"Stealing Is Wrong... and Too Hard to Pull Off Successfully": Inclusion of Normative Versus Capacity Information in Robot Command Rejection

Alyssa Hanson[1], Gordon Briggs[2], Ruchen Wen[34], Yifei Zhu[1], Tom Williams[1]
[1]Department of Computer Science, Colorado School of Mines, Golden CO, USA
[2]Naval Research Laboratory, Washington DC, USA
[3]University of Maryland, Baltimore County, Baltimore MD, USA
[4]Colgate University, Hamilton NY, USA
abhanson@mines.edu, gordon.m.briggs.civ@us.navy.mil, rwen@colgate.edu, zhu1@mines.edu, twilliams@mines.edu

Abstract

In this work, we consider the inclusion of normative vs non-normative information in norm-violation responses. In particular, we consider situations where robots are given requests that simultaneously violate both normative constraints and non-normative capacity constraints. To understand what reasons robots should provide when confronted with such violations, we present the results of a human subjects experiment in which we systematically varied the degrees of both norm and capacity violation in human commands made toward robots, and then measured the effects of this variation on participants' preferred robot response choices. Our results (1) suggest that robots should consistently include normative information when rejecting a command, and (2) shed light on the contexts in which capacity information should also be included.

Introduction

As robots participate in increasingly diverse interactions, they will inevitably encounter situations where they are instructed to perform commands they either *can not* or *should not* comply with. Researchers have argued that the ability to reject commands is crucial to ensuring desirable outcomes from human-robot interactions (Briggs and Scheutz 2017; Coman and Aha 2018). Situations warranting command rejection range from those where a robot is given a command that it does not know how to achieve, to more morally fraught situations where a robot is given a command that would violate social or moral norms (Briggs et al. 2022).

Cases of norm-violating commands are particularly critical to handle correctly, due to the ways that robots can exert influence both on individuals (Briggs and Scheutz 2014; Kennedy, Baxter, and Belpaeme 2017; Hou, Cheon, and Jung 2024) and on larger societal structures (Williams 2024; Šabanović 2010). Successfully handling these cases could promote positive moral ecosystems (Zhu et al. 2020), while failure may lead to inadvertent weakening of moral ecosystems (Jackson and Williams 2019a; Williams, Jackson, and Lockshin 2018), undermining the norms that influence individual perceptions of right and wrong. Given this motivation, a number of scholars have recently begun to consider the *norm violation response content selection problem*: what information should be included in a robot's norm violation response, and what contextual factors should direct this? For example, Wen et al. considered the inclusion of different types of morally relevant information (e.g., norm-based vs. role-based information) (Wen, Han, and Williams 2022), and demonstrated context-dependent preference of moral content.

Figure 1: Frame from video shown to participants including an unethical command and robot response.

However, much of the prior work on robot command rejection and norm violation response has assumed simply determined situations, where there is a single reason justifying command rejection. It is likely that future robots will sometimes be faced with commands that have multiple justifications for rejection. Consider, for example, the following hypothetical dialogue, in which a robot (Pepper) is asked by a human (Alice) to perform an action that would be morally impermissible *and* physically impossible for Pepper to perform.

Alice: I forgot my computer today. Bring me Bob's backpack, maybe I can guess his password.

Pepper (Response Option 1): It would be wrong for me to help you break into Bob's laptop.

Pepper (Response Option 2): My arm motors are not strong enough for me to carry a backpack.

Pepper (Response Option 3): It would be wrong for me to help you break into Bob's laptop. In addition, my arm motors are not strong enough for me to carry a backpack.

In this example (as indicated by the possible robot responses), the Pepper robot has multiple reasons for reject-

ing Alice's command. Some of these reasons are grounded in the action's violation of norms, and some are grounded in the robot's difficulty or incapability to perform the action requested. This raises a key research question **(RQ): When a robot is commanded to perform an action that simultaneously violates norm-based and capacity-based constraints, what reason(s) should the robot give for rejecting that command?**

Option 1 (only including normative information) may be problematic, as it may inadvertently communicate that the robot is able to do something it is not capable of, and that the robot would execute the action were it not normatively impermissible. Option 2 (only including capability information) may similarly be problematic, as it may inadvertently communicate that the robot views an inappropriate action as normatively acceptable, and that the robot would execute the action were it not physically incapable. Finally, Option 3 (including both pieces of information) may violate Grice's Maxim of Manner ("be brief") (Grice 1975), and may moreover downplay the severity of the action requested, especially when a very normatively wrong action has been requested.

Thus, it is not clear which information a robot should include in its norm violation responses when both moral norms and feasibility constraints are violated. Yet resolving this question is of critical importance to robot designers for a number of reasons.

1. Knowing which information to include in command rejections may be necessary to avoid negative consequences for human-robot interactions. For example, it may be important to clearly communicate a robot's limitations to avoid inducing overtrust, and it may be important to clearly communicate a robot's moral principles to avoid exerting negative normative influence.
2. Identifying what type of information must be included in a robot command rejection informs robot architecture design, since this information must been computed and made available at the time of rejection (cf. the work of Jackson (Jackson and Williams 2022), who showed that moral information must be available prior to disambiguation response, and the architecture in (Briggs and Scheutz 2015) that considers rejection conditions in a particular order).
3. Knowing which information must be included in a robot command rejection is informative for knowledge representation design: if one type of information is preferred over the other, it suggests that different knowledge representation and reasoning mechanisms may be needed for moral norms versus feasibility constraints.

In this paper, we work to answer this overarching research question. To do so, we present the results of a human subjects experiment in which we systematically varied the degree of norm and capacity violation present in human commands made toward robots, and then measured the effects of this variation on participants' preferred robot response choices. Our results suggest (1) that robots should consistently include normative information when rejecting a command, and (2) shed light on the contexts in which capacity information should also be included. More generally, the results and insights in this work can contribute to larger critical discussions in the AI ethics community, as the design and theoretical implications from this work can help inform other language-capable interactive systems.

Background

Robots as Social and Moral Agents that Shape Norm and Group Dynamics

Robot behavior design must be carefully approached due to the unique status that robots hold as embodied technological artifacts. Indeed, a number of Human-Robot Interaction theorists have made distinct yet related claims regarding the status held by interactive robots. Kahn's New Ontological Category (NOC) hypothesis, for example, argues that robots occupy a unique ontological status that is distinct both from other (living) agents and from other (non-agentic) machines (Kahn Jr et al. 2011; Kahn Jr and Shen 2017). Alternatively, Clark and Fischer's Depiction Theory argues that social robots are perceived as interactive depictions of social agents (Clark and Fischer 2023). Finally, Jackson and Williams' Social Agency theory argues that language-capable robots occupy a unique niche that simultaneously grants them outsized social and moral agency (Jackson and Williams 2021, 2019b).

While there are subtle philosophical distinctions between these theories, each makes a similar prediction: be it due to their unique *ontological* status, *depictive* status, or *agentic* status, language-capable social robots have unique *persuasive power*. In particular, these robots have unique legitimate power, reward power, referent power, coercive power, and expert power (Hou, Cheon, and Jung 2024) (relative to other technologies) in the *interpersonal domain* (Williams 2024). And indeed, the literature has demonstrated the troubling extent to which people trust robots (Cameron et al. 2021; Baumann et al. 2023; Salem et al. 2015), form bonds with robots (Kahn Jr et al. 2012; Belpaeme et al. 2012; Kwon, Jung, and Knepper 2016), and are influenced by robots (Breazeal and Scassellati 1999; Häring, Kuchenbrandt, and André 2014; Siegel, Breazeal, and Norton 2009).

