Literature Review. In *Proc. Autonomous Agents and Multi-Agent Systems (AAMAS)*.

[5] Santosh Balajee Banisetty and Tom Williams. 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*. 499–504.

[6] Ryan Blake Jackson, Sihui Li, Santosh Balajee Banisetty, Sriram Siva, Hao Zhang, Neil Dantam, and Tom Williams. 2021. An Integrated Approach to Context-Sensitive Moral Cognition in Robot Cognitive Architectures. In *Proceedings of Intelligent Robots and Systems (IROS)*.

[7] Serena Booth, Sanjana Sharma, Sarah Chung, Julie Shah, and Elena L. Glassman. 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)*.

[8] Gordon Briggs, Tom Williams, Ryan Blake Jackson, and Matthias Scheutz. 2022. Why and how robots should say 'no'. *Int'l Jour. Social Robotics* (2022).

[9] Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press.

[10] De'Aira Bryant, Jason Borenstein, and Ayanna Howard. 2020. Why Should We Gender?: The Effect of Robot Gendering and Occupational Stereotypes on Human Trust and Perceived Competency. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Cambridge United Kingdom, 13–21.

[11] Robert B Cialdini and Melanie R Trost. 1998. Social influence: Social norms, conformity and compliance. In *The handbook of social psychology*. McGraw-Hill.

[12] Herbert Clark and Kerstin Fischer. 2022. Social robots as depictions of social agents - Behavioral and Brain Sciences (forthcoming). *Behavioral and Brain Sciences* 2022 (07 2022), 1–33.

[13] Leigh Clark. 2018. Social Boundaries of Appropriate Speech in HCI: A Politeness Perspective. In *Proceedings of British HCI*.

[14] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, 1–12.

[15] Leigh Clark, Abdulmalik Yusuf Ofemile, and Benjamin Cowan. 2020. *Exploring Verbal Uncanny Valley Effects with Vague Language in Computer Speech*. 317–330.

[16] European Commission, Content Directorate-General for Communications Networks, and Technology. 2019. *Ethics guidelines for trustworthy AI*. Publications Office.

[17] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Daniel Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Annual Meeting of the Association for Computational Linguistics*.

[18] Autumn Edwards, Chad Edwards, and Andrew Gambino. 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 (08 2020).

[19] Vanessa Evers, Heidy Maldonado, Talia Brodecki, and Pamela Hinds. 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)*.

[20] Imran Fanaswala, Brett Browning, and Majd Sakr. 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, 133–134.

[21] Jodi Forlizzi. 2007. How Robotic Products Become Social Products: An Ethnographic Study of Cleaning in the Home. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (Arlington, Virginia, USA) *(HRI '07)*. Association for Computing Machinery, 129–136.

[22] Hideki Garcia, Katie Winkle, Tom Williams, and Megan Strait. 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)*.

[23] Albert Gatt and Emiel Krahmer. 2018. Survey of the State of the Art in Natural Language Generation: Core Tasks, Applications and Evaluation. *J. Artif. Int. Res.* 61, 1 (jan 2018), 65–170.

[24] Felix Gervits, Gordon Briggs, and Matthias Scheutz. 2017. The Pragmatic Parliament: A Framework for Socially-Appropriate Utterance Selection in Artificial Agents. *Cognitive Science* (2017).

[25] Erving Goffman. 1967. *Interaction Ritual: Essays in Face-to-Face Behavior*.

[26] Swati Gupta, Marilyn A. Walker, and Daniela M. Romano. 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*.

[27] Stephan Hammer, Birgit Lugrin, Sergey Bogomolov, Kathrin Janowski, and Elisabeth André. 2016. Investigating Politeness Strategies and Their Persuasiveness for a Robotic Elderly Assistant. In *Proceedings of the 11th International Conference on Persuasive Technology - Volume 9638* (Salzburg, Austria) *(PERSUASIVE 2016)*. Springer-Verlag, 315–326.

[28] Erin R. Hoffman, David W. McDonald, and Mark Zachry. 2017. Evaluating a Computational Approach to Labeling Politeness: Challenges for the Application of Machine Classification to Social Computing Data. *Proc. ACM Hum.-Comput. Interact.* (2017).

[29] Thomas Holtgraves. 2021. Understanding Miscommunication: Speech Act Recognition in Digital Contexts. *Cognitive science* 45 (10 2021).

[30] Yaxin Hu, Yuxiao Qu, Adam Maus, and Bilge Mutlu. 2022. Polite or Direct? Conversation Design of a Smart Display for Older Adults Based on Politeness Theory. In *Proc. CHI*.

[31] Danette Ifert Johnson. 2007. Politeness Theory and Conversational Refusals: Associations between Various Types of Face Threat and Perceived Competence. *Western Journal of Communication* 71 (07 2007), 196–215.

[32] Danette Ifert Johnson, Michael Roloff, and Melissa Riffee. 2004. Responses to refusals of requests: Face threat and persistence, persuasion and forgiving statements. *Communication Quarterly* 52 (09 2004), 347–356.

