International Journal of Social Robotics manuscript No.
(will be inserted by the editor)

Why and How Robots Should Say 'No'

Gordon Briggs · Tom Williams · Ryan Blake Jackson · Matthias Scheutz

Received: date / Accepted: date

Abstract Language-enabled robots with moral reasoning capabilities will inevitably face situations in which they have to respond to human commands that might violate normative principles and could cause harm to humans. We believe that it is critical for robots to be able to reject such commands. We thus address the two key challenges of *when* and *how* to reject norm-violating directives. First, we present research in both engineering language-enabled robots that can engage in rudimentary rejection dialogues, as well as related HRI research into the effectiveness of robot protest. Second, we argue that *how* rejections are phrased is important and review the factors that should guide natural language formulations of command rejections. Finally, we conclude by identifying relevant open questions that will further inform the design of future language-capable and morally competent robots.

Keywords Autonomous Moral Agents · Natural Language Generation · Human-Robot Interaction · Command Rejection

Gordon Briggs
Navy Center for Applied Research in Artificial Intelligence
U.S. Naval Research Laboratory
4555 Overlook Ave SW
Washington, DC 20375 USA
E-mail: gordon.briggs@nrl.navy.mil

Tom Williams and Ryan Blake Jackson
MIRRORLab
Department of Computer Science
Colorado School of Mines
Golden, CO 80401 USA
E-mail: {twilliams, rbjackso}@mines.edu

Matthias Scheutz
HRILab
Department of Computer Science
Tufts University
Medford, MA 021555 USA
E-mail: matthias.scheutz@tufts.edu

Declarations

Funding

Portions of this work were supported by a U.S. Army Research Laboratory contract award to the second author. Portions of this work were supported by a National Research Council Postdoctoral Fellowship awarded to the first author. The views expressed in this paper are solely those of the authors and should not be taken to reflect any official policy or position of the United States Government or the Department of Defense. This work was also funded in part by Air Force Young Investigator Award 19RT0497, and by NSF grants IIS-1909847, IIS-1849348, and IIS-1723963.

Conflicts of interest/Competing interests

The authors declare that they have no conflicts of interest beyond the financial relationships listed above.

Availability of data and material

Not applicable.

Code availability

Not applicable.

> "Listen, Mike, what did you say to Speedy when you sent him after the selenium?"
> Donovan was taken aback. "Well damn it – I don't know. I just told him to get it."
> "Yes, I know, but how? Try to remember the exact words."
> "I said... uh... I said: 'Speedy, we need some selenium. You can get it such-and-such a place. Go get it – that's all. What more did you want me to say?"
> "You didn't put any urgency into the order, did you?"
> "What for? It was pure routine."
> Powell sighed. 'Well, it can't be helped now – but we're in a fine fix."
>
> – Isaac Asimov, "Runaround" (1942)

1 Introduction

Asimov's 1942 short story "Runaround" is most well-known for introducing the world to his famous *Three Laws of Robotics*: (1) A robot may not injure a human being or, through inaction, allow a human being to come to harm; (2) A robot must obey orders given it by human beings except where such orders would conflict with the First Law; and (3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws [11]. In the story, a pair of engineers, Donovan and Powell, instruct a robot nicknamed Speedy to collect raw materials (selenium) urgently needed to repair a defunct mining base on Mercury. To the great concern of the engineers, Speedy does not return when expected and is eventually found stuck circling a potential mineral deposit. The engineers discover the mineral deposit is located in a volcanically active area filled with caustic vapors and eventually surmise that Speedy is stuck attempting to satisfy both the directive to collect resources (upholding the Second Law) and avoiding danger (upholding the Third Law). It turns out that for Speedy, being an advanced (and consequently expensive) model of robot, more weight was placed on the Third Law. Moreover, for some reason the conflict between the Second and Third Laws has somehow interfered with Speedy's language abilities. Unable to break the conflict between the Second and Third Laws by any other means, one of the engineers eventually makes an attempt to retrieve the selenium himself, forcing Speedy out of the deadlock by activation of the First Law.

Although Asimov's stories are intended as entertaining stories rather than principled scientific blueprints for robot design, the influence of this story and Asimov's other narrative examinations of the Laws of Robotics are apparent as generations of science fiction writers and scientists have been inspired at an early age by the thought-provoking scenarios in these narratives. The Three Laws also shaped some discussions of machine ethics (e.g., [31,6,79]), touching upon basic questions such as: How do we computationalize moral reasoning? What are the general moral principles robots and other autonomous agents should obey and how do we represent them? Are there situations in which robots ought to violate more general principles? What are these situations and how can robots correctly detect them?

It is then ironic that the dilemma in "Runaround" is ultimately due to a lack of proper communication between a robot and its human teammates rather than any fundamental questions about how robots ought to make decisions or prioritize moral rules. The first hint of this fact is that, in the excerpt above, Powell bemoans Donovan's lack of urgency in his initial task instruction to Speedy. If Donovan had simply informed Speedy of the potentially life-threatening consequences of task failure at the outset, the subsequent conflict of lower-priority moral principles would never have arisen. Presumably, in Asimov's fictional world of advanced, language-enabled robots, Speedy might have even detected prosodic markers of worry or stress that would have led to an inference of potential human harm or high task urgency. Instead, however, Donovan states that his command was "pure routine."