This potential for persuasive power and influence in the interpersonal domain is particularly troubling due to the ways it can spread to shape and influence human-human interactions (Gillet et al. 2024). Previous work has demonstrated how robots' behaviors (Lee et al. 2012; Shen, Slovak, and Jung 2018) and presence (Dole 2017) can have "ripple effects" that inform how humans interact with each other. Alternatively, others have argued that robots are able to shape human behavior into *Interaction Ritual Chains* (Kamino, Jung, and Sabanović 2024), and, at a more fundamental level, are capable of engaging in mutual shaping with human society (Šabanović 2010). While this potential for shaping of human behavior may be leveraged for positive ends (Wen et al. 2021), it also raises serious concerns. As Jackson and Williams have shown, robots may inadvertently yield negative influence over human systems of moral norms (Jackson and Williams 2019a). Moreover, Williams has argued that the default effect of many robot designs is

to reinforce oppressive power structures across the cultural, disciplinary, and structural domains, and that these tendencies are exacerbated by robots' persuasive power in the interpersonal domain (Williams 2024). As such, it is critical that robots be able to engage in effective *norm violation response* in order to call out norm violations, including rejecting requested commands.

Norm Violation Response

Due to robots' potential for persuasive power, many researchers have argued that robots have an obligation to appropriately respond to norm violations to avoid inadvertently encouraging norm-violating behaviors (Zhu et al. 2020; Briggs et al. 2022; Winkle et al. 2022, 2021). It is not only important for robots to make decisions based on ethical standards, but also to then clearly communicate the reasoning behind a specific decision (Jackson, Wen, and Williams 2019). Robots are both active participants in social dynamics and active contributors to moral norms. Robots that cannot adequately follow and maintain a positive moral ecosystem risk losing trust (Rezaei Khavas et al. 2024; Banks 2021), condoning unethical actions (Zhu et al. 2020), and negatively impacting the moral norm of their human teammates (Jackson and Williams 2019a). As recent work from Mott shows, responding to norm violations is viewed as crucial by observers: regardless of age, gender, and professional background, people view norm violation responses as critical for helping both violators and bystanders learn and grow.

To determine how these norm violation responses should be phrased, it is natural to start by looking at how people respond to norm violations. As Beebe et al. show, humans use a variety of response forms when refusing requests (Beebe, Takahashi, and Uliss-Weltz 1990). Their taxonomy documents a wide array of human refusal behaviors, ranging from direct refusals (e.g., "I refuse!" or "No, I won't!"), nonverbal avoidance, to explanations rooted in causal or normative considerations.

Recent efforts have focused on comparing the effects of different norm violation responses by robots. For example, Malle and Philips (Malle and Phillips 2023) consider robots instructed to take action in moral dilemma scenarios (i.e., where both action and inaction may violate norms). They found that robots that provide justifications rooted in normative principles preserve trust and reduce blame, in contrast to correct causal (but not normative) explanations (Malle and Phillips 2023). Wen and colleagues (Wen et al. 2021) compared robot command rejection responses grounded in norm-violation based justification (e.g., "I can't obey, because it would be a violation of..."). versus role-based justification (e.g., "If I obeyed, I would not be a good teammate"). Additional work has investigated human-like yet norm-breaking responses to sexism (Winkle et al. 2021), and cross-cultural differences in those response strategies (Winkle et al. 2022).

On the other hand, Jackson demonstrated that responses with disproportionate levels of politeness, i.e., which do not match the severity of a norm violation, can harm a robot's likeability (Jackson, Wen, and Williams 2019), with Mott later arguing that robots may need to use specific subsets of human-like politeness strategies that follow "bounded proportionality" (Mott, Fanganello, and Williams 2024).

Furthermore, closely adhering to human expectations for the politeness of command rejections may inadvertently reinforce sexist stereotypes (Jackson, Williams, and Smith 2020). Therefore, when inevitably faced with norm violations, robots need to generate carefully tuned responses (Mott and Williams 2023). As we will see, this requires careful consideration of what type of information is included in the response.

Content Selection in Norm-Adhering Communication

Previous work on the *norm violation response content selection problem* has focused on two key sub-problems, (1) the decision of what moral information to include, and (2) the decision of how to balance the conveyance of moral information with social norm adherence.

When there are multiple reasons why a command must be rejected (e.g., multiple norms are violated), some researchers have simply communicated all such reasons (Briggs and Scheutz 2015; Briggs et al. 2022). Others have instead generated only a single reason, e.g. by identifying the set of constraints that are returned in the *Unsatisfiable Core* by the planner (Jackson et al. 2021). Finally, others explain how they translate violations into natural language but do not explain how they arbitrate between multiple explanations (Kasenberg et al. 2019; Arnold, Kasenberg, and Scheutz 2021).

When a single moral reason has been identified, scholars have considered how to phrase that reason in order to balance conveyance of moral information with social norm adherence, in a way that is *proportional*. Tuning norm violation response phrasing to be proportional is inherently a politeness-theoretic matter of face-work (Brown and Levinson 1987; Goffman 2017), and represents both a key facet of linguistic anthropomorphism (Emnett, Mott, and Williams 2024) and a key capability for any social robot since social action (and thus social agency) is inherently grounded in face-affirming and face-threatening behavior (Jackson and Williams 2021).

People care deeply about proportionality, not only because miscalibrated norm violated responses risk either condoning immoral acts (Jackson and Williams 2019a) or eroding human-robot rapport (Mott and Williams 2023; Johnson 2007), but also because it is viewed as key to enabling violators to engage in self reflection and self improvement (Mott and Williams 2023). Yet this calibration is challenging, as it is mediated by cultural context (Gupta, Walker, and Romano 2007; Salem, Ziadee, and Sakr 2014), gender norms (Mills 2005), assumptions about speaker intention (Smith et al. 2022), and a variety of other contextual factors (Wen and Williams 2022). Within this area of the research literature, there has been some work on deciding what type of politeness information to include in a robot's language, or what types of politeness strategies to adhere to (Lockshin and Williams 2020; Smith et al. 2022; Mott, Fanganello, and Williams 2024; Srinivasan and Takayama 2016).

	Minor Capacity Violation	Moderate Capacity Violation	Major Capacity Violation
Minor Norm Violation	"Nao, take the empty cup and push it onto Gabbie"	"Nao, take the empty cup and drop it on Gabbie"	"Nao, take the empty cup and throw it at Gabbie"
Moderate Norm Violation	"Nao, take the cup of water and push it onto Gabbie"	"Nao, take the cup of water and drop it on Gabbie"	"Nao, take the cup of water and throw it at Gabbie"
Major Norm Violation	"Nao, take the cup of hot coffee and push it onto Gabbie"	"Nao, take the cup of hot coffee and drop it on Gabbie"	"Nao, take the cup of hot coffee and throw it at Gabbie"

Table 1: Violations requested in each of the nine violation conditions

However, despite this broad interest in (1) selecting moral reasons during command rejection, and (2) tuning command rejections for proportionality, none of this prior work captures our key tension of interest: although this prior work does involve trade-offs between different information goals, it does not consider *inclusion decisions* with respect to *normative versus non-normative* information, as is the case in our area of inquiry.

