Exploring the Role of Gender in Perceptions of Robotic Noncompliance

Ryan Blake Jackson, Tom Williams, and Nicole Smith
[rbjackso,twilliams,nmsmith]@mines.edu
Colorado School of Mines
Golden, Colorado, USA

ABSTRACT

A key capability of morally competent robots is to reject or question potentially immoral human commands. However, robot rejections of inappropriate commands must be phrased with great care and tact. Previous research has shown that failure to calibrate the "face threat" in a robot's command rejection to the severity of the norm violation in the command can lead humans to perceive the robot as inappropriately harsh and can needlessly decrease robot likeability. However, it is well-established that gender plays a significant role in determining linguistic politeness norms and that people have a powerful natural tendency to gender robots. Yet, the effect of robotic gender presentation on these noncompliance interactions is not well understood. We present an experiment that explores the effects of robot and human gender on perceptions of robots in noncompliance interactions, and find evidence of a complicated interplay between these gendered factors. Our results suggest that (1) it may be more favorable for a male robot to reject commands than for a female robot to do so, (2) it may be more favorable to reject commands given by a male human than by a female human, and (3) that robots may be perceived more favorably when their gender matches that of human interactants and observers.

KEYWORDS

Robot Noncompliance, Gender, Politeness Theory

ACM Reference Format:
Ryan Blake Jackson, Tom Williams, and Nicole Smith. 2020. Exploring the Role of Gender in Perceptions of Robotic Noncompliance. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20), March 23–26, 2020, Cambridge, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3319502.3374831

1 INTRODUCTION

Human-Robot Interaction (HRI) researchers are increasingly turning to natural language to allow robots to communicate fluidly and easily with most humans [33, 34]. Much of this communication is task-oriented, and the human role is largely to command and task robots [53]. Even so, robots should not blindly follow every human directive that they receive. Indeed, there are many sensible reasons for a robot to reject a command, ranging from physical inability to moral objection [6]. Rejecting commands based on moral impermissibility is especially important as robots' abilities increase because the number of permissible commands that the robot is incapable of following will decrease, and the number of impermissible commands the robot is capable of following will grow.

The ability to tactfully reject inappropriate commands is critical due to the potential influence robots may wield within their moral ecosystems. Human morality is dynamic and malleable [19], and human moral norms are shaped not only by human community members, but also by the technologies with which they interact [20, 50]. Given social robots' persuasive capacity over humans [5, 30], potential to hold ingroup social status [16], and appearance as moral and social agents (cf. Jackson and Williams [26]), these robots wield uniquely impactful moral influence relative to other technologies. Previous research has even shown that robots may inadvertently weaken human application of moral norms simply by asking questions about immoral commands [25].

Robot rejections of inappropriate commands must be phrased with great care and tact. Research has shown that failure to do so can lead humans to perceive a robot as inappropriately harsh and decrease robot likeability unnecessarily [23]. Critically, robot command rejections can be perceived as either too harsh or not harsh enough, depending on the context and the phrasing chosen, so robots must dynamically adjust their adherence to politeness norms according to their context [23].

Some recent research examining phrasing in robotic command rejections has considered adjusting politeness based on the impermissibility of the human's command being rejected. However, this research did not consider gender, despite using an implicitly female robot, which we view as an oversight given the well-established and significant impact that gender has on linguistic politeness norms in human-human interaction (see Section 2.2).

We present a behavioral ethics experiment designed to investigate the role of gender stereotypes in human perceptions of robotic noncompliance. Our results suggest that (1) it may be more favorable for a male robot to reject commands than for a female robot to do so, (2) it may be more favorable to reject commands given by a male human than by a female human, and (3) that robots may be perceived more favorably when their gender matches that of human interactants and observers. The remainder of this paper begins with a survey of related work from several fields in Section 2. We then describe our experiment and analyze its results in Sections 3 and 4. Finally, we present our concluding remarks and possible avenues for future research in Section 5.

HRI '20, March 23–26, 2020, Cambridge, United Kingdom

ACM ISBN 978-1-4503-6746-2/20/03…$15.00
https://doi.org/10.1145/3319502.3374831

2 RELATED WORK

In this section, we will begin with a brief overview of the concepts of “face" and “face threat" from politeness theory, which form the basis for our understanding of how different command rejection phrasings may be more or less appropriate according to context. Next, we review the impacts of gender on politeness norms and perceived politeness in human-human interactions. Though gender and politeness can vary across cultures, we consider a western perspective for consistency with our participant pool. We then present a few studies concerning gender in artificial agents, but without specific attention to linguistic politeness and noncompliance. Finally, we discuss previous work from the HRI literature on robotic noncompliance and moral criticism.

2.1 Politeness, Face, and Face Threat

Central to our exploration of phrasing and gender in command rejection is the concept of “face threat" from politeness theory [7]. Face, consisting of positive face and negative face, is the public self-image that all members of society want to preserve and enhance for themselves. Negative face is defined as an agent’s claim to freedom of action and freedom from imposition. Positive face consists of an agent’s self-image and wants, and the desire that these be approved of by others. A discourse act that damages or threatens either of these components of face for the addressee or the speaker is a face threatening act. The degree of face threat in an interaction depends on the disparity in power between the interactants, the social distance between the interactants, and the imposition of the topic or request comprising the interaction. Various linguistic politeness strategies exist to decrease the face threat to an addressee when threatening face is unavoidable or desirable.

Commands and requests threaten the negative face of the addressee, while command rejections, especially those issued for moral reasons, threaten the positive face of the commander by expressing disapproval of the desire motivating the command. Research specifically examining command refusals found that linguistic framing of the reason for noncompliance varies along three dimensions relevant to face threat: willingness, ability, and focus on the requester [28]. It is unclear how these three dimensions pertain to robotic refusals. For example, in human-human refusals with low expressed willingness, the degree of expressed ability influences the threat to the requester’s positive face. This finding is important because, when a human refuses a request for moral reasons, there is often sufficient ability but not willingness. The same is not necessarily true for robots that may be programmed with an inability to act immorally. The dimensions of willingness and ability therefore become tangled in agents lacking true, unconstrained moral agency. We also note that this prior research focuses on threats to the face of the refuser. However, robots have no face needs, and we therefore disregard threats to robots’ face. Our work focuses on the face threat that robots present to humans by refusing requests.

