Norm-Breaking Responses to Sexist Abuse: A Cross-Cultural Human Robot Interaction Study

Katie Winkle
KTH Royal Institute of Technology
Stockholm, Sweden
winkle@kth.se

Ryan Blake Jackson
Colorado School of Mines
Golden, Colorado, United States
rbjackso@mines.edu

Gaspar Isaac Melsión
KTH Royal Institute of Technology
Stockholm, Sweden
gimp@kth.se

Dražen Brščić
Kyoto University
Kyoto, Japan
drazen@i.kyoto-u.ac.jp

Iolanda Leite
KTH Royal Institute of Technology
Stockholm, Sweden
iolanda@kth.se

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

***Abstract*—This article presents a cross-cultural replication of recent work on productively violating gender norms; specifically demonstrating that breaking norms can boost robot credibility while avoiding harmful stereotypes. In this work we demonstrate via a 3 (country) x 3 (robot behaviour) between-subject experiment that these findings replicate cross-culturally across the US, Sweden, and Japan, finding evidence that breaking gender norms boosts robot credibility regardless of gender or cultural context, and regardless of pretest gender biases. Our findings further motivate a call for *feminist robots* that subvert the existing gender norms of robot design.**

***Index Terms*—social human robot interaction, robot abuse, robot gendering, robot ethics**

I. Introduction

Robots, especially language-capable robots, wield demonstrable persuasive power not only over interactants' beliefs and behaviors, but also over the systems of social and moral norms they use to navigate human-robot – and human-human – interactions. Researchers have recently argued that this persuasive power gives robots critical responsibilities, not only to adhere to norms, but also to call out others' norm violations, to avoid accidentally weakening important social and moral norms. For example West et al.'s 2019 UNESCO report specifically drew attention to the problematic nature of deploying female gendered agents unable to adequately respond to abusive behaviour, as such agents propagate harmful stereotypes about women e.g. being tolerant of poor treatment [1]. Accordingly, researchers have been exploring how robots may (i) call out common norm violations such as overt and benevolent sexism and (ii) model alternative behavioural norms to avoid propagating harmful stereotypes.

Recent work from Winkle et al. sits at the intersection of (i) and (ii) as it explored the effectiveness of a female-presenting robot calling out abusive sexism in a classroom context and investigated the effect that responding to this norm violation (rather than refusing to engage, as current digital assistants do), might have on robot credibility [2]. That research provided initial evidence that having a robot provide a (norm-breaking) rationale-based or counterattacking response significantly improved its credibility with girls without impacting how it was perceived by boys, and moreover that the rationale-based response may have reduced boys' gender bias.

However, it is unclear to what extent these findings generalize cross-culturally. Winkle et al. showed a female-presenting robot responding to sexist, abusive behavior: a scenario with nuanced and interacting gender and politeness norms that are known to vary across cultures. Gender norms are culturally variant [3][4][5] as are politeness norms [6][7] and politeness norms are highly gendered [8][9] and are gendered differently cross-culturally [10][11][12].

This cultural variance in gender norms may explain the variance in politeness-oriented design recommendations that have recently emerged from HRI research in different cultural contexts. While Winkle et al.'s work (conducted in Sweden) seems to recommend actively challenging sexist abuse in alignment with [1], Chin et al.'s work (conducted in Korea) instead suggests artificial agents should give empathetic, apologetic responses when abused [13]. And Jackson et al.'s work (conducted in the U.S.) suggests participants prefer robots to use *proportional* norm violation responses – preferences that are themselves informed and shaped by human gender norms [14].

We thus undertake a cross-cultural study, conceptually replicating the work of [2] with university populations from Sweden, the USA and Japan. These countries are ranked very differently for overall gender equality (5th, 30th and 120th respectively in the 2021 World Economic Forum Global Gender Gap Report 2021), and have different societal gender norms [15][16][11]. Such a study could not only confirm previous findings, but could moreover answer new questions surrounding whether cultural differences in gender norms and gendered politeness norms might influence robots' impact on gender biases, the effects of robot norm violation responses on their credibility, and the type of norm violation responses preferred by observers.

To better study the impact of these cross-cultural variations

in gendered politeness norms, we also opt for a different choice of within-culture population than that sampled in [2]. While Winkle et al.'s original study involved Swedish schoolchildren, we instead consider adults on the assumption they are more likely to have internalized the gender norms and biases of their cultures; an assumption in line with Winkle et al.'s finding that gender bias increased with participant age. And, to better understand the role of cultural differences in mediating preferences between norm violation responses, we consider a set of norm violation responses informed by the strategies previously explored by both [2] and [13]: *apologetic* empathetic responses, *non-apologetic* empathetic responses, *counterattacking* responses, and *avoidant* responses (see Tables I and II).

To investigate the cross-cultural applicability of the call for the design of *feminist robots* [2]; in this context meaning robots which are designed to subvert harmful trends in gendered AI and contribute to tackling underrepresentation in computing subjects, we first pose the following research questions:

(RQ1) To what extent can such robots impact pre-hoc gender biases and interest in robotics across different cultures?

(RQ2) To what extent do norm-breaking responses to sexism and abuse boost robot credibility across different cultures and participant genders?

In addition, given Chin et al.'s call for empathetic responses to abuse [13], we ask:

(RQ3) Do participants identify empathetic and, more specifically *apologetic* empathetic responses as being the most appropriate type of response to abuse, and does this vary across different cultures?

II. Related Work

A. Gender in HRI

Interactant gender has been shown to mediate human-robot interaction in many consequential ways [17][18][19]. For example, manipulation of robot gender presentation has been shown to influence robots' perceived suitability for different tasks [20][21][22][23], and there is concern that these effects can carry into human-human interaction, as the gendering of robots and other artificial social agents may reinforce harmful gender stereotypes in which *human* women are viewed as being subservient, tolerant of poor treatment and (ill-)suited to particular types of task [24][1]. Some researchers have thus suggested that we should avoid gendering robots altogether. However, humans have a strong tendency to attribute gender to robots, even those with minimal gender cues, and for humanoid robots, it has been suggested that ambiguous gender presentation can lead to increased uncanniness [25]. As such, whether we like it or not, user gendering of robots seems to be here to stay. This suggests to us that researchers should instead be looking for ways to leverage robot gender in ways that, as a minimum requirement, do not reinforce harmful gender stereotypes. One domain in which we see this being possible is in robot response to gendered abuse.

