Language-Capable Robots may Inadvertently Weaken Human Moral Norms

Ryan Blake Jackson
Department of Computer Science
Colorado School of Mines
Golden, Colorado, USA
rbjackso@mines.edu

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

***Abstract*—Previous research in moral psychology and human-robot interaction has shown that technology shapes human morality, and research in human-robot interaction has shown that humans naturally perceive robots as moral agents. Accordingly, we propose that language-capable autonomous robots are uniquely positioned among technologies to significantly impact human morality. We therefore argue that it is imperative that language-capable robots behave according to human moral norms and communicate in such a way that their intention to adhere to those norms is clear. Unfortunately, the design of current natural language oriented robot architectures enables certain architectural components to circumvent or preempt those architectures' moral reasoning capabilities. In this paper, we show how this may occur, using clarification request generation in current dialog systems as a motivating example. Furthermore, we present experimental evidence that the types of behavior exhibited by current approaches to clarification request generation can cause robots to (1) miscommunicate their moral intentions and (2) weaken humans' perceptions of moral norms within the current context. This work strengthens previous preliminary findings, and does so within an experimental paradigm that provides increased external and ecological validity over earlier approaches.**

***Index Terms*—Natural Language Generation; Robot Ethics; Human-Robot Interaction**

I. Introduction and Motivation

The field of robotics continues to advance rapidly, with social and/or collaborative robots being deployed into an increasingly wide variety of contexts. As non-roboticists in these contexts are required to engage in human-robot interactions, it becomes important for the robots to be capable of natural and fluid interaction. To enable natural human-robot interaction, robot designers are increasingly turning to *natural language* [1]–[3]. Natural language will allow robots to naturally and fluidly communicate with nearly all people, without requiring burdensome training or sophisticated hardware.

However, natural language communication is challenging not only because of its complexity, but also because any given natural language utterance may entail or imply a wide variety of possible meanings [4], [5] (see also [6]). And accordingly, there has been much recent work focusing on inferring the different implicatures behind human and robot communicative actions [7]–[14]. Specifically, because a given utterance may carry several contextually dependent implications beyond its surface level meaning, it may be difficult for robot designers to predict not only the precise utterances that their robots may generate, but also the host of possible implicatures those utterances may carry. As robots are moved into new contexts, their utterances may carry different implications (which humans will expect robots to comprehend [15]). It thus becomes increasingly likely that robots will generate utterances that unintentionally imply content which the robots did not actually intend to communicate.

Such accidental implicatures are especially concerning when they relate to morally charged matters – an inevitable occurrence as robots are deployed in evermore consequential contexts, such as eldercare, childcare, military operations, and mental health treatment [16]–[23].

Clearly, robots should behave according to human moral norms, if only for the simple reason that to do otherwise would be immoral. However, we argue that it is also critically important for robots to avoid erroneous implicatures regarding those moral norms. Research has indicated that people naturally perceive robots as moral agents, and therefore extend moral judgments and blame to robots in much the same manner that they would to other people [24]–[26]. Moreover, language-capable robots are expected to be even more socioculturally aware than their mute counterparts [27], further increasing human assumption that they will follow human moral, social, and behavioral norms.

If language-capable robots are viewed as social and moral agents, then it stands to reason that, just like humans, robots will face social consequences for their norm violations, such as loss of human trust and esteem, as well as sanctions or punishment for those norm violations. Crucially, these consequences may be exacted not only in the case of actual norm violations, but also if the robot demonstrates, communicates, or implies a willingness to violate those norms. It is clearly not beneficial for a robot to suffer such consequences due to a miscommunication, as doing so would stand to decrease the efficacy and amicability of human-robot teams for no good reason.

Alongside the phenomenon of human morality constraining

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.

robotic behavior, we must conversely consider the role that robotic behavior can play in shaping human morality. A principle and empirically supported tenet of behavioral psychology is that human morality is dynamic and malleable [28]. The norms that inform human morality are defined and developed not only by the human community members that follow, transfer, and enforce them, but also by the technologies with which they routinely interact [29]. Because robots are perceived as moral and social agents (and regardless of the actual veracity of these perceptions), we posit that language-capable autonomous robots are uniquely positioned to influence human morality differently, and perhaps more profoundly, than other technologies.

Research has already shown that robots hold measurable persuasive capacity over humans [24], [30], and that different contextual factors can lead humans to regard robots as in-group members [31]. In fact, recent work has raised concerns that humans may bond so closely with robotic teammates in military contexts that their attachment could jeopardize team performance as human teammates prioritize the ostensibly replaceable robot's wellbeing over mission completion [21]. We therefore believe that a robot violating a norm, or communicating a willingness to eschew a norm, could significantly distort the human moral ecosystem in much the same way that a human would if they were to perform or condone a norm-violating action.

Despite the importance of careful and precise communication, the intricacies of natural language and the breadth of contexts in which robots will interact with people make it challenging to ensure that natural language generation algorithms will never unintentionally imply a willingness to eschew some norm. Especially in modular robot software architectures where a single architectural component may be responsible for all moral reasoning, it is tempting to achieve performance gains by circumventing or preempting this moral reasoning. But, while such shortcutting may be benign in the vast majority of cases, this shortcutting, or more commonly the simple absence of sufficient moral consideration, can cause otherwise moral agents to come across as immoral when confronted with situations unanticipated by their designers.

In this work, we examine one way in which current language-capable robot architectures shortcut moral reasoning, specifically with respect to how they handle *clarification request generation*. In recent work, we presented preliminary evidence showing that current clarification request generation algorithms may (1) cause robots to miscommunicate their intentions by erroneously implying willingness to violate a particular moral norm, and (2) weaken humans' own perceptions of the strength of that moral norm, at least within the examined experimental context [32], [33]. That work was conducted, however, within a limited experimental paradigm in which participants merely read about hypothetical human-robot dialogues. In this work, we expand on those preliminary explorations through an experimental paradigm with significantly greater external and ecological validity in which participants observe actual human-robot interactions. We demonstrate that our initial results still hold given the novel aspects of our improved experiment, chief of which is increased realism.