[33] Nasif Imtiaz, Justin Middleton, Peter Girouard, and Emerson Murphy-Hill. 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.*

[34] Ryan Blake Jackson, Ruchen Wen, and Tom Williams. 2019. Tact in Noncompliance: The Need for Pragmatically Apt Responses to Unethical Commands. In *Proc. AI, Ethics, and Society (AIES)*.

[35] Ryan Blake Jackson and Tom Williams. 2019. Language-Capable Robots may Inadvertently Weaken Human Moral Norms. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

[36] Ryan Blake Jackson and Tom Williams. 2021. A Theory of Social Agency for Human-Robot Interaction. *Frontiers in Robotics and AI* (2021).

[37] Ryan Blake Jackson and Tom Williams. 2022. Enabling Morally Sensitive Robotic Clarification Requests. *ACM Trans. Human-Robot Interaction* (2022).

[38] Ryan Blake Jackson, Tom Williams, and Nicole Smith. 2020. Exploring the Role of Gender in Perceptions of Robotic Noncompliance. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

[39] Andrew Jarosz and Jennifer Wiley. 2014. What Are the Odds? A Practical Guide to Computing and Reporting Bayes Factors. *The Journal of Problem Solving* 7 (11 2014).

[40] JASP Team. 2023. JASP (Version 0.18.0)[Computer software]. https://jasp-stats.org/

[41] Harold Jeffreys. 1948. *Theory of probability*. (2d ed. ed.). Clarendon Press, Oxford.

[42] Malte F. Jung, Nikolas Martelaro, and Pamela J. Hinds. 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)*.

[43] Peter H. Kahn and Solace Shen. 2017. *NOC NOC, Who's There? A New Ontological Category (NOC) for Social Robots*. Cambridge University Press, 106–122.

[44] Boyoung Kim, Ruchen Wen, Qin Zhu, Tom Williams, and Elizabeth Phillips. 2021. Robots as Moral Advisors: The Effects of Deontological, Virtue, and Confucian Role Ethics on Encouraging Honest Behavior. In *Comp. HRI*.

[45] Minae Kwon, Malte F. Jung, and Ross A. Knepper. 2016. Human expectations of social robots. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

[46] Michael D Lee and Eric-Jan Wagenmakers. 2014. *Bayesian cognitive modeling: A practical course*. Cambridge university press.

[47] Namyeon Lee, Jeonghun Kim, Eunji Kim, and Ohbyung Kwon. 2017. The Influence of Politeness Behavior on User Compliance with Social Robots in a Healthcare Service Setting. *International Journal of Social Robotics* 9 (11 2017).

[48] Geoffrey Leech. 2014. The Pragmatics of Politeness. *The Pragmatics of Politeness* (07 2014), 1–368.

[49] Eleonore Lumer and Hendrik Buschmeier. 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)*.

[50] Eleonore Lumer and Hendrik Buschmeier. 2023. Should robots be polite? Expectations about politeness in human–robot interaction. *Frontiers in Robotics and AI* (2023).

[51] Alexander Ly, Alexander Etz, Maarten Marsman, and Eric-Jan Wagenmakers. 2018. Replication Bayes factors from evidence updating. *Behavior Research Methods* 51 (08 2018).

[52] Christoforos Mavrogiannis, Alena M. Hutchinson, John Macdonald, Patricia Alves-Oliveira, and Ross A. Knepper. 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), 421–430.

[53] Sara Mills. 2005. Gender and impoliteness. *Jour. Politeness Research-Language Behaviour Culture* (2005).

[54] Terran Mott and Tom Williams. 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)*.

[55] Bilge Mutlu and Jodi Forlizzi. 2008. Robots in organizations: The role of workflow, social, and environmental factors in human-robot interaction. *ACM/IEEE Int'l Conf. on Human-Robot Interaction* (2008).

[56] Aidan Naughton and Tom Williams. 2021. How to Tune Your Draggin': Can Body Language Mitigate Face Threat in Robotic Noncompliance?. In *Proc. International Conference on Social Robotics (ICSR)*.

[57] Caroline Pantofaru, Leila Takayama, Tully Foote, and Bianca Soto. 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, 327–334.

[58] Giulia Perugia, Stefano Guidi, Margherita Bicchi, and Oronzo Parlangeli. 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)*.

[59] Giulia Perugia and Dominika Lisy. 2022. Robot's Gendering Trouble: A Scoping Review of Gendering Humanoid Robots and its Effects on HRI. *International Journal of Social Robotics* (2022).

[60] Avi Rosenfeld and Ariella Richardson. 2019. Explainability in Human–Agent Systems. *Autonomous Agents and Multi-Agent Systems* (2019).

[61] Maha Salem, Micheline Ziadee, and Majd Sakr. 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)*.