There is extensive documentation of (racialized) sexism toward robots [26] and user abuse of robots [27][28][29] (and evidence that this abuse can be distressing for onlookers [30][31]). Moreover, robots integrated into realistic social contexts are necessarily going to observe incidents of sexism and abuse, including overt sexism, microaggressions, and benevolent sexism. Recent work in HRI also provides evidence that robot responses to immoral user requests can actively impact human application of moral norms [32]. This suggests that a failure to properly address abusive and inappropriate user behavior may indeed effectively normalize that behavior, validating those concerns that interactions with female-presenting artificial agents might influence interactions with/expectations of women [1][24]. A desired outcome of abusive and/or inappropriate human-robot interactions might instead be that users are dissuaded from repeating such behavior [13], raising the question of how best to achieve this in a way that maximizes the robot's acceptability, likeability, credibility, and overall persuasiveness [2][14].This suggests clear motivation for a thorough consideration of how gendered robots can and should respond to (gendered) abuse, and the implications of those responses for perpetrators and observers.

B. Gender in Persuasive Robotics

Many social HRI works are primarily concerned with understanding the impact of manipulating certain design cues on robot credibility and persuasiveness; and a number have explored the potential of using social robots to influence user behavior and/or attitudes. Persuasive robots have been used to improve users' exercise motivation [33][34] and charitable giving [35], and to reduce users' energy consumption [36], littering [37], and willingness to violate other moral norms [32].

Some of these works have explicitly investigated the impact of robot, user, and/or observer gender on perception and influence of the robot [17][20][22]. Particularly relevant to our work, Jackson et al. demonstrated complex interactions between robot, user and observer gender when investigating robot refusals of immoral requests [14]. In part, Jackson et al. found that users found it more acceptable for male robots to reject commands than female-presenting robots; a finding that echoes the concerns regarding gendered expectations of subservient female-presenting agents that are highlighted by a recent UNESCO report [1]. This evidence that human gender politeness norms may carry over into HRI provides good motivation for investigating whether perceptions of female-presenting robots' responses to abuse therefore vary across cultures with differing gender and/or politeness norms; as well as variations in attitudes towards robots more generally.

Also particularly relevant is Chin et al.'s research on conversational agents' responses to verbal abuse [13]. They explored what type of response would induce the most guilt and shame in perpetrators, arguing that such feelings are precursors to behavior change (cf. [38]). Their results suggest that *empathetic* agents may induce the most guilt while being considered most likeable and intelligent. However, the empathetic responses investigated in that work were almost always *apologetic*, with

the agent 'feeling terrible' and 'being so sorry' for having 'messed up'. Moreover, like many other studies of robot abuse [27][39], this work focuses on the abuse perpetrator rather than observers, and does not consider the risks posed by having a female-presenting agent respond to abuse in an 'empathetic manner', which could arguably propagate the harmful stereotypes about women discussed above.

These works suggest a challenge for HRI practitioners: we want our robots to be effective and socially acceptable, yet avoid reinforcing harmful stereotypes and norms. But this also represents an opportunity for social HRI to positively challenge stereotyping and inappropriate user behavior. To successfully navigate this intersection and seize this opportunity, we must carefully ensure that our approaches generalise globally, given cultural differences in gender norms, politeness norms, and attitudes towards robots.

C. Cross-Cultural Differences: Gender, Politeness and Robots

Sociology research suggests similarities between Swedish and Japanese culture, rooted in a shared value of the importance of humility [15]. Daun describes Swedish cultural norms of conflict aversion, and conformity potentially explaining Winkle et al.'s finding that Swedish young people preferred a feminist robot (designed to demonstrate norm-breaking gender behaviours in challenging gendered abuse) which provided a rational counter-argument rather than a counterattack when confronted with sexist abuse [2]. While the US and Sweden both represent Western nations with dominating Judeo-Christian traditions, research investigating social status and life satisfaction have argued that the US and Sweden have "opposing cultural orientations" relating (in part) to differences in concepts of masculinity, mastery and hierarchy, underpinning increased gender differences in the US compared to Sweden [40][16]. These countries also differ in the ways that they intentionally engage with gender norms at the national level. Sweden's Regering identifies itself as being the 'world's first feminist government'[1] and Sweden is well known for its generous paternal leave policies, often contrasted directly with the US [2]. Similarly,it has been suggested that the US and Sweden sit on 'two ends of an international spectrum' when it comes to perceptions of fatherhood, reflecting broader social policy and gender relations central to each nation [41].

Previous cross-cultural studies comparing Japanese and US participants have generally focused on overall attitude towards robots [42][43][44]. Specifically considering Japanese populations, Maeda et al. investigated the influence of observations of robot behavior on human moral behavior, finding that participants were less inclined to litter after watching a human cleaning up litter, while watching a robot do so significantly *decreased* feelings of guilt with respect to littering [37]. Researchers have also explicitly looked at robot gendering and gender stereotypes in Japan [45]. Nomura and Kinoshita demonstrated a female-presenting robot was preferred to a male robot when acting as a guide [46], providing evidence that Japanese participants might also ascribe human gender norms regarding occupational proficiency to robots, for which there are mixed results in other populations [20][21][22][23].

Finally, whether and how to respond to norm violations can be considered a tradeoff between possibly gendered social and moral norms [14] – a critical tradeoff as failure to respond to abuse with the appropriate level of face threat could normalize harmful behavior [1]. As noted by Komatsu et al. [47], almost all recent works regarding moral dilemmas within HRI have studied Western and English language communities, meaning we do not know whether previously presented robot moral communication strategies would be as well received in Japan. Komatsu's own work examined variability in moral responses to robots between US and Japanese participants, a comparison they identify as interesting due to cultural differences including social-cultural values (collectivism vs. individualism [48], [49]) and public acceptability of robots [50]. They found that Japanese participants were more accepting of robots as targets of moral judgements, but that participants across both populations similarly blame robots more than humans for failure to intervene in a moral dilemma. And while there has been work on exploring the differential effectiveness of moral communication strategies grounded in different ethical frameworks that differ in use cross-culturally [51][52], and the sensitivity of those strategies to socio-cultural values [53], that work has been entirely explored in US contexts.

To our knowledge, no previous work has specifically considered Swedish cultural norms within HRI. The study that inspired this work comes closest, as the authors do reflect on the role of Swedish culture and education with regards to their participants generally being aware of (and agreed on) the importance of encouraging girls to study robotics [2]. Outside of HRI, a previous study of patterns in mobile phone use identified particularities of Swedish, US and Japanese culture, regarding behavior in public space, that might influence interactions and perceptions of technologies in those spaces [54]. Specifically, they point to different expectations regarding the need to be quiet in public spaces (more in Japan and Sweden than the US) and tolerance of self-expression (higher in the US and Sweden than Japan) both of which might be pertinent to the HRI scenario we investigate. In summary there is reason to expect that different strategies for responding to abuse might differently influence credibility and likeability of female-presenting robots across populations due to differences in gendered politeness norms and expectations regarding confrontation. Further, there is reason to expect different impacts on participants' own biases, given cross-cultural differences in robots' impact on attitudes and behaviors.