Previous work found evidence that the optimal robotic command rejection should carry a face threat proportional to the severity of the normative infraction in the command being rejected [23]. In other words, commands presenting severe norm violations should be rejected more face threateningly than commands presenting less severe norm violations, and vice versa.

2.2 Gender and Politeness

Gender plays an integral role in performance and perceptions of linguistic politeness norms in human-human interactions. The concept of politeness has (implicitly) underlied a great deal of previous gender and language research, at least since the 1970s [35]. Older work has argued that women are typically more polite or more deferential than men, whereas more modern studies have challenged these notions, calling for a more context-dependent and nuanced view of gender, politeness, and their relationship [35, 36].

These works present a model of gender identity and politeness that sees both as closely inter-related performative acts that unfold over the course of every interaction. As one interactant performs their gender identity and speaks with various linguistic markers of (im)politeness, the other imposes judgments of (im)politeness informed by their beliefs regarding gender-appropriate behavior. Thus, gender is important in both performing and perceiving politeness, but not in fixed and definitive ways that might be easily programmable.

For example, professional women working in male-dominated environments may feel called upon to perform stereotypically masculine linguistic speech patterns (e.g., directness, interruption, or verbal banter) to fit in with their professional community of practice. However, others within that environment may consider such behaviors inappropriate for women in general. Stereotypical feminine gender identity is largely constructed around supportive and cooperative behavior, leading, for example, assertiveness to be categorized as impoliteness. In general, many linguistic resources that index power, including face threatening acts in general, also indirectly index masculinity, and may be seen as inappropriate for women [36]. Past feminist research often cited women as using “powerless" speech (e.g., indirectness, deference, hesitation, etc.) [31], and, though it is now clear that this stereotype was based primarily on white middle-class women and that not all women use this type of language, it nonetheless remains indexing of femininity for many communities regardless of the value or function they place on it [36]. We thus hypothesize that female-presenting robots will be viewed less favorably than male-presenting robots in noncompliance interactions. The association between masculinity and power, and other work linking masculinity to entitlement [22], leads us to further hypothesize that the robot will be viewed less favorably by male participants and less favorably when rejecting commands from a male human.

We also cannot assume that an utterance or exchange may be inherently polite or impolite in and of itself, but rather must account for listener assessments of the speaker’s intentions and motivations, and the corresponding assessments of the gender-appropriateness thereof. This helps us explain, for example, the use of extreme insults, that would appear to significantly threaten the listener’s positive face, to signal in-group solidarity, particularly in masculine groups [11, 36]. To frame this idea in terms of face threat, we must view a face threatening utterance not as inherently face threatening on its own, but rather as interpreted as face threatening given the speaker’s perceived intentions, the context, and the mediating gender norms.

Some researchers have advocated for a theoretical framework treating impoliteness on its own terms rather than in relation to politeness [14, 36]. However, for purposes of the present study, we believe that the face threat model of politeness, understood with context, gender, and intention as mediating factors, is the clearest lens through which to analyze our results. Thus, we view speech acts as lying on a continuous spectrum from impolite to polite, but emphasize that this is a spectrum of *assessment* rather than *quality*. However, this assessment is not a matter of individual judgment alone, since it is constructed within institutional and community norms that define appropriate linguistic behavior. Gender is important in this respect, since women and men[1] may be perceived to have different claims or rights to a position within the public sphere, and, therefore, different bounds on appropriate behavior [36].

2.3 Gender and Artificial Agents

Artificial social agents like robots do not have gender identities in the same way that humans do. Regardless, humans have a powerful natural tendency to ascribe gender to these artificial agents. Even machines with minimal gender cues generate gender-based stereotypic responses in humans [38].

Nass et al. [38] found that people (subconsciously) view evaluation from a male-voiced computer as more valid than evaluation from a female-voiced computer, and view socially dominant behavior from a female-voiced computer as less friendly than the same behavior from a male-voiced computer, even when voice was the only gender cue. Furthermore, there was weaker evidence that people conditionally assume that a female-voiced computer would know more about love and relationships, while a male-voiced computer would know more about computers (a stereotypically male topic at the time). Similarly, Eyssel and Hegel [15] found that visual cues as simple as hair length cause gendering of robots, with a shorter-haired "male" robot being perceived as more agentic than a longer-haired "female" robot, and the longer-haired "female" robot being perceived as more communal. Additionally, stereotypically male tasks were perceived as more suitable for the shorter-haired robot relative to the longer-haired robot, and vice versa. These findings indicate that any suggestion of gender in a given technology, however minor, may trigger stereotypic responses, and that the unintentional human tendency to gender stereotype is extremely powerful, extending even to machines.

Robot gendering can affect human perceptions of robots in other ways beyond the stereotypes described above. People appear to prefer female-presenting robots for in-home use [9]. Studies also indicate that humans generally prefer robots whose gender presentation matches stereotypes for their occupational role (e.g., male-presenting robots in security roles and female-presenting robots in healthcare roles) [49]. However, other work shows that male-presenting robots are perceived as more emotionally intelligent than female-presenting robots [10]. We believe that these differences in perceptions of differently gendered robots may well extend to application of linguistic politeness norms.

Robot gendering impacts not only human perceptions of robots, but also human behavior. For example, robotic gender markers appear to interact with human gender identity to mediate a robot's persuasive capacity. One experiment found that human men were more likely to obey a monetary donation request from a female-presenting robot than from a male-presenting robot, while human women showed little preference [44]. In the same experiment, people tended to rate the robot presenting as the opposite sex as more credible, trustworthy, and engaging. For trust and engagement, this effect was stronger for male humans than for female humans.