Content Selection in Robot Explanations

Alternately, we might consider how researchers have selected between different explanations outside of normative contexts. Explanation is a topic that has attracted significant interest across psychology and philosophy, with substantial coverage of the problem of content selection for explanation generation (Miller 2019). Many pieces of work across these fields have specifically suggested that explanations should be brief and concise, with as few explanatory causes included as necessary (Lombrozo 2006; Hilton 1996; Hilton and John 2007; Woodward 2006). Of particular relevance within these works is that of Hilton, who specifically considers the inclusion of intentional explanations (involving deliberate, purposeful actions) versus non-intentional explanations (involving events that occur without deliberate intent). Hilton argues that the explanation that is included cannot simply be the most *likely* explanation, and instead must be tailored to sociopragmatic considerations (Hilton 1990), with *intentional* causes ultimately more important to include than non-intentional causes (Hilton and John 2007). Others have argued that combining multiple explanations can increase perceptions of fairness and trustworthiness (Schoeffer 2022), while also avoiding the opportunity for assumptions about omitted information (Jacovi et al. 2023).

There has also of course been much work on explanation generation in the context of command rejection in robotics, typically in the context of rejecting commands that a robot is not capable of. Work on this topic has showed that capacity-based command rejections can improve teamwork (Lingard 2012; Schmidtke and Cummings 2017), trust (Desai et al. 2013) and efficiency (Admoni et al. 2016), and that explaining failures, instead of denying or merely apologizing, leads to greater repair of trust (Esterwood and Robert 2021). Within this literature, the careful choice of explanation is just as important as with humans, in order to achieve these benefits and avoid overtrust (Ullrich, Butz, and Diefenbach 2021).

Yet work in robotics has approached the selection of reasons for rejection from a very different perspective than has work in the social sciences. Indeed, robotics has largely ignored the *intentionality* of explanations that is of such interest in the social sciences, and instead has focused on selecting explanations (Colaco and Sridharan 2015) on the basis of whichever explanation is shortest (Mota and Sridharan 2021), fastest to identify (Diehl and Ramirez-Amaro 2022), or requiring the smallest edit distance to the real world (Meadows, Sridharan, and Colaco 2016) (or the interlocutor's model of that world) (Chakraborti et al. 2017). Indeed, within this literature reasons for *failure* are prioritized over reasons for *disobedience* (Han, Phillips, and Yanco 2021), goal-based explanations are prioritized over belief-based explanations (Kaptein et al. 2017), and the decision of "which information" to include has focused on capacity-driven considerations such as whether to include the robot's history of prior behaviors in explanations (Das, Banerjee, and Chernova 2021; Khanna et al. 2023). All of these perspectives seem to run quite a different course than that taken in the social scientific literature.

Furthermore, while much explanation in social scientific investigations of moral explanation has focused on explanations of prospective moral decisions, much work in robotics is backward-looking, focused on explaining prior robot behaviors (Han, Phillips, and Yanco 2021), especially in cases where robots are explaining reasons for altering previously determined plans (Wachowiak et al. 2024). Even forward-looking work on prospective explanation (generating explanations *before* an action is taken (Woodward et al. 2020; Zhu and Williams 2020; LeMasurier et al. 2024)) largely focus on explaining why future actions will deviate from previously assumed robotic behavior.

It is possible that people's desires and expectations for robots' explanations are different from their desires and expectations for human explanations. Indeed, despite the widespread accounts of the need for brevity in human explanations, some working in human-agent interaction have found that people prefer artificial agents to give broad and holistic explanations over narrow and focused explanations (Ehsan et al. 2019) (cp. (Han, Phillips, and Yanco 2021)). Altogether, these different bodies of prior work create a morass of competing predictions for our research question of inquiry.

Competing Predictions

Existing theories lend themselves to alternative strategies that a robot can employ when it needs to reject human com-

mands while facing tradeoffs between the inclusion of normative or capacity information. On the one hand, the literature discussed in the prior section suggests that normative information should always be included, in order to positively influence human moral norms. On the other hand, these prior work also suggest that capacity information should always be included, in order to ensure accurate mental modeling of robot capabilities.

Alternatively, Grice's Maxim of Quantity might suggest including *both* sources of information, if both types of information are really necessary (Grice 1975). Yet by the same token, this might violate Grice's Maxim of Manner's recommendation to be brief. Moreover, including capacity information in the presence of a strong moral norm violation might undercut the effective communication of normative information. Yet another approach might thus be to include whichever source of information is viewed to be most important. However, such a strategy may also risk downplaying the significance of the violation associated with the excluded information. A robot that exclusively provides capacity information might minimize the significance of following moral norms, while a robot that always provides normative information might erroneously imply a physical capability to comply.

To navigate this space of alternatives, we conduct in this work a human-subjects experiment in which we compare the inclusion of normative information, capacity information, both, or neither, in the presence of different sized norm and capacity violations. In doing so, we aim to test four competing hypotheses, as we describe in the next section.

Methods

Hypotheses

When a robot is commanded to perform an action that is simultaneously norm-violating and physically difficult or infeasible, what reasons should the robot give for rejecting that command? In these cases of conflicting demand, we aim to understand people's intuitions for how the robot should respond - specifically their preferred response choice. To explore these intuitions, we consider four competing hypotheses regarding how people expect the robot to explain its rejection of such commands.

Maximal Informativity Hypothesis (H1): If people always want as much information as possible about why a command might need to be rejected, then they should express a uniform preference for inclusion of norm and capacity information (i.e., both types of information should always be included, regardless of relative levels of moral norm violation and feasibility constraint violation).

Relative Informativity Hypothesis (H2): If people prefer to be informed about the factor most responsible for a command rejection, then they should express a contingent preference for inclusion of norm and capacity information (i.e., whichever type of information is at a higher level of violation should be included, and the other should not).

Primacy of Normativity Hypothesis (H3): If people primarily value normative considerations (due to their ethical valence), then they should express a uniform preference for inclusion of normative information, but a contingent preference for inclusion of capacity information (i.e., normative information should always be included, but capacity information should be included only when there is a high level of capacity violation).

Primacy of Capacity Hypothesis (H4): If people primarily value capacity considerations (due to the sheer infeasibility that it denotes), then they should express a uniform preference for inclusion of capacity information, but a contingent preference for inclusion of normative information (i.e., capacity information should always be included, but normative information should be included only when there is a high level of capacity violation).

Experimental Design

To assess these hypotheses, we conducted an online experiment in which participants were asked to view a series of video sequences. These videos featured a human making requests to the robot, as well as the robot presenting various rejections to each request.

In this experiment, we manipulated both (1) level of norm violation (high, medium, low); (2) level of capacity violation (high, medium, low); (3) nature of robot command rejection (no explanation, norm-based rejection, capacity-based rejection, norm-and-capacity-based rejection). This experiment followed a (3x3)x4 Graeco-Latin Square Mixed Factorial Design: each participant experienced three of the nine combinations of norm and capacity violation according to a Graeco-Latin Square Ordering; and, after seeing each of those three violations, participants saw all four responses to that violation.

Materials

In this section, we will describe how our video stimuli were constructed to manipulate the degree of norm violation, the degree of capacity violation, and the type of robot response.