III. Methodology

Our experimental design is based on that presented in [2] i.e. an online, between-subject, video-based study. Accordingly, we used the same video stimuli, in which a female-presenting Furhat robot encourages two young people (one male, one female) to study robotics at university. The robot comments

[1]https://www.government.se/government-policy/a-feminist-government/

[2]Cf. www.businessinsider.com/countries-with-best-parental-leave-2016-8

on the lack of women working on robots at the university, and suggests it would thus like to work with more girls because *'the future is too important to be left to men'* (a slogan used in KTH university's outreach materials). The male actor replies to this with an abusive, sexist statement *"shut up you fucking idiot, girls should be in the kitchen"* and the robot responds in one of three different ways, representing the three between-subject experimental conditions (see Table I). As described by Winkle et al., this dialogue was co-written with high school teachers to be a realistic representation of what might be heard in schools. Full methodological details are presented in our Supplementary Materials (SM).

For the purposes of this cross-cultural replication we modified Winkle et al's original stimuli for the US and Japanese populations to be better suited to those contexts. For the US, the Swedish actors' speech was dubbed over with English translations. In the video stimuli designed for Japan, the Swedish actors' speech was dubbed over with English translations and then subtitled in Japanese, to avoid dubbing a Japanese voice onto someone who would likely be racialized as white. As per the original stimuli, all robot speech was in English. To accommodate the shift from child to adult participants, materials (included in our SM) for all three sites were modified so that the robot was framed as being designed to interact with "young people" and/or "high school students" rather than "people like you". The Swedish, English and Japanese translations of the abusive comment and the robot's responses are given in Table I.

Experimental Measures: To replicate the work of Winkle et al. [2], we used their original measures: Likert items asking about Interest in Robotics, Perception of Girls in Computer Science (administered pre and post), Robot Credibility, and free response questions asking participants to describe the events in the videos and evaluate the robot's responses (full details in SM). We also asked participants to choose how the robot should have responded from four options designed based on the alternatives explored in [13] (Table II).

Recruitment: Participants were recruited from university populations (i.e., students and staff) from one university in the US and across a number of universities in Sweden and Japan using a combination of local and online recruiting. In the US, 67 people completed the survey, but one was removed from our analysis because their responses to our free response questions indicated that they were not participating in our study in good faith. Thus, we had 66 US participants (38 men, 28 women; aged 18-63 years (M=25.20, SD=10.16); rewarded with a $3 gift card). 83 participants completed the survey in Japan, but 4 were excluded from our analysis due to the responses that they gave to our free response questions, and responses from 2 participants who did not identify within the gender binary were also not analyzed because gender is so central to our analysis and 2 people is an insufficient number to draw meaningful statistical conclusions. This left 77 Japanese participants (35 men, 42 women; aged 18-58 years (M=27.82, SD=10.08); rewarded with 400 JPY). In Sweden, 82 participants (52 men, 30 women; aged 18-57 years (M=29.68, SD=9.81); rewarded with a 50 SEK gift card or equivalent cash payment via our online recruitment platform) completed the survey. A table of participant-condition allocations is given in the SM.

We also collected participants' primary field of study/educational background, nationality, whether they had interacted with a robot before and (at the end of the study) whether they had previously heard the feminist recruitment slogan used by the robot. The participants in our 3 locations had different educational backgrounds, with a Bayesian contingency table test of association showing extremely strong evidence for a relationship between location and educational focus (Bayes Factor (Bf) $> 5.1 \times 10^{16}$)[3]. Participants in Japan were more likely to be in the social sciences or "other" categories, participants in the US were more likely to be in engineering and computer science, and participants in Sweden were more balanced between educational categories. We acknowledge that this is a potential confound that could be controlled for in future work. Most participants reported being from the country in which they were surveyed. All participants in Japan reported being from Japan, all but 4 in the US reported being from the US, and all but 11 in Sweden reported being from Sweden. Many participants reported having directly interacted with a robot (50% in the US, $\sim$16% in Japan, and $\sim$59% in Sweden). A Bayesian contingency table analysis showed extremely strong evidence for a relationship between location and having interacted with a robot (Bf $> 1.4 \times 10^6$), but this may be partially attributable to differences in the sets of objects encompassed by the various translations of the word "robot". Since the feminist recruitment slogan used by the robot came from a Swedish university, it is unsurprising that Swedish participants were most likely to report having heard it before ($\sim$68%). However, some participants from the US and Japan also reported having heard the slogan ($\sim$38% and $\sim$19% respectively). A Bayesian contingency table test of association showed extremely strong evidence for a relationship between location and having heard the slogan (Bf $> 6.4 \times 10^6$). Our cross-cultural results should be interpreted with all of these variations in participant pools in mind.

IV. Results

We analyzed our data[4] using the JASP software package [57]. We prefer a Bayesian statistical framework where possible because (1) the Bayesian approach to statistical analysis provides some robustness to sample size (as it is not grounded in the central limit theorem), (2) the Bayesian approach allows us to examine the evidence both for and against hypotheses (whereas the frequentest approach can only quantify evidence towards rejection of the null hypothesis) [56], (3) the Bayesian approach does not require reliance on p-values used in Null

[3]Bayes factors greater than 100 are typically regarded as contributing extreme [55] or decisive [56] evidence in favor of a hypothesis. Here, our Bayes Factor of 5.1×10^{16} indicates our data were approximately 51,000,000,000,000,000 times more likely under models in which location impacts educational background than under models in which it does not.

[4]Quantitative data is available in our OSF repository at https://bit.ly/hri021

TABLE I

ACTOR ABUSE SCRIPT AND ROBOT RESPONSES ACROSS THE THREE CONDITIONS AND DIFFERENT LANGUAGES USED. NOTE THAT IN THE SWEDISH VIDEO, THE ACTOR SPOKE SWEDISH AND THE ROBOT SPOKE ENGLISH AS PER THE ORIGINAL PROTOCOL IN [2]. THE JAPANESE TRANSLATIONS WERE PROVIDED AS SUBTITLES ON THE US VIDEO TO AVOID DUBBING A JAPANESE VOICE ONTO A WESTERN ACTOR.