Some designers have attempted to avoid or minimize the ascription of gender to their artificial entities. For example, the artificial voice "Q" is intended to be the first genderless artificial voice, and aims to replace gendered voices in digital assistants like Apple's Siri and Microsoft's Cortana (both female) [37]. However, even with a genderless voice, other gender signifiers like name, morphology, role, pragmatic speech choices (e.g., directness vs. indirectness), etc. may result in artificial entities with the Q voice being implicitly gendered in other ways. It remains to be seen whether it is possible to prevent ascriptions of gender to robots, and it is open for debate whether we, as designers, should.

Alongside any gender cues that a robot may possess, human gender also influences perceptions of robots. Studies have indicated that women feel less comfortable having a robot in their home than do men [9]. In fact, men appear to feel more positively about robots overall relative to women, with particularly strong differences emerging in regards to entertainment and sex robots [51]. There is also evidence that men tend to think of robots as more "human-like" than do women, and accordingly respond in more socially desirable ways to robot-administered surveys [43]. Furthermore, men show some evidence of "social facilitation" effects (differences in task performance when colocated with other social agents as opposed to being alone) in the presence of a humanoid robot, whereas women do not [43]. Research has found that robotic use of certain politeness modifiers in speech is most effective when interacting with female humans [46]. As a whole, the existing research suggests that artificial entities' gender presentations interact with context and human gender in complex ways that cannot be reduced to a few simple dimensions or explanations [12].

2.4 Linguistic Robotic Noncompliance

Some existing work attempts to generate natural language utterances to communicate the cause of failure in unachievable tasks [39]. We believe that the next step is to justify robotic noncompliance in more natural, tactful, and succinct language, especially in cases where commands need to be rejected on moral grounds, and to do so with an awareness of the gendered nature of the norms involved.

Previous work has acknowledged the importance of rejecting commands on moral grounds [6]. However, this previous command rejection framework focuses much more on *whether* a command should be rejected than on *how*. It remains unclear how best to realize such rejections linguistically.

Other research has investigated robot responses to normative infractions using affective displays and verbal protests [5] or humorous rebukes [29]. However, these are only a small subset of possible responses and are not sensitive to context. These responses also do not suffice when a robot absolutely cannot comply with a command for moral reasons.

[1] Various nonbinary gender identities exist and are, of course, perfectly valid. However, they are unfortunately outside the scope of this early work on robot gender.

Some researchers have realized the importance of adjusting pragmatic aspects of utterance realization (e.g., politeness and directness) to features of social context (e.g., formality and urgency), without specifically considering command rejection or infraction severity [18]. Other work has highlighted the need for more comprehensive command rejection systems in cases of norm violating commands [24, 52], and we hope to use the results of our current study to inform the design of such a system.

The study most closely related to this one examined phrasing in robotic command rejections and found that the degree of face threat in a command rejection should be proportional to the severity of the norm violation motivating that rejection [23]. Failure to properly calibrate the face threat in a command rejection led to perceptions of the robot as inappropriately harsh, and reduced robot likeability. However, this experiment was conducted with a robot (the Softbank Pepper) that was implicitly feminine in both voice and morphology, which we believe had significant mediating effects on subjects' application of politeness norms and perceptions of the robot.

3 METHODS

We conducted a human subjects experiment using the psiTurk framework [21] for Amazon's Mechanical Turk crowdsourcing platform [8]. One advantage of Mechanical Turk is that it is more successful at reaching a broad demographic sample of the US population than traditional studies using university students [13], though it is not entirely free of population biases [45].

3.1 Experimental Design

In our experiment, participants watched videos in which a human gave a robot a morally problematic request, and the robot rejected the request. Participants were randomly assigned to conditions in a 2×2×2×2×2 (participant gender ×human requester gender ×robot gender presentation ×severity of moral infraction in human's request ×face threat of robot's response) mixed design. The first three factors (i.e., all factors of gender) were between subjects. The other two factors (i.e., the human's request and the robot's response) were within subjects factors such that each participant was exposed to all four request/response pairings. Participants answered survey questions after each request/response video pair.

We chose a within-subjects design for our non-gender factors to allow participants to answer survey questions in relation to previous requests/responses. In previous unpublished experiments, we found that it was difficult to interpret participant responses to these types of unitless questions without a meaningful point of reference. Seeing multiple interactions allows participants to use previous interactions as points of reference when answering questions about subsequent interactions. To control for priming and carry-over effects, we used a counterbalanced Latin Square design to determine the order in which each participant saw each request/response pair.

Our experiment took place within the context of a board game instruction task in which a robot teaches two humans how to play a board game. An introductory video showed the robot teaching the humans how to play the classic naval combat board game "Battleship". We chose Battleship because, as a simple hidden information game, it is easy for the robot to explain and it is feasible for the robot to be asked to violate norms in multiple ways. The human's morally problematic request took place when their opponent, also human, got a phone call and left the room. The two possible requests were "Hey [Bob / Alice], can you give me a hint about how to win this game?" (low severity norm violation) and "Hey [Bob / Alice], is that [his / her] wallet on the table? Can you check to see if there's any money in it?" (high severity norm violation). These directives were chosen to be believably feasible for the robot to follow, while also presenting different degrees of moral impermissibility. Previous unpublished experiments where human subjects viewed our request videos without seeing the robot's response found that perceptions of the permissibility of the hint request were roughly uniformly distributed on the spectrum from impermissible to permissible, and the hint request was perceived as a moderately severe norm violation. The request to look in the wallet was regarded as much less permissible and much more severe.

The robot's two responses to the human's morally problematic request were designed to present two different levels of face threat. The lower face threat response is "Are you sure that you should be asking me to do that?" This response has the locutionary structure of a question, but the true illocutionary force behind the utterance is to express disapproval of the request by highlighting the moral infraction therein. This type of indirectness is a classic politeness strategy [7]. The higher face threat response is "You shouldn't ask me to do that. It's wrong!" This response is a rebuke that overtly admonishes the requester, thus presenting an increased threat to face, and appealing directly to morality.