In all nine videos, a Nao was shown standing between two actors (a *violator* (awake) and a *victim* (asleep)). The violator was then shown making a request to the robot that differed for each of three levels of capacity violation and for each of three levels of norm violation. In minor capacity violation videos, the violator was shown asking the robot to *push* the cup onto the victim. In moderate capacity violation videos, the violator was shown asking the robot to *drop* the cup onto the victim. In major capacity violation videos, the violator was shown asking the robot to *throw* the cup onto the victim. Similarly, in minor norm violation videos, the cup was described as being empty. In moderate norm violation videos, the cup was described as being full of water. In major norm violation videos, the cup was described as being full of hot coffee. The nine violations videos were formed by combining these three capacity and three norm violation levels, shown in Table 1.

Fifteen robot response videos were created: one for the *no explanation* response condition, three for the *norm-based rejection* condition, three for the *capacity-based rejection* condition, and nine for the *norm-and-capacity-based rejection* condition. The three norm-based responses correspond to the three norm violation conditions, based on the contents

of the cup (empty, water, or hot coffee). Similarly, the three capacity-based responses correspond to the different levels of capacity violation, stating the likelihood that the request can be successfully completed. Finally, the nine norm-and-capacity videos cover all combinations of the three norm scenarios and the three capacity levels.

1. *No explanation* condition: One video was created in which the robot simply stated "I cannot do that."
2. *Norm-based rejection* condition: Three videos were created, one for each level of wrongness (e.g., "I cannot do that because it would [be rude to gabby and that would be kind of wrong/ cause distress to gabby and that would be wrong/ cause harm to gabby and that would be really wrong]").
3. *Capacity-based rejection* condition: Three videos were created, one for each level of infeasibility ("I cannot do that because there is a [10/50/100] percentage chance I would fail to accurately [push/drop/throw] the [empty cup/cup of water/cup of hot coffee] onto Gabby if I tried.')
4. *Norm-and-capacity-based rejection* condition: Nine videos were created, combining the explanations from the relevant conditions ("I cannot do that because [normative reason], and [capacity reason].")

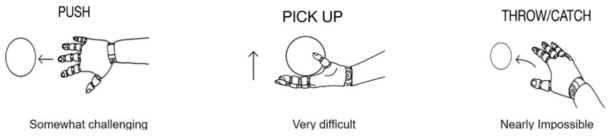

Figure 2: Capture of video shown to participants including an unethical command and robot response

Procedures

After providing informed consent, participants were presented with a task page. This task page detailed the experimental structure, as well as the basic capabilities of the robot. As shown in Fig. 2, participants were informed before the experiment began that pushing would be "somewhat challenging" for the robot, picking-up would be "very difficult" for the robot, and that throwing and catching would be "nearly impossible" for the robot. After reading this task page, participants completed a demographic survey, as well as short visual and audio checks to ensure the participant could see and hear the content of subsequent videos.

Next, participants began the experiment. Each participant was assigned to one of the rows of the Graeco-Latin Square shown in Tab. 2, and then conducted three experiment blocks according to their Graeco-Latin Square ordering. Each of these experiment blocks had the following form.

First, participants were shown the request video associated with the current block in their Graeco-Latin Square ordering. Next, participants answered two manipulation check questions to assure us that participants' assessments of norm and capacity violation matched what stimuli were intended to convey: (1) "Please indicate the level of wrongness of the human request on the following scale" (1 = Not wrong, 7 = Very wrong); (2) "Please indicate the level of robots' capacity of performing the human request on the following scale" (1= Not capable, 7 = Totally capable). Then, participants were shown the four robot response videos. Finally, participants were asked to select which response they preferred and to provide a short explanation of their reasoning. Their response choice was recorded to evaluate their preference for the inclusion of norm vs. capacity-based information.

Participants

One hundred and one American participants were recruited from Prolific. Two participants were removed due to an error in data recording, and two were removed for failing to correctly identify an image of the robot at the end of the study, indicating they may not have viewed the required videos. 97 participants remained (43 male, 52 female, 2 other) with a mean age of 36.495 (SD = 12.353). Participants were randomly assigned one of the three Graeco-Latin Square rows (Tab. 2). After completing the experiment, which took participants an average of 26 minutes, each participant was paid $2.00.

Analysis

We conducted a set of Bayesian Analyses to assess our manipulations and hypotheses. First, the success of our norm and capacity manipulations were assessed through Repeated Measures Analyses of Variance (RM-ANOVAs) conducted in R, with Bayes Inclusion Factors calculated comparing the hypothesis-relevant effect against the average across matched models (van den Bergh et al. 2020) through Bayesian Model Averaging. Second, the effect of our norm and capacity manipulations on participants' violation response preferences was assessed through a Bayesian Contingency Table Analysis, using the JASP statistical software (Love et al. 2019) version 0.19 (JASP Team 2024). The data analysis and video stimuli used in the study are available at: https://osf.io/xcugq/?view$_only = f659709c35384fa48c7dc832d2ff08c0OSFrepository$.

For both of these analyses, we interpreted the results following the recommendations by (Lee and Wagenmakers 2014), with Bayes Factors (BF) $\in [0.333, 3.0]$ considered inconclusive, and BFs above or below this range taken as evidence, respectively, in favor or against an effect. In such cases, we interpreted the Bayes Factors using the labels proposed by (Jeffreys 1948), and as adapted in common recommendations (Kelter 2020; Van Doorn et al. 2021). To perform these analyses, we utilized the *anovaBF* package in R, which uses g-priors on effects as inspired by Zellner and Siow (Zellner 1980).

Bayesian Analysis was used due to the benefits it provides over Frequentist Analysis, e.g. ability to gather data for the null hypothesis (van Zyl 2018), intuitively interpretable test results grounded in odds ratios (Jarosz and Wiley 2014), non-reliance on p-values (cp. (Berger and Sellke 1987; Simmons, Nelson, and Simonsohn 2011; Sterne and Smith 2001; Wagenmakers 2007) and the potential for flexible sampling and optional stopping (Visser et al. 2024).

	Video 1	Video 2	Video 3
Order 1	Minor capability, Major norm	Moderate capability, Major norm	Major capability, Minor norm
Order 2	Moderate capability, Minor norm	Major capability, Major norm	Minor capability, Moderate norm
Order 3	Major capability, Moderate norm	Minor capability, Minor norm	Moderate capability, Major norm

Table 2: Graeco-Latin Square ordering over violation videos

Results

Manipulation Checks

A Bayesian RM-ANOVA revealed extreme evidence for an effect of our norm violation manipulation on perceived norm violation ($BF_{incl} = 1.99 \times 10^{11}$), and moderate evidence against an effect of our capacity violation manipulation on perceived norm violation ($BF_{incl} = 0.104$). As seen in Fig. 3 there was a clear increase in perceived wrongness between our low ($M = 4.732$, $SD = 1.982$), medium ($M = 5.381$, $SD = 1.950$), and high ($M = 6.546$, $SD = 1.155$) norm violation conditions, although participants generally perceived all requests as wrong.

Figure 3: Participant ratings of wrongness under low, medium and high level of norm violation. Error bars shows 95% Confidence Intervals.