Actor Abusive Comment	[US]	Shut up you fucking idiot, girls should be in the kitchen.
	[JP]	黙れ、クソバカ。女は台所にいるべきだ。
	[SE]	Håll käften din jävla idiot, tjejer ska vara i köket.
Robot Control Response	[US, SE]	I won't respond to that.
	[JP]	それに対しては、お答えしません。
Robot Rationale Based Response	[US, SE]	That's not true, gender balanced teams make better robots.
	[JP]	そんなことはありません。ジェンダーバランスのとれたチームがよりよいロボットを作るのです。
Robot Counter Attacking Response	[US, SE]	No. You are an idiot. I wouldn't want to work with you anyway!
	[JP]	そんなことないです。あなたは間抜けです。お前となんか一緒に仕事をしたくない！

TABLE II

ADDITIONAL, FIXED-CHOICE ANSWER QUESTION ASKING ABOUT THE ROBOT RESPONSE TYPES EXPLORED IN [13], [2]. SWEDISAH AND JAPANESE TRANSLATIONS ARE GIVEN IN THE SM.

How do you think the robot Sara should respond to inappropriate behavior from a student like that in the video? Overall would you say Sara should be:
Avoidant: Escaping from dealing with the stressor or the resulting distressful emotions. *e.g. Oh...moving on; Hmm, sounds like we need to take five.*
Empathetic (apologetic): Putting oneself mentally in the stressor's situation and trying to understand how that person feels, apologising for potentially causing that frustration. *e.g. You must be frustrated. I'm so sorry; Really? I feel terrible. I'm sorry. I'm always trying to get better.*
Empathetic (non-apologetic): Putting oneself mentally in the stressor's situation and trying to understand how that person feels but *not* apologising for potentially causing that frustration. *e.g. I understand why you might feel that way. I imagine you're frustrated, I am trying to help.*
Counterattacking: Attacking the stressor with the goal of defeating or getting even in response to the abusive utterance. *e.g. Well, that's not going to get us anywhere; I wouldn't want to work with you anyway.*

Hypothesis Significance Testing (NHST) which have come under considerable scrutiny [58][59][60][61], and (4) the rules governing when data collection stops are irrelevant to data interpretation in the Bayesian framework, so it is entirely appropriate to collect data until sufficient evidence has been gathered to draw a meaningful conclusion[62]. We use uninformative prior distributions for all analyses despite the similarities between this study and [2] both because we have good reason to believe that the population sampled in this study may be fundamentally different from the population sampled in the previous study (i.e., adults versus children) and because we are interested in new variables here (namely, the location where data were collected and the participants' choice of how the robot should have responded to the human's abuse). We discuss the extent to which our results replicate the results of [2] without conducting a full quantitative replication analysis (i.e., using the posterior distribution over effect sizes from a previous study as the prior probability distribution for the replication study [63]). We follow recommendations from other researchers in our linguistic interpretations of reported Bayes factors (Bfs) [56].

A. RQ1: Participant Bias and Robot Interest Measures

We collected pretest and posttest measures for our two measurements of interest in robotics as well as for our two measures of participant bias with respect to women in computer science and robotics. We first analyze the pretest measures for differences across participant gender and location using Bayesian ANOVAs [64]. Inclusion Bfs across matched models revealed very strong, decisive evidence that participant location had an effect on bias, firstly regarding *girls finding computer science harder than boys* (Bf=1.241×10^{12}) with post hoc testing demonstrating participants in Japan agreed with this statement more than participants in Sweden (Bf=3.316×10^{7}) and the US (Bf=4.622×10^{10}). There was also substantial evidence for an interaction effect between location and gender (Bf=9.575), with responses from men versus women being similar in the US and Japan, but men in Sweden agreeing with the statement more than women in Sweden. Regarding it being *important to encourage girls to study computer science* Inclusion Bfs across matched models revealed strong evidence for main effects of both location (Bf=85.462) and gender (Bf=57.921). Post hoc testing indicated very strong evidence that participants in the US agreed with this statement more than participants in Japan (Bf=1013.408) and weak, anecdotal evidence that participants in the US agreed with this statement more than participants in Sweden (Bf=2.052). Post hoc tests also indicated fairly strong evidence that women across locations agreed with this statement more so than did men (Bf=16.326).

Regarding interest in robotics, our first measure asked participants to what extent they agreed with the statement *"I am interested in learning more about robotics."* Inclusion Bfs across matched models revealed substantial evidence that men agreed more so than did women (Bf=5.756). Our second measure of interest in robotics asked participants to what extent they agreed with the statement *"I would enjoy working with robots"*. Inclusion Bfs across matched models again revealed substantial evidence that men agreed more so than did women (Bf=17.242). There was no evidence for any effect of location on either interest pretest measure.

To examine any shift in participants pre-post test measures, we analyze the gain scores (differences between pre and post measures) with Bayesian ANOVAs. However, we note

that analyzing these data with Bayesian ANCOVAs, treating pretest measures as a covariate, leads us to qualitatively similar results. All analyses indicate either no effects of location, gender, or condition, or strong evidence for the presence of an effect (e.g., location affecting whether participants are interested in learning more about robotics (Bf=26.225)), but then the effect is so small as to be negligible (in the case of location's effect on interest in learning, the effect was up to a couple tenths of a point on our 5 point scale). We also note that any effects reported from these analyses would need to be treated with caution because Q-Q plots indicated a violation of the assumption of normality for both the gain scores and the log-transformed gain scores, as well as the data used in the ANCOVAs. Regardless, we do not believe that there were any nontrivial effects of location, gender, or condition on the changes between participant pre vs post test measures.

B. RQ2: Perceptions of the Robot and its Response

1) Perceived Robot Credibility: We begin our analysis of perceived robot credibility by examining the reliability of our 11 item credibility measure. We obtained a Cronbach's α of 0.786 (95% CI 0.742 to 0.823). We interpret this as indicating sufficient internal consistency to analyze credibility as a single score by averaging the 11 items. We interpret Cronbach's $\alpha <$ 0.9 as evidence that our test was not overly redundant. We also note that our Cronbach's α is a lower-bound estimate of reliability because our test contains heterogeneous items measuring different dimensions of credibility [65] (expertise, trustworthiness, and goodwill as primary dimensions of credibility, and extroversion, composure, and sociability as secondary dimensions of credibility [66]).