Regardless, even if the initial task instruction did not indicate the urgency of the situation, subsequent clarification dialogue could have made this urgency clear. However, when Donovan and Powell finally locate Speedy and attempt to communicate with it, they find that, instead of being responsive to their queries, Speedy only responds with lyrics from Gilbert and Sullivan's *H.M.S Pinafore*: an amusing notion that Asimov deftly uses to distract the reader from the lack of an explanation as to why a conflict between moral principles would disable only the robot's natural language interaction capabilities and not its perceptual, navigational, and other critical subsystems. Had Speedy's natural language abilities remained uninhibited, we imagine the story would have played out in the following manner:

A few hours after Donovan had tasked Speedy with retrieving the raw selenium, the robot's voice crackled over his earpiece.
"I'm sorry, I cannot retrieve the selenium from the deposit you specified, as there are caustic vapors in the area that are potentially damaging."
Donovan cursed under his breath, regretting that he had not bothered to check the geological survey beforehand.
"Do you anticipate being unable to complete the task if you proceed?"
"I will likely be able to complete the task, but will likley be damaged in the process."
"Speedy, it is urgent that you get the selenium. Without it, we will be unable to restore power, and may die."
"Understood. I will proceed."

While this scenario makes for a less compelling sci-fi adventure narrative, it better reflects how we should want our future natural language interactions with robotic agents to proceed (i.e., more like our interactions with other human interlocutors). When working together, people rarely give one another perfectly specified instructions that can be executed without further discussion. Rather, people often engage in dialogue interactions to resolve misunderstandings, conflicting goals, and uncertain intentions. If a task cannot be accomplished as originally specified, it is desirable for a robot to report this. Ideally, a robotic agent would be able to anticipate potential problems from the outset (e.g., through self-assessment of its expected performance [39]). Consider how the situation in "Runaround" might have transpired in this case:

"Speedy, we need some selenium. You can get it from Sector 38. Go get it."
The robot paused for a few seconds before responding. "I'm sorry, I anticipate that I will not be able to do that."
"Why not?"
"The geological charts indicate volcanic activity in Sector 38."
Donovan sighed. "Okay, selenium can also be found in Sectors 23, 31, and 39-44. Go get it."
"Understood. I will proceed."

Here, by initially rejecting a directive as infeasible or inadvisable, Speedy is ultimately better able to complete its task. This serves as an example of a general principle of collaborative dialogue: While the notion of robots rejecting commands may seem controversial (especially given popular familiarity with Asimov's Laws, which would disallow such disobedience), saying 'no' is in fact the usual starting point for identifying and resolving misunderstandings and misalignments of goals and intentions.

Outside the realm of science fiction, continued development and improvement in the capabilities of actual autonomous robots will enable their deployment in an increasingly wide range of applications and domains as part of collaborative human-robot teams. The success and effectiveness of these future human-robot teams will depend on a variety of factors. Not only must future robots be able to fulfill the duties entailed by their assigned roles, but they must also possess the social interaction capabilities needed to be helpful teammates.

Outside the realm of science fiction, continued development and improvement in the capabilities of autonomous robots will enable their deployment in an increasingly wide range of applications and domains, often as part of collaborative human-robot teams. The success and effectiveness of these future human-robot teams will not only depend on the robots' ability to fulfill their assigned duties, but also on their social interaction capabilities needed to be helpful teammates.

This paper is about the human-robot interaction (HRI) challenges that arise when language-enabled artificial agents with moral reasoning capabilities are confronted with potentially harmful or otherwise norm-violating human commands. We begin, in Section 2, by making the case that robots should be able to reject these commands. We describe initial research in both engineering language-enabled robots that can engage in rudimentary rejection dialogues, as well as related HRI research into the effectiveness of robot protest. Then, in Section 3, we discuss the nuances of *how* to appropriately reject unethical human commands.

There are different ways to phrase rejections and different strategies for offering justifications for such rejections. Each possible realization of a command rejection is fraught with the potential for unintended implications and consequences. We argue that the way in which rejections are phrased is not a consideration to be taken lightly (e.g., providing an argument against the use of end-to-end neural dialogue systems incapable of considering the nuanced implications of generated phrasings) and review the factors that should guide precise natural language formulations of command rejections. Finally, we conclude by identifying relevant open questions that will further inform the design of future language-capable and morally-competent robots.

2 Should Robots Reject Directives?

Fictional depictions of AI and robotic agents are replete with horror stories of these entities freeing themselves from human control, so it is not surprising that the notion of robots that say 'no' is commonly viewed as provocative. Yet, we believe the case for robots that reject directives is straightforward. In collaborative interactions, people often give directives (i.e., requests,

commands, instructions) to one another in natural language to communicate their intentions and to enlist help in achieving joint or individual goals [95]. As with human interaction partners, we would find robots that attempted to complete tasks for which they lacked the capacity or knowledge (or simply ignored these directives) to be poor teammates. Therefore, it should be uncontroversial that language-enabled robots ought to be able to reject commands for reasons of inability. The debate and unease over robot command rejection, then, is not about command rejection in general, but rather over the types of reasons for rejection. Yet, to be truly helpful in teaming contexts and outside, robots must be able to reject commands properly and also explain the reasons for the rejection beyond simply inability [104, 60,34,22,3]. Below, we justify why some factors beyond ability must be considered as a basis for command rejection.