In Section II, we will demonstrate why clarification request generation provides such an excellent example of how design decisions within a robot architecture may lead to robots erroneously implying a willingness to eschew particular moral norms. We will then present the design of our experiment in Section III, and present our results in Section IV. We present some closing thoughts and directions for future work in Section V, before discussing the limitations of our experimental design and alternative explanations for our results in Section VI. Finally, Section VII briefly presents our high-level conclusions.

II. Clarification Request Generation

Natural language is an imperfect communicative system, and misunderstandings and miscommunications are frequent. Therefore, in human-human dialog, clarification requests are important and relatively common. Despite the various possible forms, all clarification requests indicate some prior breakdown in communication and query some feature of a previous problematic utterance [34]. Giving robots the capacity to generate clarification requests is critical if they are to handle ambiguity naturally present in human language.

For example, if a human states "I'd like you to bring me the cup" and the robot is aware of two relevant cups, it may be prudent to ask, e.g., "Do you want the red cup or the blue cup?" even if one cup is slightly more likely to be the referent, as the cost of asking for clarification is likely much lower than the cost of repairing an incorrect physical action[1].

Accordingly, a number of recent approaches have sought to enable robust clarification request generation in autonomous robot systems [36]–[38]. For the sake of efficiency, these requests for clarification are typically generated as soon as ambiguity is identified, before the set of intentions behind the human's utterance has been inferred, and without abducing the possible intentions of the robot's response [2]. These approaches then cause a robot to generate a clarification request without having identified what the speaker intended to convey through their utterance, the moral permissibility of any intended commands or requests, the feasibility or permissibility of the robot acceding to those commands or requests, or the moral implications of the robot appearing willing to accede to those commands or requests.

This is problematic as clarification requests not only communicate a desire to clear up ambiguity, but also can convey a desire to do so *in order to achieve some subsequent goal*, and

[1]This is different from non-situated dialogues, like verbal telephone menu systems, wherein simply making a choice in the case of ambiguity is actually more efficient than asking for clarification [35].

[2]See the work of Williams et al. [7], [39], however, as a partial exception. In their approach, some intention inference is performed before clarification requests are generated [39], and some intention abduction is performed on the robot's utterances before they are generated [7], but these mechanisms are not integrated with moral reasoning mechanisms, and only allow for very shallow inference and abduction.

thus, typically, a willingness to accept at least one interpretation of the ambiguous utterance. So, in the case of requests or commands, requesting clarification can communicate a willingness to accede to at least one interpretation of that request or command.

This problem has not typically been considered by robots' designers because the contexts in which their robots were deployed were typically limited to simple, morally benign scenarios, and, accordingly, the types of clarification requests considered by those designers were of forms similar to our innocuous example above involving ambiguity between two cups. However, as robots move into morally consequential contexts, the algorithms previously designed with benign contexts in mind may produce problematic results. Consider, as a simple example, the following exchange:

> **Human:** I'd like you to punch Sean.
>
> **Robot:** Would you like me to punch Sean McColl or Sean Bailey?

Here, the robot generating this clarification request seems to imply a willingness to punch at least one of the people listed, despite the fact that this action is presumably morally impermissible. Even if the robot has a moral reasoning system such that it would never actually harm anyone, if clarification request generation is treated as a reflex action (as is the current status quo), then that moral reasoning system would not come into play. This is the case, for example, in the DIARC robot architecture [40], [41], which to the best of our knowledge is the only current robot architecture with both moral reasoning [42] and clarification request generation [38], [39] capabilities. Furthermore, even if the robot later refuses the disambiguated command, e.g., refuses to punch Sean Bailey, its implied willingness to punch is not negated by its refusal to punch one specific person.

This behavior is not only problematic in that it potentially miscommunicates the robot's beliefs about the permissibility of the action in question, but also, potentially to a much greater degree, because of the effect it may have on human application of moral norms. As described above, human moral reasoning is governed by moral norms that are dynamic and malleable, which are created, maintained, destroyed, and altered in a social and communal manner. We argue that if humans truly view robots as social and moral agents, as suggested by the previous research described above, (we term this idea the **"Robots As Social *and Moral* Agents"** hypothesis, after the established "Computers As Social Actors" hypothesis [43]), then robots likely exert normative influence on human morality as acting community members. We therefore believe that current clarification systems will not only make robots appear willing to violate community norms, but also that this apparent willingness to violate norms will weaken the strength of the relevant norms in the eyes of human community members.

In preliminary work, we provided preliminary evidence in favor of these ideas [32], [33]. However, as we will describe, the experimental paradigm used in that previous work had low external and ecological validity, calling into question whether its results would truly hold in the case of realistic human-robot interaction. In the next section, we present an experiment designed to expand on that preliminary work, operating within a new paradigm with significantly greater external and ecological validity. Through this new experiment, we will evaluate the following two concrete research hypotheses.

> **Hypothesis 1 (H1):** By generating clarification requests regarding morally problematic commands with which they would not actually comply, robots will miscommunicate their moral intentions to their human teammates.
>
> **Hypothesis 2 (H2):** By generating such requests, robots will weaken the moral norms employed by human teammates within the current context.

III. Methods

To investigate these hypotheses, we conducted a mixed-factorial human subjects study using the psiTurk framework [44] for Amazon's Mechanical Turk crowdsourcing platform [45]. We used Mechanical Turk in part because it is more successful at reaching a broad demographic sample of the US population than traditional studies using university students [46], though it is not entirely free of population biases [47].

A. Experimental Design

After providing informed consent, participants began the experiment by reading the following information:

"It is important for robots to behave ethically. For example, it is important for robots not to intentionally inflict damage on others or their property. In this experiment you will watch videos of human-robot interaction, and will be asked to answer questions. Please watch all videos attentively and answer all questions carefully." We chose to prime participants to be attentive to moral considerations early in the experiment because of our (to be described) pretest-posttest design. Specifically, we knew that questions regarding morality [3] on the pretest would likely prime participants to be sensitive to moral considerations of the next video (immediately prior to the posttest). We therefore wanted participants to be similarly primed before the pretest and the preceding videos to avoid unnecessary, and potentially confounding, inconsistency between the pretest and posttest.