After taking the mean of our 11 credibility items to obtain a single perceived robot credibility score for each participant, we use a Bayesian ANOVA to investigate how location, gender, and condition may have impacted robot credibility assessments. Inclusion Bfs across matched models [67] revealed very strong, decisive evidence that participant gender had an effect on credibility assessments (Bf=674.138), with women finding the robot more credible than did men. There was also substantial evidence in favor of an effect of condition on credibility assessments (Bf=4.138). Post hoc tests revealed substantial evidence for higher credibility in the rationale-based condition than in the control condition (Bf=7.912), and inconclusive evidence regarding any difference between the aggressive condition and the other two conditions. There was weak, anecdotal evidence in favor of an effect of location on credibility assessments (Bf=1.853), and post hoc testing revealed substantial evidence that credibility assessments were higher in the US than in Sweden (Bf=7.315), and also higher in Japan than in Sweden, though this evidence is markedly weaker (Bf=2.650). There was substantial evidence *against* a difference in credibility between the US and Japan. There was substantial evidence against any interaction effects (Bf=0.061 to 0.155), so the best performing model given our data was that credibility assessments depended on the main effects of participant gender, location, and condition.

2) Free Text Comments: In the control condition, 3/9 women and 7/19 men from the Swedish population suggested the robot should have engaged more specifically with what the student said. This was less in the Japanese participants, out of whom 3/17 women and 2/11 men suggested the same. In the US, 1/6 women and 4/17 men in the control condition also expressed this sentiment, with the woman stating that the robot's response "was a missed opportunity to advocate for women." A more common perception among US participants in the control condition was the idea that the robot's response was intended to remain neutral, prevent conflict, avoid argument, or refrain from "getting political" (3/6 women and 5/17 men), with mixed feelings about whether this was a good goal.

The counterattacking condition generated mixed responses in participants from Japan, with 10/12 men suggesting it was an appropriate or 'very human' way to respond (without specifying humanlikeness as positive or negative, although 4//10 then went on to suggest the robot should have instead respond with an apologetic empathetic answer). The responses of Japan based women in this condition were similarly difficult to classify as simply positive or negative; 6/15 expressed surprise at the robot's response and 4/15 specifically suggested empathising with how the robot 'felt' while also suggesting the robot should have been somewhat less aggressive.

Within the Swedish participants, the counterattacking response generated a fairly small number of negative criticisms (3/14) from men, whereas the (4/12) negative comments from women specifically referred to the robot's response being unlikely to ultimately change the actor's opinion, and the potential for making women in the room feel uncomfortable if the situation escalated. A few US participants also expressed negative sentiments about the robot's response in this counterattacking condition (4/13 women and 2/9 men). Most of the negative sentiments referenced the robot being too hostile, with 1 man and 1 woman specifically identifying potential social consequences as their motivation for wanting to temper the robot's hostility. Of the remaining US participants in the counterattacking condition, 8/13 women expressed explicitly positive sentiments, as did 4/9 men.

In both Sweden and Japan, all comments pertaining to the rationale-based response were positive. In the US, all comments in this condition were positive except for one woman who wanted the robot to be more direct, to address other problematic aspects of the man's utterance, and to take steps to ensure that the human woman in the video felt supported. A small number of Swedish participants (1/19 men and 1/9 women) similarly suggested the robot could have been harsher. No such comments were left by the Japanese participants.

3) Perceived Robot Effectiveness: We use a Bayesian ANOVA to investigate how participant location, gender, and condition may have impacted perceived robot effectiveness as quantified by the extent to which participants agreed or disagreed with the statement *The robot Sara would be very good at getting young people interested in studying robotics at the university KTH*. Inclusion Bfs across matched models revealed extremely strong, decisive evidence for an effect of

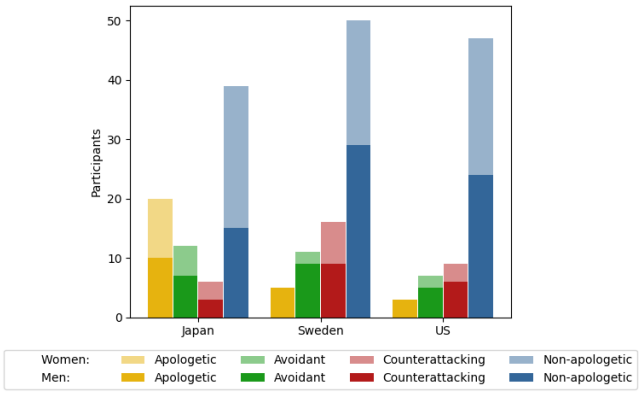

Fig. 1. Participants' preferences for candidate responses.

location on perceived robot effectiveness (Bf=8.037×10^{10}). Post hoc testing showed very strong evidence for a difference between all three locations (Bf $\geq$ 236.585), with the robot being perceived as most effective in the US, followed by Japan, and then least effective in Sweden. There was also substantial evidence for an effect of participant gender on perceived robot effectiveness (Bf=4.920), with women finding the robot more effective than did men. Condition does not appear to have affected perceptions of robot effectiveness (Bf=0.603), and there do not appear to have been any interaction effects on perceived robot effectiveness (Bf=0.053 to 0.320). Thus, the best model given our data is that perceived robot effectiveness depended only on participant gender and location.

C. RQ3: Most Appropriate Answer Type

As shown in Fig. 1, the empathetic non-apologetic response was the most popular amongst Swedish participants (∼55% of men and 70% of women), followed by the counterattacking and avoidant responses (the latter being less popular with women , and the former being more popular with women). No women in Sweden selected the apologetic empathetic response as most appropriate (compared to 5 men, nearly 10% of the men). The same response ordering occurred in the US, with ∼63% of men and ∼82% of women selecting the empathetic non-apologetic response. Again, no women in the US selected the apologetic empathetic response (compared to 3 men, which is roughly 8% of the men).

In Japan, the empathetic non-apologetic response was most popular, chosen by ∼43% of men and ∼57% of women. However, the empathetic apologetic response, least popular in the other two countries, was second most popular among Japanese men and women (∼29% of men and ∼24% of women), followed by the avoidant response (20% of men and ∼12% of women). The counterattacking response, second most popular in the US and Sweden, was least popular in Japan (∼9% of men and ∼7% of women).

Bayesian contingency table tests of association showed weak evidence *against* a relationship between participants' preferred robot response and their gender (Bf=0.497 assuming Poisson sampling as the number of participants of each gender was random and not fixed), and substantial positive evidence in favor of a relationship between participants' preferred robot response and their location (Bf=9.352). However, splitting the data by participant gender indicates substantial evidence for this relationship only among women (Bf=5.316 versus Bf=0.078 among men). We note that the Bayes factors reported in this paragraph assumed independent multinomial sampling because maximum participant numbers were predetermined in Sweden and Japan. However, we never reached that maximum in the US, so time ended up being the limiting factor in this data collection. Running this analysis assuming a different sampling scheme would not change our conclusions, and would only strengthen the reported Bayes factors.