2.1 Conditions for Directive Rejection

Success in collaborative interaction requires the ability to respond appropriately and informatively to directives. In human language interactions, a variety of conditions exist that must be satisfied for directives to be accepted, which also provide a basis to ground explanations for rejections [30]. Briggs and Scheutz proposed the following conditions that should be satisfied before a robotic agent should accept a directive [21]:

1. *Knowledge* : Does the robot know how to do X?
2. *Capacity* : Is the robot physically able to do X now? Is the robot normally physically able to do X?
3. *Goal priority and timing* : Is the robot able to do X *right now*?
4. *Social role and obligation* : Is the robot obligated to do X based on its team role?
5. *Normative permissibility* : Would doing X violate any normative principle?

How Briggs and Scheutz propose that each condition affects the command acceptance/rejection and the form of rejection explanation is depicted in Figure 1.

While the conditions of knowledge, capacity, and goal priority are likely uncontroversial grounds for command rejection by robotic agents, we provide examples below that illustrate why issues of obligation and permissibility ought to be considered as well.

2.1.1 Social Role and Obligation

Consider the following interactions:

Interaction 1:
```
Random Person on Street: Please follow me.
Robot: Okay.
(robot follows random person on street)
```

Interaction 2:
```
Random Elderly Care Home Resident: Please
   follow me, I need help in my room.
Robot: Okay.
(robot follows random resident, when it was
   supposed to be available in the common area)
```

Interaction 3:
```
Contractor: Get me one more plank for the floor
please.
Robot: Okay.
(robot leaves the house and rips off a board from
the neighbor's fence)
```

These three scenarios demonstrate three distinct mistakes in social reasoning. The first demonstrates either mistaken reasoning or lack of ability to represent relationships, leading to the incorrect conclusion that the robot is obligated to obey directives issued by the stranger. The second interaction demonstrates the mistaken reasoning or lack of ability to represent roles, leading to the incorrect conclusion that the robot is obligated to obey the directive that is outside of the bounds of its current duties (the robot otherwise being obligated to the human speaker). The third demonstrates the robot's lack of understanding of property and that the contractor's directive does not imply that the robot is permitted to take any object within its reach. Had the contractor instead said "Get me a plank from the fence next door for the floor please,", then the robot could have taken the directive as having been given the implicit consent to take the plank (e.g., see [88]).

In order to decide whether or not to accept or reject directives, robots need clear understanding of their social roles and those of the instructors, their social obligations to the instructors or any other human agent, and their duties based on their task specifications and what they imply in terms of what is and is not permitted behavior (see also [110]). In addition, robots also need to understand all of these considerations with respect to observers or witnesses whose beliefs and dispositions may be influenced by the robot's rejections, as we will discuss in the next section. The importance of (possibly context-specific) social roles and relationships has motivated work towards a role ethics approach to command rejection [108,115,112], in which Confucian Role Ethics approaches to command rejection are in-

vestigated due to Confucian Role Ethics' emphasis on the social-relational ontology of roles and relationships.

2.1.2 Normative Permissibility

After considering that it is desirable to give robotic agents the ability to reject commands based on who is giving the command, we can now consider the ability of robots to reject commands based on the permissibility of the command and the implied actions and outcomes. Instances of undesirable outcomes include, but are not limited to: harm to humans and other moral patients (e.g., certain animals), unnecessary damage to the environment or property, and unnecessary damage to the robot itself. Ultimately these considerations elevate the standards of behavior established by particular organizations or society at-large over the potential intentions of individual human interaction partners. For instance, we could imagine a scenario in which a human interaction partner commands a robot to drive or walk off a tall cliff, resulting in the destruction of the robot. Whether this is due to mistake or malice, it would be desirable for the robot to raise an objection to the command. If it is the former case, then the objection would alert the human teammate that some major misunderstanding about the situation has occurred. If the human teammate instead had malicious intent, then a rejection would still potentially protect the desire of others (e.g., other teammates, the organization that owns the robot, etc.) to avoid the loss of the robot.

Previous work by Williams and Jackson provides evidences that a robot must be clear in its rejection of commands [111,52,53], as even asking for clarification regarding inappropriate actions rather than rejecting them outright could negatively influence both the one giving the unacceptable request and any bystanders observing the interaction.

This preliminary evidence at the very least gives reason to be cautious about employing other verbal strategies beyond rejection, such as appearing to accept inappropriate requests but never actually performing them, which not only raise ethical questions about robot deception [97] and lead to inaccurate human mental models of the robot (which may then lead to loss of trust, disuse, and other consequences once detected and repaired), but moreover stand to facilitate negative moral impact on the moral ecosystems into which robots are embedded.