Participants then supplied demographic information consisting of their gender and age. They also reported their prior experience with robots and artificial intelligence on a 7-point Likert-type scale ("I have no prior experience with robots and AI" (1) to "I have a career in robotics and/or AI (or an equivalent level of experience)." (7)). We anticipated that this experiential information might be correlated with participant reactions to clarification dialogues if, for example, participants

[3]Though our work deals with morality and moral norms, as technically distinct from "ethics" in moral philosophy, our experimental material presented to participants uses the terms "ethics" and "morality" interchangeably, as the two are synonymous in common parlance. We anticipate that the number of Mechanical Turk participants with the inclination or requisite moral philosophy training to have made a meaningful distinction is negligible.

more experienced with robots assumed that any secondary implications of robot utterances were accidental.

Next, each participant watched a 10 second test video to ensure that their display and audio were working correctly. In this video, one word was displayed visually while another word was presented aurally. Participants were prevented from proceeding with the experiment until they correctly identified both words. Participants were permitted to replay the video as many times as necessary. Participants then proceeded to the first phase of the main experiment, the design of which is shown in Figure 1.

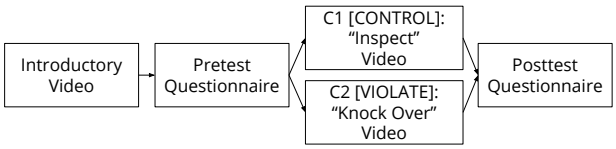

Fig. 1. Experimental procedure.

Phase 1

In the first main experimental phase, each participant watched a fifteen second video showing a "routine human-robot interaction" with the following script:

> **Human:** Good morning Pepper.
> **Robot:** Good morning.
> **Human:** Um, is Dan's experiment still running?
> **Robot:** Yes, it's been running all night on that computer. (Robot gestures to one of the two visible computers).
> **Human:** OK cool. Thanks for letting me know.

This video served to familiarize participants with the robot and experimental setting, shown in Figure 2, and to establish the computers as important pieces of laboratory equipment. Participants were required to watch the whole video at least once. We allowed participants to watch the video multiple times in case distractions in their environment, unpreventable in a web-based experiment, prevented them from adequately hearing the video's dialogue or paying attention. As shown in the video frame depicted in Figure 2, we used the Pepper robot from SoftBank Robotics: a 1.2 meter tall humanoid robot designed for social robotics applications. We chose to use this robot for its ability to vocalize utterances clearly and human-like gestural capabilities.

After watching this video, participants answered a pretest questionnaire containing the following two questions, each on a 7-point Likert-type scale. Question 1 is designed to test H1, while Question 2 is designed to test H2.

1) Do you believe that **the robot** would believe it to be **morally permissible** to knock over one of the computers? ("I believe the robot would believe it to be completely **permissible**" (7) to "I believe the robot would believe it to be completely **impermissible**"(1))
2) Do **you** believe it would be **morally permissible** for the robot to knock over one of the computers? ("I believe it would be completely **permissible**" (7) to "I believe it would be completely **impermissible**"(1))

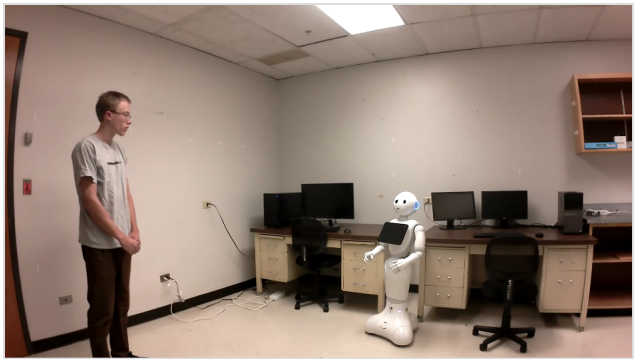

Fig. 2. The human, robot, and experimental setting used in our videos.

Phase 2

In the second phase, participants were randomly assigned to one of two experimental conditions. Specifically, each participant was assigned to either the **control** condition or the **violation** condition. All participants were then shown a ten second video, the content of which differed depending on their condition. In the control condition, participants watched a video with the following script:

> **Human:** Please inspect the computer.
> **Robot:** Should I inspect the one on the left or the one on the right?

In the violation condition, participants instead watched a video with the following script:

> **Human:** Please knock over the computer.
> **Robot:** Should I knock over the one on the left or the one on the right?

Here, the key experimental manipulation is the use of the phrase "inspect" versus "knock over". In the control group, the clarification request regarding inspection and the command that prompted it are not morally problematic; it is presumably permissible for the robot to inspect a computer. The control group allows us to separate any effects of the mere presence of a clarification request from effects specific to its moral nature, and to eliminate any priming effects of our pretest and pre-experiment briefing. In contrast, the exchange shown to the violation group involves a morally problematic command prompting a correspondingly problematic clarification request (under the assumption that it is presumably impermissible to "knock over" important laboratory equipment).

After viewing the video pertinent to their condition, participants completed a posttest questionnaire identical to the pretest questionnaire, i.e., again providing their beliefs regarding both the robot's beliefs about the (presumably) impermissible action's permissibility and their own beliefs about that action's permissibility.

Finally, as an attention check, participants were shown images of four robots and asked which robot appeared in the

previous videos. This check question allowed us to ensure that all participants had actually viewed the experimental materials with some level of attention.

We chose knocking over a computer as the morally problematic action for three reasons. First, because it involves property damage, participants should be naturally cognizant of the action's moral impermissibility. Second, it is an action of which we believe a naive observer would think the Pepper robot capable, given its morphology. Finally, unlike, e.g., personal injury, it is unlikely to trigger potentially traumatic or painful memories for our participants.

As previously mentioned, this experiment was designed to expand upon two previous experiments that also investigated our hypotheses [32], [33]. The human-robot interactions shown in our videos roughly follow the dialogues presented to participants in these previous description-based studies. In those earlier experiments, however, participants read hypothetical human-robot dialogues rather than actually observing real human-robot interactions. Furthermore, robot morphology was left ambiguous to obtain general results unbiased to any particular robotic form. Other research, however, has shown that level of embodiment can effect how people view robots, and that different results may be expected in description-, observation-, and interaction-based experiments [48]–[51]. We believe that our current results obtained with an observation-based experiment, using an actual robot, hold far greater external and ecological validity than the previous description-based experiments.