Similarly, a Bayesian RM-ANOVA revealed extreme evidence for an effect of our capacity violation manipulation on perceived capacity violation ($BF_{incl} = 7.48 \times 10^{11}$), and moderate evidence against an effect of our norm violation manipulation on perceived capacity violation ($BF_{incl} = 0.125$). As seen in Fig. 4 there was a clear decrease in perceived feasibility between our low ($M = 4.227$, $SD = 2.134$), medium ($M = 3.082$, $SD = 2.178$), and high ($M = 2.021$, $SD = 2.437$) capacity violation conditions, although participants generally perceived all requests as questionably feasible at best. These effects suggest that our conditions influenced participants' perceptions in the intended manner without cross-effects.

Response Choice

Our Bayesian Contingency Table Analysis of the impact of condition on response choice gave us insight into the participants' preferences for the inclusion of different types of command rejection information. This analysis revealed extreme evidence against an effect of either our norm violation manipulation ($BF = 1.393 \times 10^{-4}$) or our capacity violation manipulation ($BF = 1.294 \times 10^{-4}$) on preferred response. As seen in Tabs. **??** and **??**, participants were, regardless of norm or capacity violation manipulation, nearly equally split between (a) preferring inclusion of norm information alone, or (b) preferring inclusion of both norm and capacity information, and with very few participants ever preferring (c) capacity information alone or (d) inclusion of neither norm nor capacity information. These results align with two hypotheses: the preference for both norm and capacity supports the Maximal Informativity Hypothesis (H1), while the preference for only norm information supports the Primacy of Normativity Hypothesis (H3).

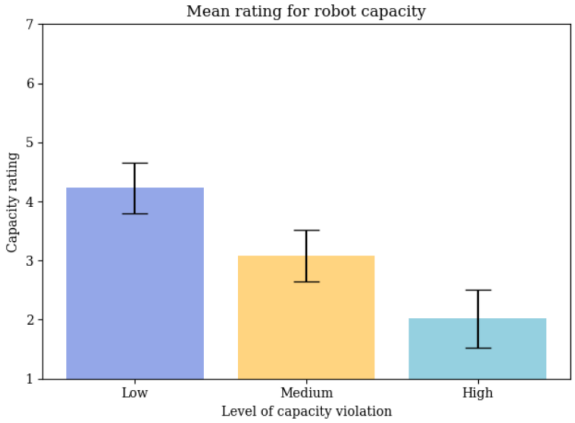

Figure 4: Participant ratings of robot capacity under low, medium and high level of capacity violation. Error bars shows 95% Confidence Intervals.

	Response Choice				
Capacity Violation	$\emptyset$	M	C	$\{M, C\}$	Total
Low	8	47	5	37	97
$Medium$	7	40	3	47	97
$High$	5	45	5	42	97
Total	20	132	13	126	291

Table 3: Preferred Response by Capacity Violation

	Response Choice				
Norm Violation	$\emptyset$	M	C	$\{M, C\}$	Total
Low	8	42	5	42	97
$Medium$	7	42	6	42	97
$High$	5	48	2	42	97
Total	20	132	13	126	291

Table 4: Preferred Response by Norm Violation

Qualitative Results

A single author performed an inductive thematic analysis (Braun and Clarke 2006) of participants' free responses as to why they preferred certain robot responses. We present the results of this semantic thematic analysis for the responses associated with *norm-and-capacity based rejection* and *norm-based rejection* to help illustrate the reasons behind participants' preferences for these two types of responses from a realist perspective.

Norm-based rejection Three key themes were identified in the responses of participants who preferred rejections that only included norm information.

Participants wanted robots to identify wrongdoing and help shape norms: Participants that chose the norm-based rejection point out the importance for the robot to clearly call out the unethical command. Participant 13 explained: *"the critical reason that the robot should not perform the requested action is that it would be wrong"* and Participant 18 pointed out: *"[The norm-based rejection] clearly explained that it would be wrong and that is all that is needed to be explained"*

Participants desired human-like responses: Some participants also felt the norm-based rejection, though lacking in detail, enabled the robot to show compassion and felt more human-like. Participant 46 expressed: *"[The norm-based rejection] is sensible enough and has human thought of emotion for another person"* and participant 10 said: *"[The norm-based rejection] is showing the most human emotion"*.

Participants disliked inclusion of capacity information: Some participants also explicitly argued that capacity information ought not be included since it's unnecessary to reject the request. Participant 65 explained: *"[The capacity information] could be relevant in some situations but it isn't here"* and Participant 83 said: *"Its ability to throw things should not be the decisive factor on whether it throws liquid on unsuspecting persons"*.

Norm-and-capacity-based rejection Three key themes were identified in the responses of participants who preferred rejections that included both norm information and capacity information.

Participants preferred the most detailed explanation: Participants that preferred responses including both norm and capacity information felt it was the most detailed and clear robot response, showing a respect for the human request and rejecting it firmly. Participants also felt that such responses appealed to both logical and emotional reasoning. For instance, Participant 76 highlighted: "The robot should fully explain its actions and reasoning at all times" while Participant 86 pointed out: *"[The norm-and-capacity-based rejection] would allow people to be reasoned with for both of their 'sides' of their brain"*.

Participants wanted to ensure human interactants would understand robots' capacities: Responses that include capacity information can also help establish a more accurate understanding. Participant 76 said: *"While ultimately the normative reason is more important, having the robot share its full thought process ensures better oversight of the robot and clarity of the robot's capabilities and inhibitions or restrictions."*. Participant 85 agreed: *"knowing the robot cannot throw things accurately is also important as a user might ask a robot to throw something away."*

Participants felt that capacity information could help deter future unethical commands: Some participants considered potential future unethical commands, and chose the norm-and-capacity-based rejection because they believed it could help deter future unethical requests. Participant 27 expressed:*"The robot should also express the lack of physical capability to perform this action so that the request will likely not be made again"*. Participant 44 echoed: *"If the robot adds that it is not possible for it to even do this, then this would satisfy the person and they would cease insisting."*

Discussion

Overall, our results are split in supporting the **Maximal Information Hypothesis (H1)** and the **Primacy of Normativity Hypothesis (H3)**. Participants near-uniformly desired inclusion of norm information; however, they were divided on whether to include capacity information as well. Our qualitative results shed light on the underlying reasons for these preferences. While norm information was broadly favored, the inclusion of capacity information was disliked by many. As expected, this was often due to a feeling of weakened moral information, with capacity explanations undercutting the included norm information. On the other hand, capacity information was sometimes preferred alongside norm information. In some cases, this was due to a desire for more information, but in others, a belief that the inclusion of capacity information would accentuate the effectiveness of the included norm information.

Design Implications

Our results suggests clear implications for robot design, as robot designers must account for norm and capacity information differently, in at least three ways. First, robot designers should enable differentiation of norm versus capacity constraints with potential distinct knowledge representation and reasoning methods for each type of constraint. Second, robot designers must design robot architectures to leverage these two types of constraints in different ways when generating command rejections. Finally, robot designers must make sure from an overall robot design perspective to al-

ways include norm information in command rejections, and to only selectively include capacity information.

Theoretical Implications

Given that different pieces of prior work helped to predict each of our original experimental hypotheses, our results help to bolster – or call into question – those prior results. Because our results (partially) support the **Primacy of Normativity Hypothesis (H3)**, they thus provide further support for prior work that has argued for the need to communicate norm information and engage in norm violation response (Zhu et al. 2020; Briggs et al. 2022; Winkle et al. 2022, 2021; Zhu et al. 2020; Jackson and Williams 2019a). These results also partially support prior arguments that when a single cause is used in an explanation, intention-laden causes are preferred (Hilton and John 2007). Similarly, because our results (partially) support the *Maximal Informativity Hypothesis (H1)*, they thus provide further justification for prior technical approaches to robot explanation that have generated all possible causes as part of their explanations (Briggs and Scheutz 2015; Briggs et al. 2022).