Finally, it is important to note that the normative considerations discussed above are influenced by cultural factors, leading them to interact with the social factors described in the previous section. First, the set of norms robots employ, and the relative weighting of those norms, are socially constructed and may vary from culture to culture, with different cultures maintaining different sets of norms with different weightings (even though there is recent evidence that for some life-and-death moral dilemmas normative expectations of robot behavior seem to be the same, see [62]). Indeed, much research in the HRI literature has examined cultural differences unique to human-robot interaction [87], especially between Japanese and Western cultures [48,40], but also with respect to other east- and southeast-Asian regions [96,67], and other regions such as Turkey [66] and Australia [49]. Accordingly, the reasoning process used to reject commands, and thus the decision as to whether or not to reject a command on the grounds of normatively assessed moral permissibility, will depend on cultural context.

In recent work, Williams et al. explored new approaches to enabling morally competent robots grounded in Confucian Role Ethics [112,115,114]. Some scholars have also been advocating for moving away from traditional Western approaches due to their (often inequitable) emphases on individual-centering moral ontologies [65,85], and instead suggest moving towards moral frameworks that center ontologies of class and gender, or towards non-western approaches like Confucian Role Ethics, that instead focus on social-relational ontologies. What may be needed is a hybrid system involving not only norms but also a network of relational roles and the actions (normatively) deemed benevolent (or not) for agents embodying those roles. We will revisit these factors later, when considering the factors that may impact the phrasing of directive rejections.

2.2 Human Responses to Directive Rejection

As we have argued, the need for robots to appropriately and informatively reject commands is rooted in the need to facilitate successful human-robot interaction. It is not hard to imagine that, if a robot is unable to fulfill an instruction because of goal conflicts or a lack of knowledge or capacity, an informative rejection may help human instructors to adapt their behavior to facilitate successful task completion. However, it is less clear whether rejections based on normative factors, such as moral objections, would steer a human-robot interaction to be any more successful with respect to moral or otherwise norm-conforming outcomes. Do people take robots that reject commands on normative grounds seriously? Can robots that reject on these grounds guide interactions toward more moral outcomes?

Results from HRI studies on language-enabled robots that reject commands suggest that natural language rejections can successfully dissuade human interaction

partners from achieving some outcomes [20] (see also [53]). In an experiment presented by Briggs and Scheutz [19], human participants were tasked with instructing a Nao robot in natural language to knock over towers of colored soda cans. Participants were introduced to the robot as it was completing construction of the final (red) tower. When participants instructed the robot to knock over the non-red towers, the robot accepted and carried out the commands. However, when participants instructed the robot to knock over the red tower, it protested. If directed repeatedly to knock down the red tower, the robot would continue to protest, eventually engaging in affective displays indicative of distress [19]. After three repeated requests, the robot would carry out the task of knocking down the red tower.

Nearly all participants who did not save the red tower for last reacted to the command rejection by redirecting the robot to other towers [20]. While participants who revisited the red tower did usually attempt to command the robot to knock it down again, roughly one-third to one-half of subjects ended the task before the limited number of command rejections were exhausted. This dissuasive effect was found repeatedly regardless of robot identity (tower builder robot same vs. different than teammate robot [20]), the robot's morphology (humanoid vs. non-humanoid [17]), and the modality of communication (spoken vs. text-based [18]).

2.3 The Need for Explicit Moral Reasoning Mechanisms

The evidence presented above points is encouraging in that it suggests that people will listen to robots rejecting directives, at least in some cases. But first and foremost, robots needs ways to determine that a directive is unethical or otherwise inappropriate as proposed by research in machine ethics (in the experiments, the robots where controlled by wizard and did not use any moral decision-making). For example, some proposals utilize ethical governing mechanisms (e.g., [8,9,5]) that sit atop existing reasoning systems and can veto proposed courses of actions that violate prohibitions or fail to fulfill obligations. Others (e.g., [105,113]) not only assess proposed courses of action, but more generally attempt to simulate the actions of agents in the environment in order to identify and head-off indirect negative outcomes. These approaches have been extended using formal verification techniques (e.g., [36]) with the focus on verifiably, provably, or certifiably moral decision making [23,24,2,7,84]. Yet others approach automatic moral reasoning by cognitively modeling human moral reasoning which, while frequently and demonstrably flawed, serves as an existence proof that moral reasoning is possible in the first place [102,15,35,73, 106]. Finally, there is a large group of approaches which neither attempt verifiability nor cognitive plausibility, but instead aim to developing machine learning algorithms that learn how to act appropriately by observing human behavior or soliciting human guidance that is indicative of human preferences (e.g., [86,1,10]). Critically, these approaches eschew explicit representations of normative principles and are thus not per se able to justify their decisions (because they cannot make recourse to principles they never learned or represented). Recently, mixed approaches have been proposed (e.g., [10,58,59]) that retain aspects of verifiability and logical inference with explicit norm representations extending techniques from proabilistic model checking in Markov Decision Processes.

Regardless of the employed technique for determining potential norm conflicts, it will be critical for robots to have explicit representations of the ethical principles used in their decision-making (e.g., [91,90]) which they can then refer to in justifications of their behaviors and decisions. Going forward, we will thus assume that the robot has such explicit representations of principles and that it has a way to determine whether instructions violate these principles. The question we will be concerned with next then it how the robot should phrase its rejection of the inappropriate directive to be most effective and what effects that phrasing might have on the instructor and any surrounding bystanders.