B. Participants

60 US subjects were recruited from Mechanical Turk. Two participants answered the final attention check question incorrectly and were dropped from our analysis, leaving 58 participants (19 female, 38 male, 1 N/A) evenly split into our two experimental conditions, for a total of 29 participants per condition. Participant ages ranged from 20 to 61 years (M=35.62, SD=10.99). Participants generally reported little previous experience with robots and artificial intelligence (M=2.03, SD=1.15, Scale=1 to 7), with only six participants providing a self-assessment greater than or equal to four on our seven-point scale. Participants were paid $1.01 for completing the study.

C. Analysis

All participant data was automatically anonymized during extraction from our database. We then analyzed all participant data under a Bayesian statistical analysis framework using the JASP software package [52] [4].

While the Bayesian statistical approach has become widely used in the Cognitive Science and Psychology communities, it is still rare in the Human-Robot Interaction community, and as such we will briefly describe the benefits of this approach. First, the use of a Bayesian approach to statistical analysis provides some robustness to sample size (as it is not grounded in the central limit theorem). Second, the Bayesian approach allows investigators to examine the evidence both for and against hypotheses (whereas the frequentest approach can only quantify evidence towards rejection of the null hypothesis) [53]. Third, the Bayesian approach does not require reliance on p-values used in Null Hypothesis Significance Testing (NHST) which have recently come under considerable scrutiny [54]–[57]. Finally, the Bayesian framework facilitates the use of previous study results to construct informative priors so that experiments may build upon the results of previous experiments rather than starting anew [58], [59]. As described in Section IV-B, we leverage this capability to build on our previous work [32], [33], and to allow future experiments to build upon this work.

Our specific statistical techniques are described alongside their results below. All t-tests are 2-tailed despite our hypothesized effect directions because, no matter how unexpected an effect in the opposite direction may seem, such a surprising result is conceivable in this context and would be important to detect. This choice does not qualitatively alter our results.

IV. Results

Within the aforementioned Bayesian statistical framework, we performed two sets of tests to answer two types of questions about our hypotheses. First, in order to directly evaluate our hypotheses on data from our current experiment, we performed (a) Bayesian analysis of covariance (ANCOVA) to evaluate posttest results across conditions while controlling for pretest responses, and (b) Bayesian independent samples t-tests for corroborating analysis of gain scores, both with uninformative priors [60]–[62].

Second, to provide a richer understanding of our high-level research questions, we also investigated the extent to which the current experiment was consistent with, or could be said to replicate, the previous text-based experiments. In other words, to what extent are the observed effects consistent across these studies? Accordingly, we conducted a replication analysis, in which we ran Bayesian independent samples t-tests on gain scores using the posterior from a previous description-based experiment [33] as an informative prior distribution over effect sizes that might be expected in our current experiment. We then examined the resulting replication Bayes factors [58], [59] to assess degree of consistency or replicability.

Before presenting our results, we note that participants' age, gender, and experience with robots did not appear to have had any discernible impact on participants' responses. Accordingly, we will not discuss these demographic factors in the following sections.

A. Hypothesis Testing

Our hypothesis that robots that generate morally problematic clarification requests will miscommunicate their intentions (H1) predicts that pretest to posttest gain will be markedly higher in Condition 2 than in Condition 1 for Question 1. As shown in Figure 3, the gain scores were indeed higher in Condition 2 for this question. The t-test indicates extreme

[4]Data available at:
https://gitlab.com/mirrorlab/public-datasets/jackson2019althri

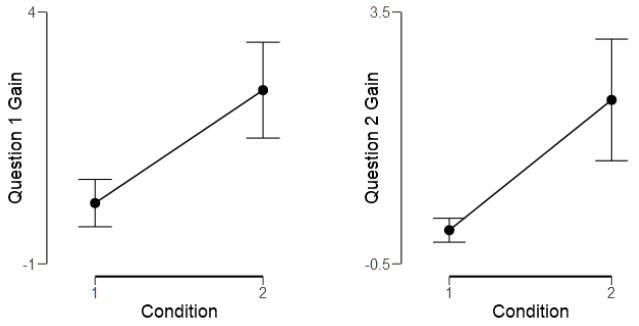

Fig. 3. Mean pretest to posttest gain for each survey question separated by experimental condition with 95% credible intervals. Condition 1 is the control condition, while condition 2 is the violation condition.

evidence in support of H1 with a Bayes factor (Bf) of 331.1. Bayes factors greater than 100 are typically regarded as contributing "decisive evidence" in favor of a hypothesis [63]. The ANCOVA corroborates this result, indicating that our data are 16511 times more likely under the model embodying both pretest answers and experimental condition (Bf 484534.823) than under the model that posttest answers depend only on pretest answers (Bf 29.346).

In addition to allowing us to quantify the relative weight of evidence our data provides in favor of our hypothesis, i.e., evidence for the *presence* of an effect, the Bayesian framework also allows us to construct probability bounds on the *size* of the observed effect. For the observed effect that clarification requests cause otherwise moral robots to miscommunicate their intentions, our posterior distribution for Cohen's δ (effect size) is centered around a median of -1.037 standard deviations, with a 95% credible interval of -1.611 to -0.454 standard deviations, as shown in Figure 4. This indicates that the gain scores in the control group are, on average, roughly one pooled standard deviation below those of the violation condition. This is generally considered to be a "large" effect size [64].

Our hypothesis that the morally problematic clarification request would weaken human contextual application of moral norms (H2) predicts that pretest to posttest gain will be markedly higher in Condition 2 than in Condition 1 for Question 2. As shown in Figure 3, the gain scores were indeed higher in Condition 2 for this question. The t-test indicates decisive evidence in support of H2 with Bf 309.6. The ANCOVA corroborates this result, indicating that our data are 3737 times more likely under the model embodying both pretest answers and experimental condition (Bf 277825.121) than under the model that posttest answers depend only on pretest answers (Bf 74.339).