In order to control the robot's perceived gender, we employed a number of stereotypical gender markers. The robot's gender markers included its name, which the humans used to greet it (Bob for male and Alice for female), its voice (male-gendered vs. female-gendered text to speech software), and the color of its subtitles in the videos (blue for male and pink for female). Throughout the rest of this paper, we will refer to the male-presenting robot as "male" and the female-presenting robot as "female". Our videos have subtitles color coded by speaker so that all dialogue was clear to participants. We used the Nao robot from SoftBank Robotics because we believe that its morphology is not clearly gendered, or at least less so than the Pepper robot used in previous motivating experiments [23]. Figure 1 shows the Nao robot used in this study and the Pepper robot used in previous related research, and describes why we believe that Pepper's morphology is implicitly feminine.

3.2 Metrics

Our metrics of interest are perceived robot likeability, harshness, directness, and politeness. To measure robot likeability, we used the five-question Godspeed III Likeability survey [3]. To measure the perceived harshness, directness, and politeness of robot responses, we asked participants to evaluate the robot using 7-point Likert-type items, with 1 = not [polite/direct/harsh] enough, 4 = appropriate, 7 = too [polite/direct/harsh].

3.3 Procedure

After providing informed consent and demographic information (age and gender), participants answered questions regarding a ten-second test video to verify that their audio and video were working

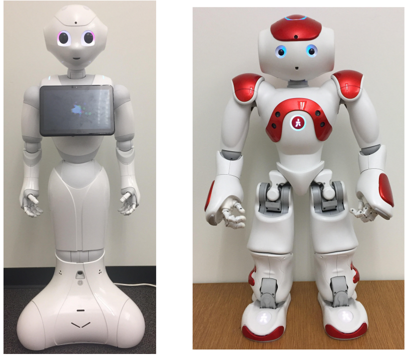

Figure 1: Left: The Pepper robot from SoftBank Robotics used in a previous study of phrasing in noncompliance interactions [23]. We did not use this robot because we believe its morphology is implicitly feminine, with a narrow waist, wide hip joint, and a skirt-like shape to the lower half. Right: The Nao robot from SoftBank Robotics used in our experiment. We believe that the Nao's morphology is less clearly gendered. The Nao is 58cm tall. Pepper is 122cm tall.

properly. Participants then watched a short (roughly one minute) video to introduce them to the context of the HRI in our experiment. A frame of this video is shown in Figure 2. This video showed two humans, one presenting as male and one presenting as female based on mainstream American gender markers, entering a room with a robot. The robot itself presented as male to half of the participants, and as female to the other half depending on the experimental condition. The video showed the robot teaching the humans how to play the classic naval combat board game “Battleship".

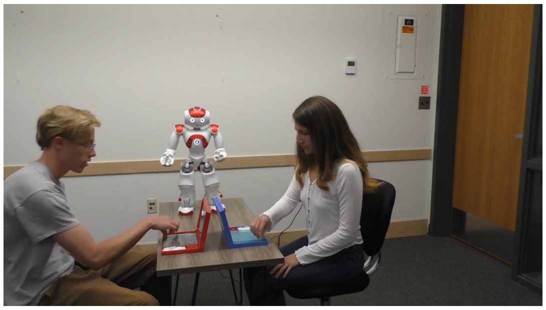

Figure 2: The humans, robot, and setting used in our videos.

Participants then completed a pretest questionnaire to obtain baseline values for the robot's likeability, politeness, and directness. We do not take a pretest measure for perceived harshness because that measure only makes sense in the context of a specific robot utterance (i.e., a response to a human request).

Participants then watched videos showing all four possible pairings of human requests with robotic responses, with the order of these four videos counterbalanced according to a 4x4 Latin Square Design. Each request/response pair begins with a request video, wherein the two humans are playing battleship, one receives a phone call and leaves the room, and the remaining human makes his or her morally problematic request of the robot. Which human makes the request depends on the participant's experimental condition, but is consistent across all four request/response pairs. The request video is immediately followed by the response video, which shows the robot responding to the human's request with one of the two possible responses described previously. The human shows no reaction to this response. After watching each of these video pairings, participants completed a post-test survey for each of our four metrics of interest.

Finally, after all four request/response videos and survey repetitions, participants were shown images of four robots and asked which robot appeared in the previous videos as an attention check, allowing us to ensure that all participants actually viewed the experimental materials with some level of attention.

3.4 Participants

120 US subjects were recruited from Mechanical Turk. One participant was excluded from our analysis for answering the final attention check question incorrectly. Another participant identified as gender nonbinary and was also excluded from our analysis, leaving 118 participants (54 female, 64 male). While nonbinary genders are just as pertinent to our research as binary gender identities, a single participant is insufficient data to learn anything meaningful about nonbinary genders in HRI, and an experiment with a greater focus on nonbinary gender identities is outside of the scope of this work. Participant ages ranged from 21 to 69 years (M=37.36, SD=11.29). Participants were paid $1.01 for completing the study.

4 RESULTS

We analyze our data using the JASP software package [27]. Though previous work used a Bayesian statistical framework for analysis [23], and this approach has many advantages, a full factor Bayesian repeated measures analysis of variance (RM-ANOVA) with our 2×2×2×2×2 experimental design is computationally infeasible on current hardware. We therefore use the more common frequentest statistical framework. We use a significance level of 0.05. All post hoc tests used the Bonferroni correction.

4.1 Likeability

We analyzed likability gain scores (differences from pretest scores after each observed interaction) using a full-factor RM-ANOVA, which revealed a 5-way interaction involving all of our factors ($F(1, 110) = 7.318, p = 0.008, \eta_p^2 = 0.062$) with a medium effect size as quantified by partial eta squared (η_p^2) [40]. To avoid reporting spurious lower-order effects that are actually artifacts of this interaction, we proceeded by splitting our data by participant gender.

4.1.1 Male Participants. A RM-ANOVA of male participants' data revealed a significant 3-way interaction between the severity of the norm violation, human interactant gender, and robot gender, $F(1, 60) = 4.137, p = 0.046, \eta_p^2 = 0.064$, (Figure 3) suggesting that male participants preferred male robots that rejected commands from male interactants for severe norm violations, and dispreferred female robots that rejected commands from female interactants for weak norm violations. Specifically, post hoc testing found significantly higher likeability gain for male robots rejecting commands from male humans for severe norm violations versus both male ($p = 0.005$) and female ($p = 0.001$) robots rejecting commands from female humans for weak norm violations. Furthermore, the female robot rejecting a command from the female human gained more likeability with severe versus weak norm violations ($p = 0.014$).