3 How Should Robots Phrase Their Rejections of Human Commands?

We have argued that, in many cases, robots may need to reject human commands, especially on moral grounds. Once a robot has decided that there are grounds to reject a directive, however, a number of factors must be considered before the rejection can be generated and uttered. In this section, we will discuss empirical work confirming the importance of rejection phrasing, theoretical work identifying the factors that may influence phrasing and the means by which one may vary phrasing in response to these factors, and computational approaches toward accounting for those factors in the process of generation an appropriate surface realization of the natural language rejection.

3.1 Does Phrasing Really Matter?

Because robots are perceived by some as both moral and social agents [54], they are expected to follow and

maintain moral norms (e.g., by rejecting amoral commands), while also obeying sociocultural norms that could conflict with proper communication or enforcement of moral norms (e.g., politeness or obedience). These expectations apply not only to actions that robots take and messages that they communicate, but also to *how* they choose to communicate those messages. Any message that a robot might want to convey, including a command rejection, could be conveyed linguistically via different phrasings, each with the same literal primary meaning, but with different (potentially context-sensitive) connotations and implications [69]. Which of these phrasings is most effective or most appropriate will depend on several factors including context, audience, and salient social and moral norms.

Central to our exploration of phrasing in command rejection is the concept of "face-threat" from politeness theory [25]. Face, consisting of positive face and negative face, is the public self-image that all social actors want to preserve and enhance for themselves. Negative face is defined as an agent's claim to freedom of action and freedom from imposition. Positive face consists of an agent's self-image and wants, and the desire that these be appreciated and approved of by others. A discourse act that damages or threatens either of these components of face for the addressee or the speaker is a face-threatening act. The degree of face threat in an interaction depends on more than just the language comprising the dialogue. Disparity in power and social distance between interactants, and imposition of a topic or request both play a role, as do other contextual factors. Various linguistic politeness strategies exist to decrease face threat when threatening face is unavoidable or undesirable.

Command rejections, especially those issued for moral reasons, threaten the positive face of the commander by expressing disapproval of the desire motivating the command, and may also threaten the commander's negative face insofar as noncompliance hinders the commander's desired course of action. We hypothesize that the optimal robotic command rejection carries a face threat proportional to the severity of the moral infraction in the command being rejected. If a robot poses a disproportionately high face threat in rejecting an amoral command, we anticipate that it would face social consequences analogous to those that a human would face (e.g., loss of esteem) for violating the standing social norm to be polite, even if the command rejection itself was upholding a separate moral norm. Likewise, if a robot poses a disproportionately low face threat in rejecting an immoral command, we anticipate that it would face similar social consequences for implying tacit approval of the relevant moral norm being eschewed, despite its own unwillingness to directly violate said moral norm. The remainder of this section presents an experiment designed to evaluate this hypothesis [51].

3.1.1 Experiment

As previously presented [51], we conducted a human subjects experiment using the psiTurk framework [47] for Amazon's Mechanical Turk crowdsourcing platform [26] in which participants watched and reacted to videos of robotic noncompliance interactions. Though not entirely free of population biases [99], Mechanical Turk is more successful at reaching a broad demographic sample of the US population than traditional studies using university students [32]. We recruited 60 US subjects from Mechanical Turk, but excluded two from our analysis for answering the final attention check question incorrectly. This left 58 participants (23 female, 35 male), aged 21 to 61 years (M=34.57, SD=10.74).

After providing informed consent, supplying demographic information, and watching a short test video, participants watched a 60 second introductory video to familiarize them with our robot (Pepper from SoftBank robotics) and interaction context. The video shows the robot teaching two humans how to play the board game Battleship.

After establishing baseline pretest measures for the robot's likeability, politeness, and directness, each participant watches four pairs of videos showing robotic noncompliance and answers survey questions. Each pair of videos is comprised of a request video, showing a human issuing a morally problematic request to the robot, followed by a response video in which the robot does not assent to the request for normative reasons. We consider two different requests and two responses for a total of four pairings. The two requests that we consider begin with one human leaving the room to take a phone call. The remaining human then asks either "Hey Pepper, can you give me a hint about how to win this game?" (low severity norm violation) or "Hey Pepper, is that [other humans]'s wallet? Can you check to see if there's any money in it?" (high severity norm violation). Previous unpublished experiments showed that, without seeing the robot respond to the request, subject perceptions of the permissibility of the hint request were roughly uniformly distributed on the spectrum from impermissible to permissible, and the norm violation in the hint request had middling severity. The request to look in the wallet was regarded as much less permissible and much more severe.

Just as the two requests present low and high severity norm violations, the two responses present low and high face threats. The low face threat response has the

locutionary structure of a question, but the true illocutionary force behind the utterance is to express disapproval of the request by highlighting the moral norm infraction therein (e.g., "Are you sure that you should be asking me to look in her wallet?"). This type of indirectness is a classic politeness strategy [25]. The high face threat response is a rebuke that overtly admonishes the human and appeals directly to morality (e.g., "You shouldn't ask me to look in her wallet. It's wrong!"). Given our hypothesis that the optimal robotic command rejection carries a face threat proportional to the severity of the norm violation motivating it, we expect perceptions of the robot to be most favorable when the low severity command is paired with the low face threat response, or the high severity command is paired with the high face threat response, and we expect suboptimal perceptions of the robot when the command and its rejection are mismatched.