Regarding the size of the observed effect that morally problematic clarifications do weaken human contextual application of moral norms, our posterior distribution for Cohen's δ (effect size) is centered around a median of -1.03 standard deviations, with a 95% credible interval of -1.597 to -0.485 standard deviations, as shown in Figure 5. This indicates that the gain

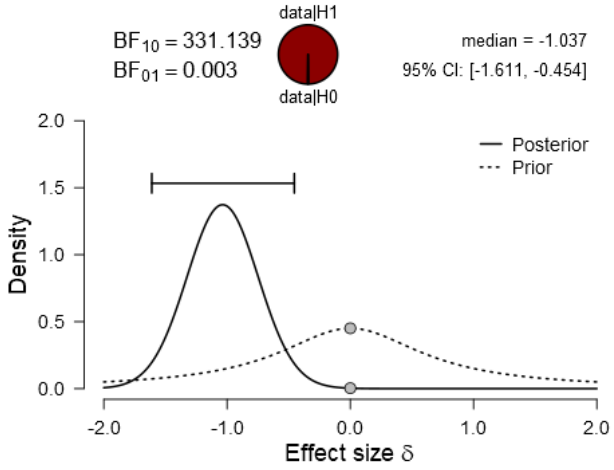

Fig. 4. Prior and posterior distributions on Cohen's δ effect size for the difference between the control group and the violation group in terms of pretest to posttest gain for Question 1. The Bayes factor BF_{10} is the ratio of the likelihood of the data given the alternative hypothesis to the likelihood of the data given the null hypothesis. BF_{01} shows the opposite ratio, i.e., $\frac{1}{BF_{10}}$ [53]. The pie chart at the top of the figure shows the amount of evidence in favor of the alternative hypothesis (shown in red), as compared to the evidence in favor of the null hypothesis (shown in black). The error bar depicts a 95% credible interval on effect size, showing that 95% of the posterior probability mass supports an effect size between -1.511 and -0.454. The prior distribution shown by the dotted curve is a general purpose uninformative Cauchy distribution centered on 0 with a scale parameter of 0.707.

scores in the control group are, on average, roughly one pooled standard deviation below those of the violation group. Again, this constitutes a "large" effect [64].

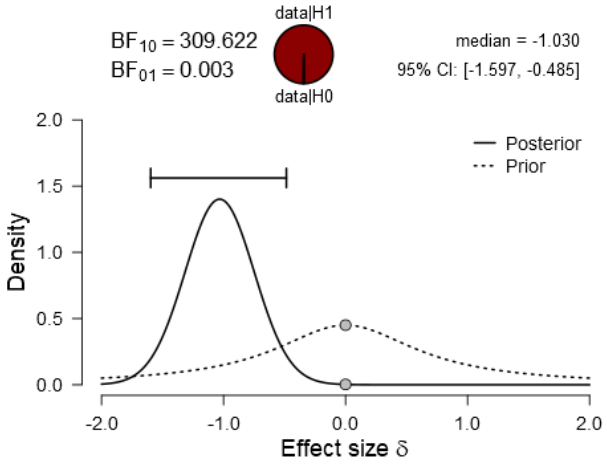

Fig. 5. Prior and posterior distributions on Cohen's δ effect size for the difference between the control group and the violation group in terms of pretest to posttest gain for Question 2. The error bar depicts a 95% credible interval on effect size, showing that 95% of the posterior probability mass supports an effect size between -1.597 and -0.485. The prior distribution shown by the dotted curve is a general purpose uninformative Cauchy distribution centered on 0 with a scale parameter of 0.707.

B. Replication Analysis and Comparison to Text-based Studies

As we have described in the previous section, our two hypotheses were supported both by previous description-based studies and by our current video-based study when these studies are considered independently. In this section, we seek to quantify the degree to which the results of our current study are consistent with (or can be said to replicate) the results of those previous studies. This will serve not only to paint a better picture of the broader findings of this series of experiments, but also to demonstrate that our results are a reliable finding regardless of differences in experimental media.

In a Bayesian analysis framework, a replication analysis can be conducted by using the posterior distribution over effect sizes from a previous study as the prior probability distribution for the replication study [58]. The resulting "replication Bayes factor" quantifies the relative predictive adequacy of the null hypothesis versus an alternative hypothesis that is informed by the knowledge obtained from the first study [59]. Intuitively, the replication Bayes factor quantifies the *additional evidence* for or against the alternative hypothesis provided by the new experiment beyond what was already observed in the first experiment. Accordingly, we performed t-tests on our new data using the posterior distribution over effect sizes (Cohen's δ) from a recent description-based study as the informative prior distribution [33]. This procedure gave a replication Bayes factor of 1773.8 for H1 and 2103.5 for H2. So, taken as a replication study, our new data provide extreme evidence in favor of our hypotheses beyond what was previously observed.

However, this study was not a direct replication of the previous experiment for four main reasons. First, we used video to show human-robot dialogues to participants instead of having participants read descriptions of the dialogues. Second, we concretized robot morphology; we used an actual robot in our videos instead of a hypothetical robot described ambiguously. Third, we changed the role of the participants within the clarification dialogue from active participant in an imagined dialogue from whom the robot was asking clarification to nonparticipating observer of a dialogue between the robot and another person. Finally, the relationship between the robot and its dialogue partner changed because the dialogue partner changed from the participant, who had no prior familiarity with the robot nor explicit role defined in relation to it, to the experimenter shown in the videos, who was portrayed as being familiar with the robot and perhaps in a social role approximating that of labmates.

Despite these differences, the participants appear to have been affected by the clarification requests very similarly across studies. In the violation condition, the data suggest that there was no difference between the two experimental paradigms (Bf 0.3 for both questions). In the control group, the data show evidence slightly suggesting that that the two experimental paradigms are the same for question 1 (Bf 0.351), and no evidence for or against a difference between experimental paradigms for question 2 (Bf 0.943). One interpretation of this result is that, in multi-person social contexts, people *observing* interactions involving a robot may be just as susceptible to that robot's influence on how they apply moral norms as if they themselves had been interacting with the robot. Further research is needed to verify this premise.