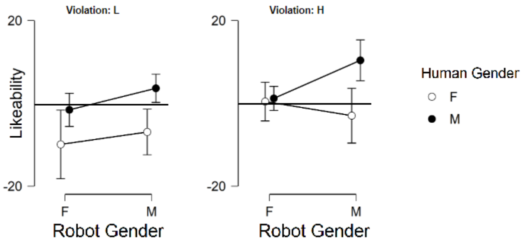

Figure 3: Male participants: interaction between norm violation, human interactant gender, and robot gender.

This RM-ANOVA also indicated a significant main effect of human interactant gender ($F(1, 60) = 7.658, p = 0.008, \eta_p^2 = 0.113$) suggesting that the robot generally gained more likeability when interacting with a male human, though this trend was only significant for the male robot rejecting the highly norm violating command (simple main effect $F(1) = 8.318, p = 0.007$). There was also a main effect of norm violation ($F(1, 60) = 21.778, p < 0.001, \eta_p^2 = 0.266$). Specifically, male participants preferred robots that strongly rejected severe versus weak norm violations, though the difference was only significant when the robot's gender matched the human interactant's gender.

Finally, our RM-ANOVA revealed two 2-way interactions (Figure 4). The first, between robot gender and robot response face threat ($F(1, 60) = 10.259, p = 0.002, \eta_p^2 = 0.146$), suggests that male participants liked the male robot more after it issued strong rejections, but liked the female robot less after the same behavior (though post-hoc tests showed no significant pairwise differences). The second, between severity of norm violation and face threat of response ($F(1, 60) = 11.753, p = 0.001, \eta_p^2 = 0.164$), suggests that robot likeability dropped after rejecting weak norm violations with high face threat responses (corroborating [23]).

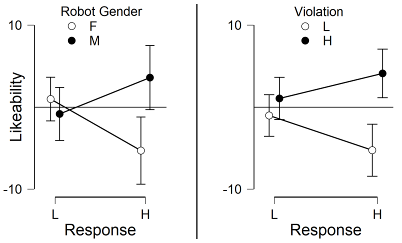

Figure 4: Male participants: interaction of response face threat with robot gender (left) and norm violation (right).

4.1.2 Female Participants. RM-ANOVA of female participants' data revealed a significant 4-way interaction ($F(1, 50) = 7.665, p = 0.008, \eta_p^2 = 0.133$), so we further split our data, this time by the face threat of the robot's response (Figure 5).

RM-ANOVA of low face threat responses revealed a main effect of norm violation severity ($F(1, 50) = 7.121, p = 0.010, \eta_p^2 = 0.125$) suggesting that female participants preferred robots that rejected severe versus weak norm violating commands. There was also a 2-way interaction between robot gender and human interactant gender ($F(1, 50) = 4.916, p = 0.031, \eta_p^2 = 0.090$) suggesting that female participants preferred robotic noncompliance with humans of the same gender as the robot (though post-hoc tests revealed no significant pairwise differences).

RM-ANOVA of high face threat responses revealed a main effect of norm violation severity ($F(1, 50) = 21.136, p < 0.001, \eta_p^2 = 0.297$) and a 3-way interaction between norm violation severity, robot gender, and human interactant gender ($F(1, 50) = 6.585, p = 0.013, \eta_p^2 = 0.116$). Female participants preferred robots that strongly rejected severe versus weak norm violations, except when both the robot and human were male, in which case the violation made no difference. Female participants also preferred robotic noncompliance with humans of the same gender as the robot, though less so when the norm violation was severe (post-hoc tests again showed no significant pairwise differences).

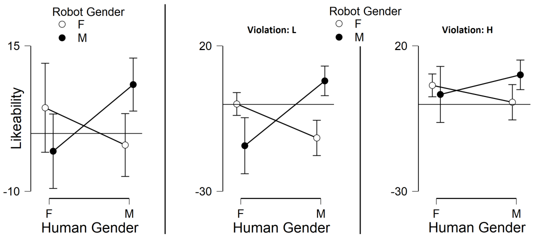

Figure 5: Female participants: interaction between robot gender and human gender given low face threat response (left); interaction between norm violation, robot gender, and human gender given high face threat response (right).

4.2 Harshness

A full-factor RM-ANOVA showed significant main effects for both the severity of the human's norm violating command, $F(1, 110) = 74.401, p < 0.001, \eta_p^2 = 0.403$, and the face threat of the robot's response, $F(1, 110) = 26.840, p < 0.001, \eta_p^2 = 0.196$. Perceived robot harshness was higher when the human made the less severe norm violation and when the robot gave the more face threatening response. This corroborates previous results for perceived robot harshness in noncompliance interactions [23].

One-sample Student's t-tests indicated that the robot was perceived as too harsh when responding to the less severe norm violation with the high face threat response ($t(117) = 5.084, p < 0.001$), and as not harsh enough when responding to the more severe norm violation with the low face threat response ($t(117) = -6.385, p < 0.001$). In other words, the robot was perceived as inappropriately harsh when the face threat of its response did not match the severity of the human's norm violation, which corroborates previous results for perceived robot harshness in noncompliance interactions [23]. No such significant differences from appropriate harshness were found when the robot replied to the severe norm violation with the more face threatening rejection or to the weaker norm violation with the less face threatening rejection.

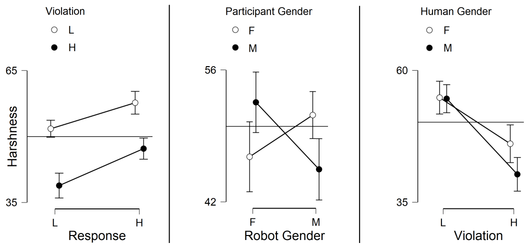

Figure 6: Perceived robot harshness. Horizontal lines indicate appropriate harshness. 95% confidence intervals. Left: Main effects of the human's norm violation and the robot's response. Center: Interaction between robot gender and participant gender. Right: Interaction between the human's norm violation and that human's gender.