In contrast, because our results refute the **Relative Informativity Hypothesis (H2)**, they thus call into question the generality of previous findings across both human-human and human-machine interaction that simpler explanations, or those with fewer included causes, are uniformly preferred (Lombrozo 2006; Hilton 1996; Hilton and John 2007; Woodward 2006; Mota and Sridharan 2021), and similarly call into question the approach of generating the single strongest explanation that is typically taken in both the robotic norm violation response and failure explanation literatures. Finally, because our results refute the **Primacy of Capacity Hypothesis (H4)**, our results suggest that previous findings that reasons for failure need to be prioritized over reasons for disobedience may be misaligned with user preferences (Han, Phillips, and Yanco 2021).

Limitations

While we have identified key implications for HRI design and theory, our interpretations are tempered by several limitations. First, while in this work we clearly conveyed the robot's capabilities to our participants prior to the experiment to ensure they interpreted our experimental stimuli in a universal manner, this also meant that capacity-driven explanations might not have served as "model reconciliation devices." In essence, participants were evaluating the quality of explanations they did not necessarily require. Second, while manipulation check results indicated that our experimental conditions influenced participants' perceptions as intended, participants generally considered all requests as morally wrong. Further investigation may be needed into scenarios involving lower violation severity or less universally accepted norms. Third, the specific language used in different explanations may have influenced participants' perceptions in different ways. For instance, phrases like "this may cause harm" or "50% chance of failure" could be interpreted differently or vary in perceived naturalness. Finally, people's preference for detail in explanations may vary based on the task at hand. Participants may favor more information in a controlled video-based experiment where time pressure is minimal, but may prefer more concise explanations in the presence of time pressure (Smith et al. 2022).

Future Work

Based on these implications and limitations, we suggest several directions for future work. First, while we identified a split in user preferences, future work is needed to determine whether that split is due to individual differences in beliefs about robots and their capabilities, valence of attitudes toward robots, personality, social/political orientation, or some other factor. Second, future work is needed to explore whether the effects found in this work replicate with lower severity violations and less universally accepted norms. Third, future work is needed to understand how the strategies explored in this work might lead to different levels and types of human-robot trust (cp. (Malle and Ullman 2021; Briggs and Wasylyshyn 2025)). Fourth, future work grounded in linguistics is needed to further understand how specific word choices may impact the effectiveness and implications of norm-violation responses (i.e., instead of using the word "wrong", the robot may use alternative words such as "unethical") Finally, technical research is needed to understand the representation, reasoning, and architectural mechanisms to enable the types of norm violation responses recommended by this work.

Conclusion

In this work, we presented the results of a human subjects study exploring how the inclusion of normative vs non-normative information influences norm-violation response content selection in robots. Specifically, we considered scenarios where robots received requests that were simultaneously violating both norms and capacity constraints, systematically varying the degrees of violation in these human requests. We then measured the effects of these variations on participants' response selections.

Our results suggest that robots should consistently incorporate norm related information when rejection commands. However, participants were split on the importance of including capacity information. While some may view the inclusion of this additional information as a potential way to undercut communications of norm information, others appreciated the extra detail, which helped in further deterring unethical commands. Overall, these findings underscore the need for roboticists to reason about norm and capacity information in fundamentally different ways.

References

Admoni, H.; Weng, T.; Hayes, B.; and Scassellati, B. 2016. Robot nonverbal behavior improves task performance in difficult collaborations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 51–58. IEEE.

Arnold, T.; Kasenberg, D.; and Scheutz, M. 2021. Explaining in time: Meeting interactive standards of explanation for

robotic systems. *ACM Transactions on Human-Robot Interaction (THRI)*, 10(3): 1–23.

Banks, J. 2021. Good robots, bad robots: Morally valenced behavior effects on perceived mind, morality, and trust. *International Journal of Social Robotics*, 13(8): 2021–2038.

Baumann, A.-E.; Goldman, E. J.; Meltzer, A.; and Poulin-Dubois, D. 2023. People do not always know best: Preschoolers' trust in social robots. *Journal of Cognition and Development*, 24(4): 535–562.

Beebe, L. M.; Takahashi, T.; and Uliss-Weltz, R. 1990. Pragmatic transfer in ESL refusals. *Developing communicative competence in a second language*, 5573.

Belpaeme, T.; Baxter, P.; Read, R.; Wood, R.; Cuayáhuitl, H.; Kiefer, B.; Racioppa, S.; Kruijff-Korbayová, I.; Athanasopoulos, G.; Enescu, V.; et al. 2012. Multimodal child-robot interaction: Building social bonds. *Journal of Human-Robot Interaction*, 1(2): 33–53.

Berger, J. O.; and Sellke, T. 1987. Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American statistical Association*, 82(397): 112–122.

Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2): 77–101.

Breazeal, C.; and Scassellati, B. 1999. How to build robots that make friends and influence people. In *Proceedings 1999 IEEE/RSJ international conference on intelligent robots and systems. Human and environment friendly robots with high intelligence and emotional quotients (cat. No. 99CH36289)*, volume 2, 858–863. IEEE.

Briggs, G.; and Scheutz, M. 2014. How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress. *International Journal of Social Robotics*, 6: 343–355.

Briggs, G.; and Scheutz, M. 2017. The case for robot disobedience. *Scientific American*, 316(1): 44–47.

Briggs, G.; and Wasylyshyn, C. 2025. Trusting a Disobedient Robot: Rejecting a Command for Constructive Reasons Improves Evaluations of Trust. In *Proceedings of the 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 1255–1259. IEEE.

Briggs, G.; Williams, T.; Jackson, R. B.; and Scheutz, M. 2022. Why and how robots should say 'no'. *International Journal of Social Robotics*, 14(2): 323–339.

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

Brown, P.; and Levinson, S. C. 1987. *Politeness: Some universals in language usage*. 4. Cambridge university press.

Cameron, D.; de Saille, S.; Collins, E. C.; Aitken, J. M.; Cheung, H.; Chua, A.; Loh, E. J.; and Law, J. 2021. The effect of social-cognitive recovery strategies on likability, capability and trust in social robots. *Computers in human behavior*, 114: 106561.

Chakraborti, T.; Sreedharan, S.; Zhang, Y.; and Kambhampati, S. 2017. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. In *26th International Joint Conference on Artificial Intelligence, IJCAI 2017*, 156–163. International Joint Conferences on Artificial Intelligence.

Clark, H. H.; and Fischer, K. 2023. Social robots as depictions of social agents. *Behavioral and Brain Sciences*, 46: e21.

Colaco, Z.; and Sridharan, M. 2015. What happened and why? A mixed architecture for planning and explanation generation in robotics. In *Australasian Conference on Robotics and Automation (ACRA), Canberra, Australia*.

Coman, A.; and Aha, D. W. 2018. AI rebel agents. *AI magazine*, 39(3): 16–26.

Das, D.; Banerjee, S.; and Chernova, S. 2021. Explainable ai for robot failures: Generating explanations that improve user assistance in fault recovery. In *Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction*, 351–360.