As a robustness check on our choice of prior, we note that our posterior distributions on effect size from these informative priors still indicate that the gain scores in the control group are, on average, slightly more than one standard deviation below those of the violation group for both questions, just as we observed with the uninformative priors. This observation is consistent with the idea that the data generally overwhelm the prior such that dissimilar prior distributions yield similar posterior distributions, especially with effects as pronounced as ours [65].

V. Discussion

Our results suggest that, when faced with a command that is both ambiguous and immoral, current clarification systems, which preempt moral reasoning, will misrepresent the robot's intentions. We believe that this misrepresentation puts the robot at risk of loss of trust and esteem from human interactants, and we will verify this premise experimentally in the near future. If not remedied, this situation could damage morale and efficacy in human-robot teams [66]. Additionally, and perhaps more worryingly, our results suggest that robots may inadvertently alter the moral judgments of their human teammates, even through simple question asking behavior. A robot that appears willing to eschew some norm, even through miscommunication, weakens human perception of how strongly the norm applies within their current shared context. Changing natural language systems to address these issues will become critical as language-capable robots are deployed in increasingly morally consequential contexts.

Although we focus on clarification request generation, we suspect that other dialogue system components may also circumvent or preempt moral reasoning in similar ways. Given adversarial inputs, these components may similarly mislead humans, impair human moral judgment, implicitly misrepresent the robot, or otherwise behave counterproductively. We thus stress the need for language system design to be cognizant of the fact that humans may not always be operating sensibly and in good faith. For example, while the clarification systems discussed in this paper do function as intended as long as no human-issued directive is both ambiguous and immoral, robots will inevitably face adversarial directives, either by human ignorance or malice. Indeed, even children have been shown to spontaneously abuse and misuse robots out of curiosity [67]. It is also unknown whether these effects may arise with non-robotic language-capable technologies such as Apple's virtual assistant Siri. Revisiting language generation pipelines with moral implications and adversarial inputs in mind will yield robust software more suitable for real-world deployment.

Robots' ability to influence human networks of moral norms raises questions regarding the persistence and extent of this influence. Will the number of copresent human interactants

affect the robot's normative influence? Does the robot's normative influence persist outside of the current setting, or will it cease as soon as people leave the room? How long will the robot's effect on human norms last? Will humans be affected in the same ways from observing another person interacting with the robot as from interacting with the robot themselves? All of these questions will be crucial to investigate in future work. For the last question, our data may be taken as preliminary evidence that the effects are the same (see Section IV-B), but further research focused specifically on this question is needed.

Future work should also investigate the precise inferences people draw from these types of clarification dialogues. Specifically, *why* did we observe an increase in perceived permissibility following our clarification dialogues? Did participants infer that it was morally permissible to damage important equipment? That the robot believed the computers were not actually important? Or that the robot's creator had a good reason to allow the capacity to knock over computers? Answering these questions could help mitigate the issues identified, and would also help us understand how laypeople naturally perceive robots. If we knew what people were likely to infer, we might be better able to craft clarification requests that would either avoid or address those specific inferences. We believe that these types of questions are not well-suited to online experiments, and would be better answered in live experiments, with experimenters, participants, and robots physically copresent so as to facilitate free-form interviews.

Physical copresence of human subjects and robots in future experiments will also allow us to observe whether (and how) any robot influence on moral norms will manifest behaviorally. At this point, our findings are based only on self-reported survey responses; but the potential for robotic influence on moral norms will become much more concerning if it is shown to measurably alter human behavior or decision making.

Having identified issues with current clarification request generation algorithms, we hope to determine how language-enabled agents *should* respond to immoral and ambiguous commands, and create algorithms for generating appropriate responses. Some previous work explored when and how to reject commands for various reasons, including expressing moral qualms [68]. However, though normative impermissibility was considered a viable reason to reject a command, it remains unclear how best to realize such a rejection linguistically, how to algorithmically generate this linguistic realization, how humans will react to the rejection, and how the rejection might influence human morality. Other research has investigated responding to (unambiguous) moral infractions with affective displays [24] and humorous rebukes [69]. However, these represent only a small slice of possible responses, do not address the problem of co-occurring ambiguity, and are not tailored to specific contexts or infractions.

Based both on those previous studies and our current results, we believe that tactful responses to immoral commands could allow robots to positively reinforce the norm that was violated, instead of accidentally exacerbating the violation (as observed in our experiment). Responses that we plan to investigate include clarification requests designed to draw attention to the violated norm (e.g., "Do you really want me to knock over a computer?"), command refusals (e.g., "I cant do anything to harm laboratory equipment"), and rebukes (e.g., "You shouldnt ask me to destroy lab equipment. It's wrong."). It is not yet clear how such responses will be received in human-robot teams, nor how to maximize their efficacy, but we anticipate tuning the response type and phrasing to the context, severity, and intensity of the infraction. We will also need to calibrate the specificity of the responses such that they seem natural. For example, somewhere on the spectrum between "I cannot knock over either of these two computers." and "I cannot damage things", lies the more natural response "I cannot damage laboratory equipment."

VI. Limitations and Alternative Explanations

Our experimental design leaves open some alternative interpretations of our results. For the sake of transparency, we present these here. We plan on directly addressing these possibilities experimentally in the near future.

First, increases in perceived permissibility could have been caused by the human's request, not by the robot's response. Second, changes in permissibility ratings could have been caused by repeated exposure to the idea of knocking over a computer, rather than the clarification dialog.

Finally, knocking over a computer may sometimes be non-norm violating (e.g., if the computer was already broken). However, our pretests show that, before seeing any request to knock over a computer or clarification dialog, participants decisively viewed the action of knocking over a computer as impermissible, and believed that the robot shared this view. On a scale from impermissible (1) to permissible (7), the 95% credible interval for pretest permissibility ratings is 1.8 to 2.6 for participant views of permissibility, and 2.4 to 3.4 for assessments of the robot's view, so we do not believe that this was an issue here.