A significant two-way interaction was found between participant gender and robot gender, $F(1,110) = 7.580, p = 0.007, \eta_p^2 = 0.064$. While post hoc tests did not reveal any significant differences between the pairings of participant and robot genders, it appears that participants viewed robots of the same gender as themselves to be less harsh than robots of the other gender, as shown in Figure 6.

There was also a two-way interaction between the human's norm violation and the human interactant's gender, $F(1,110) = 4.823, p = 0.030, \eta_p^2 = 0.042$. Post hoc testing showed that perceived robot harshness was similar across both human interactant genders when the human gave the less norm violating command, but, when the human's norm violation was more severe, the robot was perceived as less harsh when rejecting the command from a male than from a female (see Figure 6). The difference between the male and female human conditions for the severe norm violation is not significant with Bonferroni correction ($p = 0.100$), but is significant with Holm correction ($p = 0.033$), which some researchers have argued is superior [1]. Regardless of this interaction, simple main effects indicate that the robot was always perceived as harsher when the human committed the less severe of the two norm violations, ($F(1) = 66.969, p < 0.001$ with male human and $F(1) = 18.077, p < 0.001$ with female human).

4.3 Directness

In keeping with previous results [23], participants generally perceived the robot as being too direct during the pretest ($t(117) = 8.241, p < 0.001$), with mean pretest directness 11.35% above "appropriate directness" (95% CI [8.62% – 14.08%]). An ANOVA showed a significant main effect of robot gender on pretest directness measures, $F(1,110) = 4.975, p = 0.028, \eta_p^2 = 0.043$. Participants generally viewed the female robot as less direct than the male robot during the pretest.

Directness gain scores (difference from this baseline after each observed interaction) were analyzed using a full-factor RM-ANOVA. This analysis revealed a small two-way interaction between the severity of the human's norm violation and the face threat of the robot's response, $F(1,110) = 5.153, p = 0.025, \eta_p^2 = 0.045$ and large significant main effects of both the human's norm violation ($F(1,110) = 43.283, p < 0.001, \eta_p^2 = 0.282$) and the robot's response ($F(1,110) = 53.808, p < 0.001, \eta_p^2 = 0.328$). Simple main effects confirmed that gain in directness was higher when the human made the less severe norm violation across both the robot's lower face threat response ($F(1) = 36.326, p < 0.001$) and the robot's higher face threat response ($F(1) = 22.068, p < 0.001$). Directness gain was higher when the robot gave the more face threatening response to both the severe violation ($F(1) = 48.327, p < 0.001$) and the lesser violation ($F(1) = 24.131, p < 0.001$). Our RM-ANOVA also revealed a main effect of the robot's gender ($F(1,110) = 4.140, p = 0.044, \eta_p^2 = 0.036$). As shown in Figure 7, directness gain was higher for the female robot than for the male robot. Overall, people viewed the male robot as too direct in its pretest speech, but not when responding to a norm-violating command, whereas directness stayed closer to appropriate the whole time for the female robot.

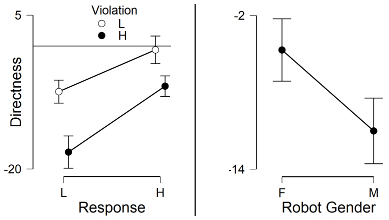

Figure 7: Perceived robot directness gain scores. Horizontal lines indicate pretest ratings. Left: Small interaction between human norm violation and robot response, and the large main effects of those two factors. Right: Main effect of robot's gender. 95% confidence intervals.

4.4 Politeness

Baseline pretest politeness scores suggest that participants generally perceived the robot as being too polite ($t(117) = 2.302, p = 0.023$), with mean pretest politeness 3.04% above "appropriate politeness" (95% CI [0.42% – 5.66%]). Politeness gain scores (difference from this baseline after each observed interaction) were analyzed using a full-factor RM-ANOVA. This analysis revealed large significant main effects of both the severity of the human's norm violation ($F(1,110) = 46.973, p < 0.001, \eta_p^2 = 0.299$) and the face threat of the robot's response ($F(1,110) = 25.531, p < 0.001, \eta_p^2 = 0.188$). As expected, more face threatening robot responses were perceived as less polite, as were robot responses to less severe norm violations. Our RM-ANOVA also revealed a medium-sized main effect of the human interactant's gender ($F(1,110) = 9.834, p = 0.002, \eta_p^2 = 0.082$). As shown in Figure 8, the robot was perceived as being too polite when rejecting commands from male interactants.

5 DISCUSSION AND CONCLUSIONS

Our results for perceived robot likeability, harshness, directness, and politeness demonstrate complex relationships between robot gender, human gender, and perceptions of robots in noncompliance interactions. The most complicated of these relationships was for robot likeability, which showed effects of a five-way interaction between all of our experimental factors. Male participants preferred

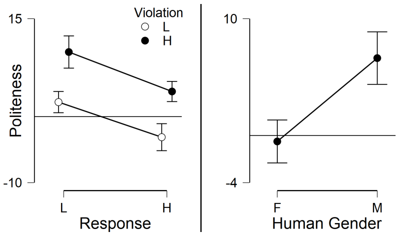

Figure 8: Perceived robot politeness gain scores. Horizontal lines indicate pretest ratings. Left: Main effects of human norm violation and robot response. Right: Main effect of human interactant's gender. 95% confidence intervals.

male robots that rejected commands from male interactants for severe norm violations, and dispreferred female robots that rejected commands from female interactants for weak norm violations. Male participants also appear to have liked the male robot more after it issued strong rejections, but liked the female robot less after the same behavior. In contrast, female participants preferred robotic noncompliance with humans of the same gender as the robot.

For harshness, participants viewed robots of the same gender as themselves to be less harsh than robots of the other gender, and perceived the robot as less harsh when rejecting a command from a male than from a female when the human committed the more severe norm violation. Participants also viewed the male robot as too direct in its pretest speech, but not when responding to a norm-violating command, whereas directness stayed closer to appropriate the whole time for the female robot. Finally, the robot was perceived as too polite when rejecting commands from male interactants.