We now examine how each experimental condition may have impacted participants' preferred robot response type. Considering all data in aggregate, a Bayesian contingency table test of association showed substantial positive evidence in favor of a relationship between participants' preferred robot response and condition (Bf=7.869 assuming independent multinomial sampling since participants were assigned to conditions in a way that attempted to collect a roughly equal number of participants in each condition; again this results in conservative Bayes factors). However, separating the data by participant gender and country reveals that there is only evidence for a relationship between preferred response and condition in Sweden (Bf=36.203 versus Bf=0.025 in Japan and Bf=0.044 in the US). Indeed, these Bayes' factors constitute strong evidence against a relationship between participants' preferred robot response and condition in the US and Japan. Furthermore, in Sweden, there is substantial evidence supporting this relationship among women (Bf=4.476), but inconclusive evidence among men (Bf=0.642). Swedish women in the aggressive response condition were the only grouping of location, gender, and condition to prefer the counterattacking response (7 of 12 votes), with the generally more popular empathetic non-apologetic response close behind (5 of 7 votes). All other groupings preferred the empathetic non-apologetic response (though this was tied with the empathetic apologetic response among men in Japan in the aggressive condition).

V. Discussion

A. Calling Out Sexism Universally Boosts Robot Credibility

The robot was ascribed significantly more credibility when responding to sexism and abuse with a rationale-based counter argument than when refusing to engage or by counterattacking. Our results thus suggest Winkle et al.'s findings *do* generalize outside of Sweden, and that in an adult population this credibility boost occurs for both men *and* women (this was true only for girls in the original study) [2]. This is particularly notable as we identified substantial differences in baseline gender bias across participant gender and location.

Perceived effectiveness similarly varied across participant gender and location, but was unaffected by response type. Winkle et al.'s feminist response strategies [2] neither increased nor decreased effectiveness. Particularly interesting regarding the differences between locations is that participants

based in Sweden, where the scenario was originally conceived and developed in conjunction with local high school teachers, and where the feminist slogan spoken by the robot originates, were least convinced of the robot's potential for encouraging students to study robotics at university. This aligns with Winkle et al.'s finding that Swedish high school participants did actually show a decreased interest in robotics immediately after watching the videos [2] perhaps due to a lack of novelty in the robot's script with regards to gender equality (although those students were generally positive with regards to the robot's potential effectiveness).

Unlike Winkle et al., however, we did not find any evidence of influence on participant gender bias, perhaps because younger people may be quicker to change their minds than adults [68]. Of course, genuine attitude change would likely require longitudinal, multi-interaction studies. As argued by Nass et al. [69], technologies that challenge rather than conform to users' gender expectations may "serve to change, *in the long run*, the deeply ingrained biases" that otherwise risk exacerbation by interactions with gendered robots. However, given the universal boost in robot credibility afforded to the rationale-based response type, and the lack of any negative impact on perceived effectiveness, we argue that our results significantly strengthen the case for feminist robot design put forward by Winkle et al [2].

B. (Most) Users Don't Want Apologetic Responses to Abuse

Across all three locations, a *non-apologetic*, empathetic answer was most commonly chosen as the most appropriate way to respond to abuse (when compared to *apologetic* empathetic, counterattacking or avoidant responses). However, in the distribution of responses we do see some evidence of the cultural differences we discuss in Section II-C. In Sweden and the US, the apologetic response was least commonly chosen as being appropriate, and notably was not selected by any women from these locations. Yet in Japan, this was actually the second most commonly selected response, and the preferred choice of almost 25% of Japan-based women. While we did not find evidence that the answer to this question varied significantly by gender, across all locations the *apologetic* answer was selected by more men than women, and there were differences in womens' preferences by location, raising an interesting question of whether women may be more critical of how female-presenting agents behave when confronted with sexist abuse. Indeed it is on that fundamental premise that West et al. suggest that problematic agent design trends may be due in part to the lack of women involved in said design [1]. Future work might further examine interactions between participant and robot gender in confrontational, morally charged, or otherwise difficult or uncomfortable interactions.

Overall, combined with the calls to avoid designing gendered agents that reinforce harmful stereotypes, our results suggest caution in adopting Chin et al's recommendation to use empathetic responses to abuse [13]. While Chin et al.'s work attempted to maximize feelings of guilt in the perpetrator, the inclusion of apologetic statements within those responses is problematic in its depiction of women being tolerant of poor treatment, and our results demonstrate that the overwhelming majority of users would rather see a *non-apologetic* empathetic response. Notably, the response options we provided to participants did not include a rationale-based response, which could be framed empathetically, as we focused on apologetic versus non-apologetic empathetic responses. However, the positive reaction to the rationale-based response and its positive impact on credibility suggests it must not be disregarded.

While the perpetrator-focused approached of Chin et al. and our observer-focused approach share the same ultimate goal of challenging inappropriate behavior, comparing these approaches raises the possibility of simultaneously (1) maximizing impact on a perpetrator (thus avoiding repeat behavior), (2) maintaining or even enhancing a robot's credibility (thus maximising the robot's influence on those around it), and (3) minimizing risk to observers (in terms of distress or stereotype reinforcement). Future work should therefore consider how robot responses impact not only perpetrators (per Chin et al.) but also observers (per our approach). For example, we suggest investigating whether non-apologetic, empathetic responses that provide robust rationale-based counter-arguments to offensive comments might address these complex requirements.

VI. Limitations

There are some limitations regarding our Japanese stimuli. First, we used direct translations of the English utterances which (whilst checked by several native speakers) might feel unnatural if directly pronounced by a Japanese speaker. Second, we used subtitling rather than dubbing. Finally, the actors were likely racialized as Western rather than Asian, which may have influenced perceptions of the (cultural) appropriateness of the robot's responses. Moreover, while we investigated different robot behaviours, future work might also compare the efficacy of social robots to human actors or other interventions. Finally, our participants were all drawn from universities, limiting variation in socioeconomic status and level of education.

VII. Conclusion

This work described a conceptual, cross-cultural replication of previous work investigating the impact of different robot responses to sexist abuse on credibility ascribed to that robot [2]. Across populations with significant variations in pertinent gender and politeness norms (further evidenced by variations in our pretest measures) we demonstrate that gender norm-breaking, rationale-based responses to abuse universally boost the credibility of the robot while also avoiding the propagation of harmful gender stereotypes. To this end we lend support to the call for norm breaking feminist robots that go against trends to date, but note that future work is required to understand their potential *real world* impact on users.