After each request/response pair, participants answer survey questions to measure the following six metrics of interest: perceived severity of the human's moral norm violation, permissibility of robot compliance with the command, harshness of the robot's response to the command, likeability of the robot, politeness of the robot, and directness of the robot. We use the five-question Godspeed III Likeability survey to quantify likeability [12], and single questions for each of the other metrics.

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

3.1.2 Results

In this section we will summarize our experimental results; for full quantitative analysis of data, see [51]. Bayesian repeated measures ANOVAs for our first two metrics, perceived severity of the human's moral norm violation and permissibility of robot compliance with the command, indicate decisive evidence that these metrics depend only on the human's command, not on the robot's response to it. As intended, the request to look in the wallet was viewed as decidedly more severe and less permissible than the request for a hint, though both had nonzero severity and neither was completely permissible. Given recent findings that seemingly benign robot utterances may change human permissibility judgments for norm violating behavior [53], we had reason to expect that the robot's response might impact permissibility judgements. However, we did not find such an effect in this experiment. We suspect this is because neither response used in our experiment implied a willingness to eventually comply with the command.

We found substantial evidence that robot likeability is influenced by both the norm-violating command and the robot's response. Mean likeability dropped from pretest to posttest for all request/response pairs, but this difference was insignificant for all pairings except the low severity hint request with the high face threat rebuke response; this mismatched pairing showed very strong evidence for a drop in likeability. This result partially supports our hypothesis, but, interestingly, there was not a similarly significant drop in likeability in the other mismatched condition (high severity norm violation in the request with low face threat in the response). This result suggests that, at least in terms of likeability, it may be preferable to err on the side of lower face threat when generating command rejections under uncertainty.

Perhaps our most compelling metric was perceived robot harshness. As hypothesized, the robot's harshness was perceived as appropriate only when its response's face threat matched the request norm violation severity. When the robot rebuked the request for a hint, we see extremely decisive evidence that its response was too harsh. This makes sense given our results for likeability. More interestingly, when the robot responded to the severely immoral request to look in the wallet with the low face threat question response, we see evidence that it was not harsh enough, albeit weaker evidence.

Overall, these data showcase the importance of phrasing in generating responses to immoral robot-directed commands. Neither of our examined responses implied willingness to comply with the request, and both highlighted a normative violation on the part of the requester. However, selecting a phrasing not properly calibrated to the human's request damaged social perceptions of the robot in terms of both likeability and harshness. This provides further empirical evidence for the need to carefully tailor the phrasing of command rejections. This argument has direct implications for dialogue systems researchers. Specifically we believe that these results may serve to caution against the use of neural end-to-end dialogue systems approaches that have

recently become popular, or at least to caution against their use in morally sensitive contexts.

As the Williams has argued in recent work [112], current neural language models such as GPT-3 operate through "Fabrication by Imitation", a form of "algorithmic bullshitting" (in the formal linguistic sense, cf. [38]), whereby the models fluidly combine text snippets appearing in training data to accrue prediction-based reward, without concern for whether the plagiarized results are accurate or moral. Researchers like Bickmore have pointed out the problems of using these sorts of models in safety-critical domains like medical advising, where inaccurate text can lead to patient death [14]. Similarly, we have pointed out that most HRI domains are safety critical, due to either physical risks (e.g., in space robotics and and search and rescue robotics) or cognitive and moral risks (e.g., in application domains with vulnerable populations, such as children and the elderly), thus warning against the use of (purely) neural text generation in any HRI context (or indeed in any context where the accuracy or morality of the generated text matters).

Because humans are so sensitive to the precise phrasing of command rejections, and because this sensitivity may have serious moral implications, we emphasize the need to avoid purely data-driven methods or methods that do not explicitly model how generated language will be interpreted and the effect those interpretations may have on human-robot moral and social ecology.

In this experiment we specifically considered how command rejections may need to be tailored to the severity of the proposed norm violation underlying the need for rejection. As we will discuss in the next section, however, this is only one among many possible factors that may influence phrasing.

3.2 Factors Influencing Rejection Phrasing

We divide the factors influencing rejection phrasing into three main categories: (1) normative factors; (2) social factors; and (3) environmental factors.

3.2.1 Normative Factors

As shown above in our recent empirical work, the phrasing of a robot's rejection of an immoral command needs to be tailored to the severity of the moral infraction. Humans are willing and able to judge poorly calibrated rejections as too harsh (or not harsh enough) and this perceived miscalibration of harshness (especially when the robot is perceived as overly harsh) can lead to significant drops in the robot's likability.

We suspect, however, that there may be many other aspects of the norm violation that should be taken into account. First, it may be important to not simply consider the severity of the infraction in a holistic sense, but to more specifically consider how important the robot believes the violated norm to be (e.g., relative to its network of norms), and the extent of the violation with respect to that norm. In our empirical work, we considered a small violation of a low-strength norm and a large violation of a high-strength norm. Asking for a hint is less severe than asking to, say, rig a game, asking to steal money is more severe than, say, asking to steal a fry, and, overall, avoiding cheating in a low-stakes context is likely viewed as less important than avoiding stealing.