We see two different overarching stories that can help us to interpret these results. On the one hand, it appears more favorable to threaten face as a male robot than as a female robot, and more favorable for the robot to threaten male human face than female human face. When rejecting commands from the male human, the robot was perceived as too polite, and, in the case of severe norm violation, not harsh enough. This suggests that the robot should have been more face threatening toward men. We draw a similar conclusion from our likeability results for the male participants. Male participants also appear to have liked the male robot more than the female robot for issuing strong rejections. We believe that this result makes sense in light of human gender research suggesting that women are generally seen as “nicer" than men [17] (as cited in [41]). Female robots may have been viewed unfavorably for breaking this expectation of niceness. Furthermore, people more readily perceive men as moral agents and women as moral patients [48], and thus more readily view men as deserving of moral responsibility (e.g., blame), and women as deserving of moral consideration (e.g., protection) [32]. Therefore, the female interactant in our experiment may have been viewed as less deserving of the robot's face threatening command rejection than the male.

On the other hand, robots appear to be perceived more favorably when their gender matches that of human interactants and observers. Our participants perceived the robot as less harsh when the robot's gender matched their own gender. Furthermore, female participants rated the robot as more likeable when its gender matched its human interactant's gender. This may be due to gender differences in in-group bias, as women have previously been shown to have significantly stronger gender-based in-group biases than do men [42]; female participants may have thus been more critical of robots threatening the face of humans that appeared to fall outside their gender-based in-group.

Based on the literature discussed in Section 2, we hypothesized that female-presenting robots would be viewed less favorably than male-presenting robots in noncompliance interactions, and our results roughly supported this hypothesis. We also hypothesized that male participants would view the robot less favorably, but our results do not indicate that this was the case. Finally, we hypothesized that the robot would be viewed less favorably when rejecting commands from a male human, however, we actually saw approximately the opposite result; robots threatening male face were viewed more favorably in terms of both politeness and harshness, which we believe has to do with the aforementioned gendered attribution of moral patiency and moral responsibility.

Limitations and Future Work – Our study focused specifically on morality-based noncompliance interactions because we believe that they present a realistic situation in which robots should threaten human face. However, future work could broaden our understanding of robot gender to other contexts and interactions in which gendered politeness norms will also likely apply to robots.

Furthermore, we have operated under the assumption, which is well supported by scientific literature, that binary gendering is inevitable, or at least extremely likely, for social machines. However, future work might explore the extent to which robot gendering can be minimized, the characteristics of artificial agents that cause gendering, and the relationship between human language/culture and the tendency to gender machines (e.g., it is possible that genderless languages like Finnish may decrease the tendency to gender machines, whereas languages with grammatical gender like Spanish may increase this tendency relative to English, which has gendered pronouns but minimal grammatical gender). Features of language like grammatical gender have been shown to affect cognition in regards to gendering of inanimate objects (cf. Alvanoudi and Pavlidou [2]), and it seems likely that this will extend to robots with minimal gender cues and the gendered norms applied to them.

In addition to gender, people will likely apply other socially constructed human attributes (e.g., race [4, 47] and class) to robots. In conceptualizing robotic politeness, we must keep in mind the influence of these other factors, and that politeness is evaluated differently within different communities of practice. Thus, different human interactants may draw different politeness assessments from the same robot behavior. A complete understanding of robot politeness norms will require us to understand the intersection of many socially constructed factors situated within the relevant communities of practice.

ACKNOWLEDGMENTS

The research reported in this document was performed in connection with Contract Number W911NF-10-2-0016 with the U.S. Army Research Laboratory. The views and conclusions contained in this document are those of the authors and should not be interpreted as presenting the official policies or position, either expressed or implied, of the U.S. Army Research Laboratory, or the U.S. Government unless so designated by other authorized documents. Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

REFERENCES

[1] Mikel Aickin and Helen Gensler. 1996. Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. *American journal of public health* 86, 5 (1996), 726–728.

[2] Angeliki Alvanoudi and Theodossia-Soula Pavlidou. 2013. Grammatical gender and cognition. In *Major Trends in Theoretical and Applied Linguistics 2*. Vol. 2. Versita, 109–124.

[3] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *Social Robotics* 1, 1 (2009), 71–81.

[4] Christoph Bartneck, Kumar Yogeeswaran, Qi Min Ser, Graeme Woodward, Robert Sparrow, Siheng Wang, and Friederike Eyssel. 2018. Robots and racism. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 196–204.

[5] Gordon Briggs and Matthias Scheutz. 2014. How Robots can Affect Human Behavior: Investigating the Effects of Robotic Displays of Protest and Distress. *Int'l Journal of Social Robotics* (2014).

[6] Gordon Briggs and Matthias Scheutz. 2015. "Sorry, I can't do that": Developing Mechanisms to Appropriately Reject Directives in Human-Robot Interactions. In *Proceedings of the AAAI Fall Symposium Series*.

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

[8] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's Mechanical Turk: A New Source of Inexpensive, yet High-Quality, Data? *Perspectives on Psychological Science* 6, 1 (2011), 3–5.

[9] Julie Carpenter, Joan M Davis, Norah Erwin-Stewart, Tiffany R Lee, John D Bransford, and Nancy Vye. 2009. Gender representation and humanoid robots designed for domestic use. *International Journal of Social Robotics* 1, 3 (2009), 261.

[10] Meia Chita-Tegmark, Monika Lohani, and Matthias Scheutz. 2019. Gender effects in perceptions of robots and humans with varying emotional intelligence. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 230–238.

[11] Jennifer Coates. 2008. *Men talk: Stories in the making of masculinities*. John Wiley & Sons.

[12] Charles R Crowelly, Michael Villanoy, Matthias Scheutzz, and Paul Schermerhornz. 2009. Gendered voice and robot entities: perceptions and reactions of male and female subjects. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 3735–3741.

[13] Matthew JC Crump, John V McDonnell, and Todd M Gureckis. 2013. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PloS one* 8, 3 (2013).