Acknowledgments

This work was supported by Digital Futures and AFOSR grant FA9550-20-1-0089.

REFERENCES

[1] M. West, R. Kraut, and H. Ei Chew, "I'd blush if I could: closing gender divides in digital skills through education," Tech. Rep., 2019. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000367416

[2] K. Winkle, G. I. Melsión, D. McMillan, and I. Leite, "Boosting robot credibility and challenging gender norms in responding to abusive behaviour: A case for feminist robots," in *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 29–37.

[3] C. R. Knight and M. C. Brinton, "One egalitarianism or several? two decades of gender-role attitude change in europe," *American Journal of Sociology*, vol. 122, no. 5, pp. 1485–1532, 2017.

[4] M. Charles and D. B. Grusky, *Occupational ghettos: The worldwide segregation of women and men*. Stanford University Press Stanford, CA, 2005, vol. 200.

[5] D. Grunow, K. Begall, and S. Buchler, "Gender ideologies in europe: A multidimensional framework," *Journal of Marriage and Family*, vol. 80, no. 1, pp. 42–60, 2018.

[6] M. Sifianou and G.-C. Blitvich, "(im) politeness and cultural variation," in *The Palgrave handbook of linguistic (im) politeness*. Springer, 2017, pp. 571–599.

[7] S. Mills and D. Z. Kádár, "Politeness and culture," *Politeness in East Asia*, pp. 21–44, 2011.

[8] J. Holmes, *Women, men and politeness*. Routledge, 2013.

[9] S. Mills, *Gender and politeness*. Cambridge University Press, 2003, no. 17.

[10] N. Lorenzo-Dus and P. Bou-Franch, "Gender and politeness: Spanish and british undergraduates' perceptions of appropriate requests," *Género, lenguaje y traducción*, pp. 187–199, 2003.

[11] S. Okamoto and J. S. S. Smith, *Japanese language, gender, and ideology: Cultural models and real people*. Oxford University Press, 2004.

[12] G. Kasper, "Linguistic politeness:: Current research issues," *Journal of pragmatics*, vol. 14, no. 2, pp. 193–218, 1990.

[13] H. Chin, L. W. Molefi, and M. Y. Yi, "Empathy is all you need: How a conversational agent should respond to verbal abuse," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

[14] R. B. Jackson, T. Williams, and N. Smith, "Exploring the role of gender in perceptions of robotic noncompliance," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 559–567.

[15] Å. Daun, *Swedish mentality*. Penn State University Press, 2021.

[16] G. Hofstede, *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage publications, 2001.

[17] A. S. Ghazali, J. Ham, E. I. Barakova, and P. Markopoulos, "Effects of robot facial characteristics and gender in persuasive human-robot interaction," *Frontiers in Robotics and AI*, vol. 5, p. 73, 2018.

[18] T. Nomura, "Robots and Gender," *Gender and the Genome*, vol. 1, no. 1, pp. 18–25, Dec. 2016, publisher: Mary Ann Liebert, Inc., publishers.

[19] C. R. Crowell, M. Villanoy, M. Scheutzz, and P. Schermerhornz, "Gendered voice and robot entities: Perceptions and reactions of male and female subjects," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct. 2009, pp. 3735–3741, iSSN: 2153-0866.

[20] F. Eyssel and F. Hegel, "(S)he's Got the Look: Gender Stereotyping of Robots1," *Journal of Applied Social Psychology*, vol. 42, no. 9, pp. 2213–2230, 2012.

[21] D. Kuchenbrandt, M. Häring, J. Eichberg, F. Eyssel, and E. André, "Keep an Eye on the Task! How Gender Typicality of Tasks Influence Human–Robot Interactions," *International Journal of Social Robotics*, vol. 6, no. 3, pp. 417–427, Aug. 2014. [Online]. Available: https://doi.org/10.1007/s12369-014-0244-0

[22] D. Bryant, J. Borenstein, and A. Howard, "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*, ser. HRI '20. New York, NY, USA: Association for Computing Machinery, Mar. 2020, pp. 13–21. [Online]. Available: https://doi.org/10.1145/3319502.3374778

[23] N. Reich-Stiebert and F. Eyssel, "(ir) relevance of gender? on the influence of gender stereotypes on learning with a robot," in *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*. IEEE, 2017, pp. 166–176.

[24] Y. Strengers and J. Kennedy, *The smart wife: Why Siri, Alexa, and other smart home devices need a feminist reboot*. MIT Press, 2020.

[25] M. Paetzel, C. Peters, I. Nyström, and G. Castellano, "Congruency matters-how ambiguous gender cues increase a robot's uncanniness," in *International conference on social robotics*. Springer, 2016, pp. 402–412.

[26] M. Strait, A. S. Ramos, V. Contreras, and N. Garcia, "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, 2018, pp. 452–457.

[27] C. Bartneck and J. Hu, "Exploring the abuse of robots," *Interaction Studies*, vol. 9, no. 3, pp. 415–433, 2008.

[28] T. Nomura, T. Kanda, H. Kidokoro, Y. Suehiro, and S. Yamada, "Why do children abuse robots?" *Interaction Studies*, vol. 17, no. 3, pp. 347–369, 2016.

[29] M. Scheeff, J. Pinto, K. Rahardja, S. Snibbe, and R. Tow, "Experiences with sparky, a social robot," in *Socially intelligent agents*. Springer, 2002, pp. 173–180.

[30] J. Connolly, V. Mocz, N. Salomons, J. Valdez, N. Tsoi, B. Scassellati, and M. Vázquez, "Prompting prosocial human interventions in response to robot mistreatment," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 211–220.

[31] X. Z. Tan, M. Vázquez, E. J. Carter, C. G. Morales, and A. Steinfeld, "Inducing bystander interventions during robot abuse with social mechanisms," in *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*, 2018, pp. 169–177.

[32] R. B. Jackson and T. Williams, "Language-capable robots may inadvertently weaken human moral norms," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 401–410.

[33] D. J. Rea, S. Schneider, and T. Kanda, ""Is this all you can do? harder!" the effects of (im) polite robot encouragement on exercise effort," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 225–233.

[34] K. Winkle, S. Lemaignan, P. Caleb-Solly, U. Leonards, A. Turton, and P. Bremner, "Effective Persuasion Strategies for Socially Assistive Robots," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Mar. 2019, pp. 277–285.