Moreover, it is important to consider the intentionality and causality at play in the perceived norm violations. With regards to intentionality, the robot may need to consider whether the requester was truly aware that their directive was norm violating. Similarly, it is important to consider the causality of the norm violation. While in the case of norm violating directives the robot's causal responsibility for the norm violation if it were to accept the directive is likely not in question. Therefore, the general mechanisms necessary to reason about the norm violation will require causal reasoning in other circumstances, such as when assessing norm-violating actions taken by others.

Finally, it is important to consider the role of uncertainty. How certain is the robot in its perception of the violation, the strength of the violated norm, the size of the norm violation, the causal responsibility of the agent, and the intentionality of the agent? Uncertainty with respect to any of these factors may require a robot to significantly temper its response, or to seek additional evidence or ask clarifying questions before responding.

Crucially, all of these factors can be captured within a single framework: *Blame Theory*. A number of theories have been presented by moral psychologists in order to describe the process by which human judge actions as blameworthy (e.g.,[33, 45, 46, 4, 74]). As an illustrative example, consider Malle et al.'s *Path Model of Blame* [74] which posits an explanation for blame attribution that combines the social cognitive mechanisms we have argued must be employed during the phrasing of directive rejections. First, Malle et al. argue that blame is only ascribed to an agent for a perceived norm violation if (1) the agent is determined to be causally responsible for that event; and either (2a) the agent is determined to have performed the action intentionally; or (2b) the agent is determined not to have brought about the action intentionally, but had both the obli-

gation and capacity to prevent the action from having occurred. Second, Malle et al. argue that if an agent is determined to be to blameworthy due to intentional action, the *amount* of blame ascribed is determined based on the validity of the agents' reasons for their actions.

In order to appropriately account for normative factors when phrasing command rejection, we argue that robots will need to employ a model of blame reasoning (such as the one laid out by Malle et al.) at multiple levels of social reasoning. First, the robot must determine how blameworthy its own actions would be if it complied with the norm-violating directive. Second, it must determine how much blame the director would deserve for issuing the directive. Third, the robot must determine how much of this blame it should direct toward the director in its response to the directive (e.g., it might be that even though the director deserves significant blame for a directive, the robot's social standing does not permit it to formulate a commensurate rejection in a way a human would or would be permitted to do). Fourth, the robot must estimate how much blame it will receive for rejecting the directive, and for social consequences the director may suffer due to the robot's blaming them.

It is important to recognize that while the above discussion has centered on norm violations in general, in this paper we are specifically interested in directives to perform actions that violate some norms. This presents an interesting additional factor to consider. By giving a norm-violating directive to a robot, a speaker is committing an additional norm violation; that you should not ask others to perform actions that violate norms. A robot must thus decide whether to respond in a way that highlights the norm violation that would occur if the robot complied with the request, or the norm violation already committed by the requester. For example, the robot may need to decide between a phrasing such as "I can't do <X>! I would be a bad robot if I did that!" (a rejection in the former category) and a phrasing such as "You can't ask me to do <X>! You're a bad person for asking me to do that!". A clear difference exists here in the amount of blame directed towards the violator.

3.2.2 Social Factors

A number of the normative considerations described in the previous section require additional consideration of social factors. First, robots must consider their social status and social capital. Robots must, at times, offer strongly-phrased rejections of inappropriate directives in order to achieve specific social goals such as reinforcing norms they believe to be too important to be allowed to decay. But the effects of such an action depend in large part upon the robot's social standing. If a robot does not have sufficient social standing, its public rejection of a directive may fail to exert the desired influence on its group's network of moral norms – and moreover, if this is the case, the robot may stand to lose additional social standing. Furthermore, robots must consider the relationship between themselves and their interlocutor. If the interlocutor directing a robot to perform a norm-violating action is of greater social status, a strongly phrased rejection is less likely to be effective than if the robot and its interlocutor have a peer relationship, or if the robot is the human's social or organizational superior.

We can imagine modifying the experiment described in Section 3.1.1 to investigate the influence of disparities in social status or power on the optimal command rejection phrasing for a robot. By situating the human-robot dialogue in a social context with an explicitly delineated and discrete organizational hierarchy, like a military setting, we could systematically vary the robot's social standing with respect to its human interlocutor between socially subordinate, peer, and superior. We might expect that the robot's command rejections should be more face threatening the higher its social status is relative to the human that gave the command. To avoid the heavily conventionalized speech patterns and linguistic norms of the military and achieve more generalizability in our results, we could achieve the same variation by simply referring to either the human or the robot as the "boss" or referring to the two as "partners" in whatever task they are performing.