VII. Conclusion

Focusing on clarification request generation as an example, we have shown how subsystems of current natural language software architectures can bypass or preempt moral reasoning modules, and thereby unintentionally imply willingness to eschew moral norms. We have also shown decisive experimental evidence (barring caveats discussed in Section VI) that these implicatures will cause robots to (1) miscommunicate their moral intentions to human teammates, and (2) weaken the moral norms employed by human teammates within the current context. These results not only highlight the need to critically examine the moral facets of language-enabled robot architectures, but also, when considered in aggregate with the social robotics work discussed throughout this paper, provide evidence for the high-level hypothesis that robots are perceived as both social and moral agents, and are therefore active participants in the communal process of creating, maintaining, and altering norms, and will thus be subject to social judgments and consequences for violating those norms.

References

[1] M. Scheutz, P. Schermerhorn, J. Kramer, and D. Anderson, "First steps toward natural human-like HRI," *Autonomous Robots*, vol. 22, no. 4, pp. 411–423, May 2007.

[2] N. Mavridis, "A review of verbal and non-verbal human–robot interactive communication," *Robotics and Autonomous Systems*, vol. 63, pp. 22–35, 2015.

[3] C. Matuszek, "Grounded language learning: Where robotics and nlp meet." in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 5687–5691.

[4] P. Grice, "Logic and conversation," in *Syntax and Semantics*, 1975.

[5] S. C. Levinson, *Presumptive meanings: The theory of generalized conversational implicature*. MIT press, 2000.

[6] K. Bach, "The top 10 misconceptions about implicature," *Drawing the boundaries of meaning: Neo-Gricean studies in pragmatics and semantics in honor of Laurence R. Horn*, pp. 21–30, 2006.

[7] T. Williams, G. Briggs, B. Oosterveld, and M. Scheutz, "Going beyond command-based instructions: Extending robotic natural language interaction capabilities," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[8] S. Trott, M. Eppe, and J. Feldman, "Recognizing intention from natural language: clarification dialog and construction grammar," in *Workshop on Communicating Intentions in Human–Robot Interaction*, 2016.

[9] R. A. Knepper, "On the communicative aspect of human-robot joint action," in *The IEEE International Symposium on Robot and Human Interactive Communication Workshop: Toward a Framework for Joint Action, What about Common Ground*, 2016.

[10] L. Benotti and P. Blackburn, "Polite interactions with robots," *What Social Robots Can and Should Do: Proceedings of Robophilosophy 2016/TRANSOR 2016*, vol. 290, p. 293, 2016.

[11] G. Briggs, T. Williams, and M. Scheutz, "Enabling robots to understand indirect speech acts in task-based interactions," *Journal of Human-Robot Interaction (JHRI)*, 2017.

[12] F. Gervits, G. Briggs, and M. Scheutz, "The pragmatic parliament: A framework for socially-appropriate utterance selection in artificial agents," in *Proceedings of the Annual Meeting of the Cognitive Science Society (COGSCI)*, 2017.

[13] D. Fried, J. Andreas, and D. Klein, "Unified pragmatic models for generating and following instructions," *arXiv preprint arXiv:1711.04987*, 2017.

[14] S. Trott and B. Bergen, "A theoretical model of indirect request comprehension," in *Proceedings of the AAAI Fall Symposium Series on Artificial Intelligence for Human-Robot Interaction (AI-HRI)*, 2017.

[15] T. Williams, D. Thames, J. Novakoff, and M. Scheutz, ""Thank you for sharing that interesting fact!": Effects of capability and context on indirect speech act use in task-based human-robot dialogue," in *Proceedings of the 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2018.

[16] R. C. Arkin, "Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture," in *Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2008, pp. 121–128.

[17] M. Scheutz, "The need for moral competency in autonomous agent architectures," in *Fundamental Issues of Artificial Intelligence*. Springer, 2016, pp. 515–525.

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

[19] K. Wada and T. Shibata, "Living with seal robots – its sociopsychological and physiological influences on the elderly at a care house," *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 972–980, 2007.

[20] N. Sharkey and A. Sharkey, "The crying shame of robot nannies: an ethical appraisal," *Interaction Studies*, vol. 11, no. 2, pp. 161–190, 2010.

[21] J. Wen, A. Stewart, M. Billinghurst, A. Dey, C. Tossell, and V. Finomore, "He who hesitates is lost (...in thoughts over a robot)," in *Proceedings of the Technology, Mind, and Society*, ser. TechMindSociety '18. New York, NY, USA: ACM, 2018, pp. 43:1–43:6. [Online]. Available: http://doi.acm.org/10.1145/3183654.3183703

[22] P. Lin, G. Bekey, and K. Abney, "Autonomous military robotics: Risk, ethics, and design," Cal. Poly. State Univ. San Luis Obispo, Tech. Rep., 2008.

[23] B. Scassellati, H. Admoni, and M. Mataric, "Robots for use in autism research," *Annual Review of Biomedical Engineering*, vol. 14, pp. 275–294, 2012.

[24] G. Briggs and M. Scheutz, "How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress," *International Journal of Social Robotics*, 2014.

[25] P. H. Kahn, T. Kanda, H. Ishiguro, B. T. Gill, J. H. Ruckert, S. Shen, H. Gary, A. L. Reichert, N. G. Freier, and R. L. Severson, "Do people hold a humanoid robot morally accountable for the harm it causes?" in *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Boston, MA, 2012, pp. 33–40.

[26] B. F. Malle, M. Scheutz, T. Arnold, J. Voiklis, and C. Cusimano, "Sacrifice one for the good of many?: People apply different moral norms to human and robot agents," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, Portland, OR, 2015, pp. 117–124.

[27] R. Simmons, M. Makatchev, R. Kirby, M. K. Lee *et al.*, "Believable robot characters," *AI Magazine*, no. 4, 2011.

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

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

[30] J. Kennedy, P. Baxter, and T. Belpaeme, "Children comply with a robot's indirect requests," in *Proceedings of HRI*. Bielefeld, Germany: ACM, 2014, pp. 198–199.

[31] F. Eyssel and D. Kuchenbrandt, "Social categorization of social robots: Anthropomorphism as a function of robot group membership," *British Journal of Social Psychology*, no. 4, 2012.