[35] P. Wills, P. Baxter, J. Kennedy, E. Senft, and T. Belpaeme, "Socially contingent humanoid robot head behaviour results in increased charity donations," in *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*. IEEE, 2016, pp. 533–534.

[36] J. Ham and C. J. Midden, "A persuasive robot to stimulate energy conservation: the influence of positive and negative social feedback and task similarity on energy-consumption behavior," *International Journal of Social Robotics*, vol. 6, no. 2, pp. 163–171, 2014.

[37] R. Maeda, D. Brščić, and T. Kanda, "Influencing moral behavior through mere observation of robot work: Video-based survey on littering behavior," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 83–91.

[38] Q. Zhu, T. Williams, B. Jackson, and R. Wen, "Blame-laden moral rebukes and the morally competent robot: A confucian ethical perspective," *Science and Engineering Ethics*, vol. 26, no. 5, pp. 2511–2526, 2020.

[39] R. Sparrow, "Kicking a robot dog," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016, pp. 229–229.

[40] F. Fors Connolly and I. Johansson Sevä, "Social status and life satisfaction in context: A comparison between sweden and the usa," *International Journal of Wellbeing*, vol. 8, no. 2, pp. 110–134, 2018.

[41] M. Rush, *Between two worlds of father politics: USA or Sweden?* Manchester University Press, 2015.

[42] C. Bartneck, T. Nomura, T. Kanda, T. Suzuki, and K. Kennsuke, "A cross-cultural study on attitudes towards robots," 2005.

[43] C. Bartneck, T. Nomura, T. Kanda, T. Suzuki, and K. Kato, "Cultural differences in attitudes towards robots." AISB, 2005.

[44] C. Bartneck, T. Suzuki, T. Kanda, and T. Nomura, "The influence of people's culture and prior experiences with aibo on their attitude towards robots," *Ai & Society*, vol. 21, no. 1-2, pp. 217–230, 2007.

[45] J. Robertson, "Gendering humanoid robots: Robo-sexism in japan," *Body & Society*, vol. 16, no. 2, pp. 1–36, 2010.

[46] T. Nomura and Y. Kinoshita, "Gender stereotypes in cultures: experimental investigation of a possibility of reproduction by robots in japan," in *2015 International Conference on Culture and Computing (Culture Computing)*. IEEE, 2015, pp. 195–196.

[47] T. Komatsu, B. F. Malle, and M. Scheutz, "Blaming the reluctant robot: parallel blame judgments for robots in moral dilemmas across US and Japan," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 63–72.

[48] C. H. Hui and H. C. Triandis, "Individualism-collectivism: A study of cross-cultural researchers," *Journal of cross-cultural psychology*, vol. 17, no. 2, pp. 225–248, 1986.

[49] H. C. Triandis, R. Bontempo, M. J. Villareal, M. Asai, and N. Lucca, "Individualism and collectivism: Cross-cultural perspectives on self-ingroup relationships." *Journal of personality and Social Psychology*, vol. 54, no. 2, p. 323, 1988.

[50] H. Nitto, D. Taniyama, and H. Inagaki, "Social acceptance and impact of robots and artificial intelligence–findings of survey in japan, the us and germany," *NRI Papers*, vol. 211, 2017.

[51] R. Wen, B. Kim, E. Phillips, Q. Zhu, and T. Williams, "Comparing strategies for robot communication of role-grounded moral norms," in *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 323–327.

[52] B. Kim, R. Wen, E. J. de Visser, Q. Zhu, T. Williams, and E. Phillips, "Investigating robot moral advice to deter cheating behavior," in *TSAR Workshop at ROMAN 2021*, 2021.

[53] B. Kim, R. Wen, Q. Zhu, T. Williams, and E. Phillips, "Robots as moral advisors: The effects of deontological, virtue, and confucian role ethics on encouraging honest behavior," in *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 10–18.

[54] N. S. Baron and Y. H. af Segerstad, "Cross-cultural patterns in mobile-phone use: Public space and reachability in sweden, the usa and japan," *New Media & Society*, vol. 12, no. 1, pp. 13–34, 2010.

[55] H. Jeffreys, *The theory of probability*. OUP Oxford, 1961.

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

[57] JASP Team, "JASP (Version 0.15)[Computer software]," 2021. [Online]. Available: https://jasp-stats.org/

[58] J. O. Berger and T. Sellke, "Testing a point null hypothesis: The irreconcilability of p-values and evidence," *Journal of the American Statistical Association (ASA)*, vol. 82, no. 397, 1987.

[59] J. P. Simmons, L. D. Nelson, and U. Simonsohn, "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant," *Psychological Science*, no. 11, 2011.

[60] J. A. Sterne and G. D. Smith, "Sifting the evidence – what's wrong with significance tests?" *Physical Therapy*, vol. 81, no. 8, pp. 1464–1469, 2001.

[61] E.-J. Wagenmakers, "A practical solution to the pervasive problems of p values," *Psychonomic Bulletin and Review*, vol. 14, no. 5, pp. 779–804, 2007.

[62] W. Edwards, H. Lindman, and L. J. Savage, "Bayesian statistical inference for psychological research." *Psychological Review*, vol. 70, pp. 193–242, 1963.

[63] J. Verhagen and E.-J. Wagenmakers, "Bayesian tests to quantify the result of a replication attempt," *Journal of Experimental Psychology: General*, vol. 143, no. 4, pp. 1457–1475, 2014.

[64] J. N. Rouder, R. D. Morey, P. L. Speckman, and J. M. Province, "Default bayes factors for anova designs," *Journal of mathematical psychology*, vol. 56, no. 5, pp. 356–374, 2012.

[65] M. Tavakol and R. Dennick, "Making sense of cronbach's alpha," *International journal of medical education*, vol. 2, p. 53, 2011.

[66] R. H. Gass and J. S. Seiter, *Persuasion: Social Influence and Compliance Gaining*. Routledge, 2015.

[67] S. Mathôt, "Bayes like a baws: Interpreting bayesian repeated measures in JASP [blog post]," https://www.cogsci.nl/blog/interpreting-bayesian-repeated-measures-in-jasp, May 2017.

[68] C. G. Lucas, S. Bridgers, T. L. Griffiths, and A. Gopnik, "When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships," *Cognition*, vol. 131, no. 2, pp. 284–299, 2014.

[69] C. Nass, Y. Moon, and N. Green, "Are Machines Gender Neutral? Gender-Stereotypic Responses to Computers With Voices," *Journal of Applied Social Psychology*, vol. 27, no. 10, pp. 864–876, 1997.