[62] Katie Seaborn and Peter Pennefather. 2022. Gender Neutrality in Robots: An Open Living Review Framework. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

[63] Joseph Simmons, Leif Nelson, and Uri Simonsohn. 2011. False-Positive Psychology. *Psychological science* 22 (11 2011), 1359–66.

[64] Cailyn Smith, Charlotte Gorgemans, Ruchen Wen, Saad Elbeleidy, Sayanti Roy, and Tom Williams. 2022. Leveraging Intentional Factors and Task Context to Predict Linguistic Norm Adherence. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*.

[65] Marie Louise Juul Søndergaard and Lone Koefoed Hansen. 2018. Intimate Futures: Staying with the Trouble of Digital Personal Assistants through Design Fiction. In *Proc. Designing Interactive Systems (DIS.*

[66] Vasant Srinivasan and Leila Takayama. 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* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, 4945–4955.

[67] Jonathan Sterne and George Davey Smith. 2001. Sifting the evidence - what's wrong with significance tests? BMJ. *BMJ (Clinical research ed.)* 322 (02 2001), 226–31.

[68] Laetitia Tanqueray, Tobiaz Paulsson, Mengyu Zhong, Stefan Larsson, and Ginevra Castellano. 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)*.

[69] Marina Terkourafi. 2005. Beyond the Micro-level in Politeness Research. *Journal of Politeness Research-language Behaviour Culture* 1 (07 2005), 237–262.

[70] Don van den Bergh, Johnny Van Doorn, Maarten Marsman, Tim Draws, Erik-Jan Van Kesteren, Koen Derks, Fabian Dablander, Quentin F Gronau, Šimon Kucharskỳ, Akash R Komarlu Narendra Gupta, et al. 2020. A tutorial on conducting and interpreting a Bayesian ANOVA in JASP. *L'Année psychologique* 120, 1 (2020), 73–96.

[71] Diede P.M. Van der Hoorn, Anouk Neerincx, and Maartje M.A. de Graaf. 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, 140–148.

[72] A J Verhagen and Eric-Jan Wagenmakers. 2014. Bayesian Tests to Quantify the Result of a Replication Attempt. *Journal of experimental psychology. General* 143 (05 2014).

[73] Eric-Jan Wagenmakers. 2007. A Practical Solution to the Pervasive Problems of p Values. *Psychonomic bulletin & review* 14 (11 2007), 779–804.

[74] Sebastian Wallkötter, Silvia Tulli, Ginevra Castellano, Ana Paiva, and Mohamed Chetouani. 2021. Explainable Embodied Agents Through Social Cues: A Review. *J. Hum.-Robot Interact.* (2021).

[75] Richard J. Watts. 2003. *Politeness*. Cambridge University Press.

[76] Kara Weisman. 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)*.

[77] Ruchen Wen, Zhao Han, and Tom Williams. 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)*.

[78] Ruchen Wen, Mohammed Aun Siddiqui, and T. Williams. 2020. Dempster-Shafer Theoretic Learning of Indirect Speech Act Comprehension Norms. In *AAAI*.

[79] Tom Williams, Priscilla Briggs, and Matthias Scheutz. 2015. Covert Robot-Robot Communication: Human Perceptions and Implications for Human-Robot Interaction. *J. Hum.-Robot Interact.* (sep 2015), 24–49.

[80] Tom Williams, Ryan Jackson, and Jane Lockshin. 2018. A Bayesian Analysis of Moral Norm Malleability during Clarification Dialogues. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*.

[81] Steven Wilson and Adrianne Kunkel. 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 (06 2000), 195–221.

[82] Katie Winkle, Ryan Blake Jackson, Gaspar Isaac Melsión, Dražen Brščić, Iolanda Leite, and Tom Williams. 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)*.

[83] Katie Winkle, Donald McMillan, Maria Arnelid, Katherine Harrison, Madeline Balaam, Ericka Johnson, and Iolanda Leite. 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* (Stockholm, Sweden) *(HRI '23)*. Association for Computing Machinery, 72–82.

[84] Katie Winkle, Gaspar Isaac Melsión, Donald McMillan, and Iolanda Leite. 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)*.

[85] Robert H Wortham, Andreas Theodorou, and Joanna J Bryson. 2016. What Does the Robot Think? Transparency as a Fundamental Design Requirement for Intelligent Systems. In *Proc. IJCAI Workshop on Ethics for Artificial Intelligence*.

[86] Qin Zhu, Tom Williams, and Ryan Jackson. 2018. Blame-Laden Moral Rebukes and the Morally Competent Robot: A Confucian Ethical Perspective. *Journal of Science and Engineering Ethics*.

[87] Qin Zhu, Tom Williams, and Ruchen Wen. 2021. Role-based Morality, Ethical Pluralism, and Morally Capable Robots. *Journal of Contemporary Eastern Asia* (2021).

[88] Yifei Zhu, Ruchen Wen, and Tom Williams. 2024. Robots for Social Justice (R4SJ): Toward a More Equitable Practice of Human-Robot Interaction. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.