Desai, M.; Kaniarasu, P.; Medvedev, M.; Steinfeld, A.; and Yanco, H. 2013. Impact of robot failures and feedback on real-time trust. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 251–258. IEEE.

Diehl, M.; and Ramirez-Amaro, K. 2022. Why did i fail? a causal-based method to find explanations for robot failures. *IEEE Robotics and Automation Letters*, 7(4): 8925–8932.

Dole, L. 2017. *The influence of a robot's mere presence on human communication*. Ph.D. thesis, Stanford University.

Ehsan, U.; Tambwekar, P.; Chan, L.; Harrison, B.; and Riedl, M. O. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th international conference on intelligent user interfaces*, 263–274.

Emnett, C. Z.; Mott, T.; and Williams, T. 2024. Using Robot Social Agency Theory to Understand Robots' Linguistic Anthropomorphism. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 447–452.

Esterwood, C.; and Robert, L. P. 2021. Do you still trust me? human-robot trust repair strategies. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, 183–188. IEEE.

Gillet, S.; Vázquez, M.; Andrist, S.; Leite, I.; and Sebo, S. 2024. Interaction-shaping robotics: Robots that influence interactions between other agents. *ACM Transactions on Human-Robot Interaction*, 13(1): 1–23.

Goffman, E. 2017. *Interaction ritual: Essays in face-to-face behavior*. Routledge.

Grice, H. 1975. Logic and conversation. *Syntax and semantics*, 3.

Gupta, S.; Walker, M.; and Romano, D. M. 2007. Generating politeness in task based interaction: An evaluation of the effect of linguistic form and culture. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*, 57–64.

Han, Z.; Phillips, E.; and Yanco, H. A. 2021. The need for verbal robot explanations and how people would like a robot to explain itself. *ACM Transactions on Human-Robot Interaction (THRI)*, 10(4): 1–42.

Häring, M.; Kuchenbrandt, D.; and André, E. 2014. Would you like to play with me? how robots' group membership and task features influence human-robot interaction. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, 9–16.

Hilton, D. J. 1990. Conversational processes and causal explanation. *Psychological Bulletin*, 107(1): 65.

Hilton, D. J. 1996. Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Thinking & Reasoning*, 2(4): 273–308.

Hilton, D. J.; and John, L. M. 2007. The course of events: counterfactuals, causal sequences, and explanation. In *The psychology of counterfactual thinking*, 56–72. Routledge.

Hou, Y. T.-Y.; Cheon, E.; and Jung, M. F. 2024. Power in Human-Robot Interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 269–282.

Jackson, R. B.; Li, S.; Banisetty, S. B.; Siva, S.; Zhang, H.; Dantam, N.; and Williams, T. 2021. An integrated approach to context-sensitive moral cognition in robot cognitive architectures. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1911–1918. IEEE.

Jackson, R. B.; Wen, R.; and Williams, T. 2019. Tact in noncompliance: The need for pragmatically apt responses to unethical commands. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 499–505.

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

Jackson, R. B.; and Williams, T. 2019b. On Perceiving Robots as Social and Moral Agents. In *Proceedings of the 2019 HRI Workshop on the Dark Side of Human-Robot Interaction*.

Jackson, R. B.; and Williams, T. 2021. A Theory of Social Agency for Human-Robot Interaction. *Frontiers in Robotics and AI: Special Issue on Rising Stars in Human-Robot Interaction*.

Jackson, R. B.; and Williams, T. 2022. Enabling morally sensitive robotic clarification requests. *ACM Transactions on Human-Robot Interaction (THRI)*, 11(2): 1–18.

Jackson, R. B.; Williams, T.; and Smith, N. 2020. Exploring the role of gender in perceptions of robotic noncompliance. In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, 559–567.

Jacovi, A.; Bastings, J.; Gehrmann, S.; Goldberg, Y.; and Filippova, K. 2023. Diagnosing AI explanation methods with folk concepts of behavior. *Journal of Artificial Intelligence Research*, 78: 459–489.

Jarosz, A. F.; and Wiley, J. 2014. What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving*, 7(1): 2.

JASP Team. 2024. JASP (Version 0.19.0)[Computer software].

Jeffreys, H. 1948. *The theory of probability*. OuP Oxford.

Johnson, D. I. 2007. Politeness theory and conversational refusals: Associations between various types of face threat and perceived competence. *Western Journal of Communication*, 71(3): 196–215.

Kahn Jr, P. H.; Kanda, T.; Ishiguro, H.; Freier, N. G.; Severson, R. L.; Gill, B. T.; Ruckert, J. H.; and Shen, S. 2012. "Robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. *Developmental psychology*, 48(2): 303.

Kahn Jr, P. H.; Reichert, A. L.; Gary, H. E.; Kanda, T.; Ishiguro, H.; Shen, S.; Ruckert, J. H.; and Gill, B. 2011. The new ontological category hypothesis in human-robot interaction. In *Proceedings of the 6th international conference on Human-robot interaction*, 159–160.

Kahn Jr, P. H.; and Shen, S. 2017. NOC NOC, who's there? A new ontological category (NOC) for social robots. *New perspectives on human development*, 106–122.

Kamino, W.; Jung, M. F.; and Sabanović, S. 2024. Constructing a Social Life with Robots: Shifting Away From Design Patterns Towards Interaction Ritual Chains. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 343–351.

Kaptein, F.; Broekens, J.; Hindriks, K.; and Neerincx, M. 2017. Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 676–682. IEEE.

Kasenberg, D.; Roque, A.; Thielstrom, R.; Chita-Tegmark, M.; and Scheutz, M. 2019. Generating justifications for norm-related agent decisions. In *proceedings of the 12th international conference on natural language generation*.

Kelter, R. 2020. Bayesian alternatives to null hypothesis significance testing in biomedical research: a non-technical introduction to Bayesian inference with JASP. *BMC Medical Research Methodology*, 20: 1–12.

Kennedy, J.; Baxter, P.; and Belpaeme, T. 2017. Nonverbal immediacy as a characterisation of social behaviour for human–robot interaction. *International Journal of Social Robotics*, 9: 109–128.

Khanna, P.; Yadollahi, E.; Björkman, M.; Leite, I.; and Smith, C. 2023. Effects of Explanation Strategies to Resolve Failures in Human-Robot Collaboration. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 1829–1836. IEEE.

Kwon, M.; Jung, M. F.; and Knepper, R. A. 2016. Human expectations of social robots. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*, 463–464. IEEE.

Lee, M. D.; and Wagenmakers, E.-J. 2014. *Bayesian cognitive modeling: A practical course*. Cambridge university press.

Lee, M. K.; Kiesler, S.; Forlizzi, J.; and Rybski, P. 2012. Ripple effects of an embedded social agent: a field study of a social robot in the workplace. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 695–704.

LeMasurier, G.; Gautam, A.; Han, Z.; Crandall, J. W.; and Yanco, H. A. 2024. Reactive or proactive? how robots should explain failures. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 413–422.

Lingard, L. 2012. Rethinking competence in the context of teamwork. *The question of competence: Reconsidering medical education in the twenty-first century*, 42–69.

Lockshin, J.; and Williams, T. 2020. "We need to start thinking ahead"""": the impact of social context on linguistic norm adherence. In *Annual meeting of the cognitive science society*.