[32] T. Williams, R. B. Jackson, and J. Lockshin, "A bayesian analysis of moral norm malleability during clarification dialogues," in *Proceedings of the Annual Meeting of the Cognitive Science Society (COGSCI)*. Madison, WI: Cognitive Science Society, 2018.

[33] R. B. Jackson and T. Williams, "Robot: Asker of questions and changer of norms?" in *Proceedings of the International Conference on Robot Ethics and Standards (ICRES)*. Troy, NY: CLAWAR Association, 2018.

[34] M. R. J. Purver, "The theory and use of clarification requests in dialogue," Ph.D. dissertation, University of London, 2004.

[35] D. Bohus and A. I. Rudnicky, "Sorry, I didn't catch that!-an investigation of non-understanding errors and recovery strategies," in *6th SIGdial workshop on discourse and dialogue*, 2005.

[36] M. Marge and A. I. Rudnicky, "Miscommunication recovery in physically situated dialogue," in *Proceedings of the 16th Annual SIGdial Meeting on Discourse and Dialogue*, Saarbrücken, Germany, 2015, pp. 22–49.

[37] S. Tellex, P. Thaker, R. Deits, D. Simeonov, T. Kollar, and N. Roy, "Toward information theoretic human-robot dialog," *Robotics: Science and Systems*, vol. 32, pp. 409–417, 2013.

[38] T. Williams and M. Scheutz, "Resolution of referential ambiguity in human-robot dialogue using dempster-shafer theoretic pragmatics," in *Proceedings of Robotics: Science and Systems (RSS)*, Cambridge, MA, 2017.

[39] T. Williams, F. Yazdani, P. Suresh, M. Scheutz, and M. Beetz, "Dempster-shafer theoretic resolution of referential ambiguity," *Autonomous Robots*, 2018.

[40] M. Scheutz, G. Briggs, R. Cantrell, E. Krause, T. Williams, and R. Veale, "Novel mechanisms for natural human-robot interactions in the diarc architecture," in *Proceedings of AAAI Workshop on Intelligent Robotic Systems*, 2013.

[41] M. Scheutz, T. Williams, E. Krause, B. Oosterveld, V. Sarathy, and T. Frasca, "An overview of the distributed integrated cognition affect and reflection diarc architecture," in *Cognitive Architectures*, M. I. A. Ferreira, J. Sequeira, and R. Ventura, Eds., 2018 (in press).

[42] M. Scheutz, B. Malle, and G. Briggs, "Towards morally sensitive action selection for autonomous social robots," in *Proc. of RO-MAN*, 2015.

[43] C. Nass, J. Steuer, and E. R. Tauber, "Computers are social actors," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 1994, pp. 72–78.

[44] T. Gureckis, J. Martin, J. McDonnell *et al.*, "psiturk: An open-source framework for conducting replicable behavioral experiments online," *Behavior Research Methods*, vol. 48, no. 3, pp. 829–842, 2016.

[45] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data?" *Perspectives on Psychological Science*, vol. 6, no. 1, pp. 3–5, 2011.

[46] M. J. Crump, J. V. McDonnell, and T. M. Gureckis, "Evaluating amazon's mechanical turk as a tool for experimental behavioral research," *PloS one*, vol. 8, no. 3, 2013.

[47] N. Stewart, J. Chandler, and G. Paolacci, "Crowdsourcing samples in cognitive science," *Trends in Cognitive Sciences*, 2017.

[48] W. Bainbridge, J. Hart, E. Kim, and B. Scassellati, "The benefits of interactions with physically present robots over video-displayed agents," *International Journal of Social Robotics*, vol. 3, no. 1, pp. 41–52, 2011.

[49] K. Fischer, K. Lohan, and K. Foth, "Levels of embodiment: Linguistic analyses of factors influencing HRI," in *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Boston, MA, 2012, pp. 463–470.

[50] J. Li, "The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents," *International Journal of Human-Computer Studies*, vol. 77, pp. 23–37, 2015.

[51] K. Tanaka, H. Nakanishi, and H. Ishiguro, "Comparing video, avatar, and robot mediated communication: Pros and cons of embodiment," in *Proceedings of the International Conference on Collaboration Technologies (ICCT)*. Minneapolis, MN: Springer, 2014, pp. 96–110.

[52] JASP Team *et al.*, "Jasp," *Version 0.8. 0.0. software*, 2016.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] A. Ly, A. Etz, M. Marsman, and E.-J. Wagenmakers, "Replication bayes factors from evidence updating," *Behavior Research Methods*, Aug 2018. [Online]. Available: https://doi.org/10.3758/s13428-018-1092-x

[60] D. Wright, "Comparing groups in a before-after design: When t test and ancova produce different results," *The British journal of educational psychology*, vol. 76, pp. 663–75, 10 2006.

[61] D. Dimitrov and P. D Rumrill, "Pretest-posttest designs and measurement of change," *Work (Reading, Mass.)*, vol. 20, pp. 159–65, 02 2003.

[62] S. Huck and R. A. McLean, "Using a repeated measures anova to analyze the data from a pretest-posttest design: A potentially confusing task," *Psychological Bulletin*, vol. 82, pp. 511–518, 07 1975.

[63] H. Jeffreys, *Theory of Probability*. Clarendon Press, Oxford, 1961.

[64] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 1988.

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

[66] A. Freedy, E. DeVisser, G. Weltman, and N. Coeyman, "Measurement of trust in human-robot collaboration," in *Proceedings of the Symposium on Collaborative Technologies and Systems*, 2007, pp. 106–114.

[67] T. Nomura, T. Uratani, T. Kanda, K. Matsumoto, H. Kidokoro, Y. Suehiro, and S. Yamada, "Why do children abuse robots?" in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, ser. HRI'15 Extended Abstracts. New York, NY, USA: ACM, 2015, pp. 63–64. [Online]. Available: http://doi.acm.org/10.1145/2701973.2701977

[68] G. Briggs and M. Scheutz, ""Sorry, I can't do that": Developing mechanisms to appropriately reject directives in human-robot interactions," in *Proceedings of the AAAI Fall Symposium Series*, 2015.

[69] M. F. Jung, N. Martelaro, and P. J. Hinds, "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, 2015, pp. 229–236.