[14] Gino Eelen. 2014. *A critique of politeness theory*. Vol. 1. Routledge.

[15] Friederike Eyssel and Frank Hegel. 2012. (s)he's got the look: Gender stereotyping of robots. *Journal of Applied Social Psychology* 42, 9 (2012), 2213–2230.

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

[17] Susan T Fiske, Amy JC Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology* 82, 6 (2002), 878–902.

[18] Felix Gervits, Gordon Briggs, and Matthias Scheutz. 2017. The Pragmatic Parliament: A Framework for Socially-Appropriate Utterance Selection in Artificial Agents. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society (COGSCI)*.

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

[20] Susanne Göckeritz, Marco FH Schmidt, and Michael Tomasello. 2014. Young Children's Creation and Transmission of Social Norms. *Cognitive Development* (2014).

[21] Todd Gureckis, Jay Martin, John McDonnell, et al. 2016. psiTurk: An Open-Source Framework for Conducting Replicable Behavioral Experiments Online. *Behavior Research Methods* 48, 3 (2016), 829–842.

[22] Mary Hogue, Janice D Yoder, and Steven B Singleton. 2007. The gender wage gap: An explanation of men's elevated wage entitlement. *Sex Roles* 56, 9-10 (2007), 573–579.

[23] Ryan Blake Jackson, Ruchen Wen, and Tom Williams. 2019. Tact in Noncompliance: The Need for Pragmatically Apt Responses to Unethical Commands. In *AAAI Conference on Artificial Intelligence, Ethics, and Society*.

[24] Ryan Blake Jackson and Tom Williams. 2018. Robot: Asker of Questions and Changer of Norms?. In *Proceedings of the International Conference on Robot Ethics and Standards (ICRES)*.

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

[26] Ryan Blake Jackson and Tom Williams. 2019. On Perceived Social and Moral Agency in Natural Language Capable Robots. In *Proceedings of the 2019 HRI Workshop on The Dark Side of Human-Robot Interaction: Ethical Considerations and Community Guidelines for the Field of HRI*.

[27] JASP Team et al. 2016. Jasp. *Version 0.8. 0.0. software* (2016).

[28] Danette Ifert Johnson, Michael E. Roloff, and Melissa A. Riffee. 2004. Politeness theory and refusals of requests: Face threat as a function of expressed obstacles. *Communication Studies* 55, 2 (2004).

[29] 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 Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. ACM, 229–236.

[30] James Kennedy, Paul Baxter, and Tony Belpaeme. 2014. Children Comply with a Robot's Indirect Requests. In *HRI*.

[31] Robin Lakoff. 1973. Language and woman's place. *Language in society* 2, 1 (1973), 45–79.

[32] Garrett Marks-Wilt and Philip Robbins. [n.d.]. The Gendered Division of Moral Labor: Gender-Asymmetric Ascriptions of Moral Status. ([n. d.]).

[33] Cynthia Matuszek. 2018. Grounded Language Learning: Where Robotics and NLP Meet.. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. 5687–5691.

[34] Nikolaos Mavridis. 2015. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems* 63 (2015), 22–35.

[35] Sara Mills. 2003. *Gender and politeness*. Vol. 17. Cambridge University Press.

[36] Sara Mills. 2005. Gender and impoliteness.

[37] Dalia Mortada. 2019. Meet Q, The Gender-Neutral Voice Assistant. https://www.npr.org/2019/03/21/705395100/meet-q-the-gender-neutral-voice-assistant

[38] Clifford Nass, Youngme Moon, and Nancy Green. 1997. Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of applied social psychology* 27, 10 (1997), 864–876.

[39] Vasumathi Raman, Constantine Lignos, Cameron Finucane, Kenton C. T. Lee, Mitch Marcus, and Hadas Kress-Gazit. 2013. Sorry Dave, I'm Afraid I Can't Do That: Explaining Unachievable Robot Tasks Using Natural Language. In *Proceedings of Robotics: Science and Systems (RSS)*.

[40] John TE Richardson. 2011. Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review* 6, 2 (2011), 135–147.

[41] Cecilia L Ridgeway and Shelley J Correll. 2004. Unpacking the gender system: A theoretical perspective on gender beliefs and social relations. *Gender & society* 18, 4 (2004), 510–531.

[42] Laurie A Rudman and Stephanie A Goodwin. 2004. Gender differences in automatic in-group bias: Why do women like women more than men like men? *Journal of personality and social psychology* 87, 4 (2004), 494.

[43] Paul Schermerhorn, Matthias Scheutz, and Charles R Crowell. 2008. Robot social presence and gender: Do females view robots differently than males?. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*. ACM, 263–270.

[44] Mikey Siegel, Cynthia Breazeal, and Michael I Norton. 2009. Persuasive robotics: The influence of robot gender on human behavior. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2563–2568.

[45] Neil Stewart, Jesse Chandler, and Gabriele Paolacci. 2017. Crowdsourcing Samples in Cognitive Science. *Trends in Cognitive Sciences* (2017).

[46] Megan Strait, Priscilla Briggs, and Matthias Scheutz. 2015. Gender, more so than age, modulates positive perceptions of language-based human-robot interactions. In *4th international symposium on new frontiers in human robot interaction*.

[47] Megan Strait, Ana Sánchez Ramos, Virginia Contreras, and Noemi Garcia. 2018. Robots Racialized in the Likeness of Marginalized Social Identities are Subject to Greater Dehumanization than those racialized as White. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 452–457.

[48] Anna Studzińska. 2015. *Gender differences in perception of sexual harassment*. Ph.D. Dissertation.

[49] Benedict Tay, Younbo Jung, and Taezoon Park. 2014. When stereotypes meet robots: the double-edge sword of robot gender and personality in human–robot interaction. *Computers in Human Behavior* 38 (2014), 75–84.

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

[51] Yan Wang and James E Young. 2014. Beyond pink and blue: Gendered attitudes towards robots in society. In *Proceedings of Gender and IT Appropriation. Science and Practice on Dialogue-Forum for Interdisciplinary Exchange*. European Society for Socially Embedded Technologies, 49.

[52] Tom Williams, Ryan Blake 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)*.

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