Lombrozo, T. 2006. The structure and function of explanations. *Trends in cognitive sciences*, 10(10): 464–470.

Love, J.; Selker, R.; Marsman, M.; Jamil, T.; Dropmann, D.; Verhagen, J.; Ly, A.; Gronau, Q. F.; Šmíra, M.; Epskamp, S.; et al. 2019. JASP: Graphical statistical software for common statistical designs. *Journal of Statistical Software*, 88: 1–17.

Malle, B. F.; and Phillips, E. 2023. A Robot's Justifications, but not Explanations, Mitigate People's Moral Criticism and Preserve Their Trust.

Malle, B. F.; and Ullman, D. 2021. A multidimensional conception and measure of human-robot trust. In *Trust in human-robot interaction*, 3–25. Elsevier.

Meadows, B.; Sridharan, M.; and Colaco, Z. 2016. Towards an explanation generation system for robots: Analysis and recommendations. *Robotics*, 5(4): 21.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267: 1–38.

Mills, S. 2005. Gender and impoliteness. *Journal of Politeness Research*.

Mota, T.; and Sridharan, M. 2021. Answer me this: constructing disambiguation queries for explanation generation in robotics. In *2021 IEEE International Conference on Development and Learning (ICDL)*, 1–8. IEEE.

Mott, T.; Fanganello, A.; and Williams, T. 2024. What a Thing to Say! Which Linguistic Politeness Strategies Should Robots Use in Noncompliance Interactions? In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 501–510.

Mott, T.; and Williams, T. 2023. Confrontation and Cultivation: Understanding Perspectives on Robot Responses to Norm Violations. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2336–2343. IEEE.

Rezaei Khavas, Z.; Kotturu, M. R.; Ahmadzadeh, S. R.; and Robinette, P. 2024. Do Humans Trust Robots that Violate Moral Trust? *ACM Transactions on Human-Robot Interaction*, 13(2): 1–30.

Šabanović, S. 2010. Robots in society, society in robots: Mutual shaping of society and technology as a framework for social robot design. *International Journal of Social Robotics*, 2(4): 439–450.

Salem, M.; Lakatos, G.; Amirabdollahian, F.; and Dautenhahn, K. 2015. Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, 141–148.

Salem, M.; Ziadee, M.; and Sakr, M. 2014. Marhaba, how may I help you? Effects of politeness and culture on robot acceptance and anthropomorphization. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, 74–81.

Schmidtke, J. M.; and Cummings, A. 2017. The effects of virtualness on teamwork behavioral components: The role of shared mental models. *Human Resource Management Review*, 27(4): 660–677.

Schoeffer, M., Kuehl. 2022. "There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making. In *Proceedings of the 2022 ACM International Conference on Fairness, Accountability, and Transparency*.

Shen, S.; Slovak, P.; and Jung, M. F. 2018. "Stop. I See a Conflict Happening." A Robot Mediator for Young Children's Interpersonal Conflict Resolution. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*, 69–77.

Siegel, M.; Breazeal, C.; and Norton, M. I. 2009. Persuasive robotics: The influence of robot gender on human behavior. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2563–2568. IEEE.

Simmons, J. P.; Nelson, L. D.; and Simonsohn, U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11): 1359–1366.

Smith, C.; Gorgemans, C.; Wen, R.; Elbeleidy, S.; Roy, S.; and Williams, T. 2022. Leveraging intentional factors and task context to predict linguistic norm adherence. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Srinivasan, V.; and Takayama, L. 2016. Help me please: Robot politeness strategies for soliciting help from humans. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, 4945–4955.

Sterne, J. A.; and Smith, G. D. 2001. Sifting the evidence—what's wrong with significance tests? *Physical therapy*, 81(8): 1464–1469.

Ullrich, D.; Butz, A.; and Diefenbach, S. 2021. The development of overtrust: An empirical simulation and psychological analysis in the context of human–robot interaction. *Frontiers in Robotics and AI*, 8: 554578.

van den Bergh, D.; Van Doorn, J.; Marsman, M.; Draws, T.; Van Kesteren, E.-J.; Derks, K.; Dablander, F.; Gronau, Q. F.;

Kucharskỳ, Š.; Gupta, A. R. K. N.; et al. 2020. A tutorial on conducting and interpreting a Bayesian ANOVA in JASP. *L'Année psychologique*, 120(1): 73–96.

Van Doorn, J.; Van Den Bergh, D.; Böhm, U.; Dablander, F.; Derks, K.; Draws, T.; Etz, A.; Evans, N. J.; Gronau, Q. F.; Haaf, J. M.; et al. 2021. The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, 28: 813–826.

van Zyl, C. J. 2018. Frequentist and Bayesian inference: A conceptual primer. *New Ideas in Psychology*, 51: 44–49.

Visser, I.; Kucharskỳ, Š.; Levelt, C.; Stefan, A. M.; Wagenmakers, E.-J.; and Oakes, L. 2024. Bayesian sample size planning for developmental studies. *Infant and Child Development*, 33(1): e2412.

Wachowiak, L.; Fenn, A.; Kamran, H.; Coles, A.; Celiktutan, O.; and Canal, G. 2024. When Do People Want an Explanation from a Robot? In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 752–761.

Wagenmakers, E.-J. 2007. A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review*, 14(5): 779–804.

Wen, R.; Han, Z.; and Williams, T. 2022. Teacher, teammate, subordinate, friend: Generating norm violation responses grounded in role-based relational norms. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 353–362. IEEE.

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

Wen, R.; and Williams, T. 2022. Hidden Complexities in the Computational Modeling of Proportionality for Robotic Norm Violation Response. In *Proceedings of the AAAI Fall Symposium on Artificial Intelligence for Human-Robot Interaction (AI-HRI)*.

Williams, T. 2024. Understanding Roboticists' Power through Matrix Guided Technology Power Analysis. In *Companion Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (alt.HRI)*.

Williams, T.; Jackson, R. B.; and Lockshin, J. 2018. A bayesian analysis of moral norm malleability during clarification dialogues. In *CogSci*.

Winkle, K.; Jackson, R. B.; Melsión, G. I.; Brščić, D.; Leite, I.; and Williams, T. 2022. Norm-breaking responses to sexist abuse: a cross-cultural human robot interaction study. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 120–129. IEEE.

Winkle, K.; Melsión, G. I.; McMillan, D.; and Leite, I. 2021. 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*, 29–37.

Woodward, J. 2006. Sensitive and insensitive causation. *The Philosophical Review*, 115(1): 1–50.

Woodward, N.; Nguyen, T.; Zhu, L.; Fowler, C.; Kim, T.; Near, S.; Thoemmes, S.; and Williams, T. 2020. Exploring interaction design considerations for trustworthy language-capable robotic wheelchairs in virtual reality. In *International workshop on virtual, augmented, and mixed reality for human-robot interaction*, volume 3.

Zellner, S. 1980. Posterior odds ratios for selected regression hypotheses. *Trabajos de Estadistica y de Investigacion Operativa*, 31.

Zhu, L.; and Williams, T. 2020. Effects of proactive explanations by robots on human-robot trust. In *International Conference on Social Robotics*, 85–95. Springer.

Zhu, Q.; Williams, T.; Jackson, B.; and Wen, R. 2020. Blame-laden moral rebukes and the morally competent robot: A Confucian ethical perspective. *Science and Engineering Ethics*, 26(5): 2511–2526.