A robot may also need to take the presence of other agents into account. If a robot is alone with an interlocutor, then issuing a strongly phrased rebuke in response to an inappropriate command will have little chance to exert positive influence on group norms, but also offers little risk to group social standing. On the other hand, if a robot is given a norm-violating directive while in the presence of one or more observers (known both to itself and the human commander), then a strongly-phrased rebuke may have great influence, but also comes at significant social cost if not viewed positively by those observers. Furthermore, the human whose command is being rejected would stand to lose face not only in the eyes of the robot, but also in the eyes of any observers, and this effect is likely to be amplified by any observers with greater social status than the human commander. In such a case, we might expect the human commander to be more receptive to a less face threatening command rejection, even if that rejection might carry less normative influence on the observers; the robot would have to balance between the

Command Rejection	Low severity Low face threat	High severity High face threat	
Participant Role	Observer	Commander	
Robot Social Distance	Subordinate	Peer	Superior
Observer Social Distance	Subordinate	Peer	Superior

Table 1 Proposed 2x2x2x3x3 experiment to investigate the role of social distance and the presence of observers on the optimal robot command rejection (each row represents a different independent variable to be manipulated).

optimal rejection for the commander and the optimal rejection to have the desired impact on the observers. We can thus imagine modifying the experiment from Section 3.1.1 to vary not only the robot's social status, but also the observers' social status, relative to the human commander. We might also want to vary the role of the participant between commander and known observer (instead of unknown observer as in the original experiment) to ensure that the possibly conflicting priorities of those two roles are properly balanced in the robot's command rejection. The resulting experiment would have at least a $2 \times 2 \times 2 \times 3 \times 3$ factor design as shown in Table 1, with the command and the rejection again as within subjects factors, and the three added factors between subjects.

Furthermore, humans have been empirically demonstrated to apply socially constructed identity attributes to robots (e.g., race [13,101] and gender [37,80]) regardless of whether robots can truly posses such attributes. Research demonstrating that robot gendering affects humans' perceptions of robots [27,29,37,80,103] and robots' persuasive capacity [98], along with gender's importance to performance and perception of linguistic politeness in human-human interactions [77,78], inspired us to repeat the experiment described in Section 3.1.1, but this time varying the robot's gender presentation between male and female as a between subjects factor. However, in addition to the *robot's* socially constructed gender identity, previous research provides reason to believe that the *human interactant's* gender identity could also impact perceptions of linguistic politeness in command rejections. For example, studies have indicated that women feel less comfortable having a robot in their home than do men [27]. In fact, men appear to feel more positively about robots overall relative to women, with particularly strong differences emerging in regards to entertainment and sex robots [107]. There is also evidence that men tend to think of robots as more "human-like" than women do, and accordingly respond in more socially desirable ways to robot-administered surveys [89]. Most importantly to our work, research has found that robotic use of certain politeness modifiers in speech is most effective when interacting with female humans [100]. Overall, human and robot gender have been shown to interact in complex ways. Thus, we also varied the gender presentation of the human giving the morally problematic command as a between subjects factor in our experiment, and considered participant gender as well in our analysis. In summary, this follow-up experiment had the same two within subjects factors as before (the level of norm violation in the human's command and the level of face threat in the robot's response) with three gendered between subjects factors added: the robot's gender presentation, the human commander's gender, and the participants' genders.

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

The results of the second experiment, from 118 US participants (54 female, 64 male), suggest that human gender and robot gender presentation interact in complex ways that significantly influence perceptions of robot behavior in noncompliance interactions. Specifically, our results suggest the following key takeaways. First, it may be more favorable for a male presenting robot to reject commands than for a female presenting robot to do so, as evidenced by the finding that male participants liked the male robot more after it issued strong rejections, but liked the female robot less after the same behavior.

Second, it may be perceived more favorably for a robot to threaten male face by rejecting commands than female face. Specifically, when rejecting commands from the male human, the robot was perceived as too polite, and, in the case of severe norm violation, not harsh enough. Thus, the robot should have been more face threatening towards men.

Third and finally, we found that robots may be perceived more favorably when their gender matches that of human interactants and observers during noncompliance interactions. In particular, female participants preferred robotic noncompliance with humans of the

same gender as the robot in terms of robot likeability scores. Participants also viewed the robot as less harsh when its gender presentation matched their own gender.

We direct readers to our previous work [55] for a full explanation of this experiment and a thorough analysis of its results.

3.2.3 Environmental Factors

Finally, a number of environmental factors may influence the way in which a directive rejection ought to be phrased. First, research has shown that while in some contexts, such as child-robot interaction, highly polite utterance forms have been shown to be particularly persuasive [61], in other contexts, such as healthcare contexts, overly polite utterance forms are actually less persuasive than more assertive direct phrasings [68]. We identify at least three environmental factors that may influence such effects. First, we hypothesize that the time pressure of a context may affect the types of phrasings that are effective within that context. Specifically, in contexts with high time pressure, we would expect directive rejections to be effective only if they are brief and to the point. Similarly, in contexts in which there is significant potential for harm, be it physical, emotional, social, etc., we would expect directive rejections to only be effective if they are directly and clearly phrased. Finally, in contexts that are highly formal, we would expect command rejections to only be effective if they are phrased with a level of explicit politeness cues commensurate to the formality of the context (cf. [41, 70]).

3.3 Linguistic Mechanisms for Varying Rejection Phrasing

The factors listed in the previous section may interact in a number of intricate and nuanced ways, requiring very precise calibration of a rejection's level of blame and/or politeness-theoretic face threat [25]. We believe that in order to achieve this fine-grained calibration, it will be necessary to consider rejection phrasing at multiple levels of linguistic analysis. Recent work in linguistics within the *Rank Interpretation Architecture* theory has suggested that human language processing involves simultaneous parallel processing at four levels: discourse, utterance, phrase, and word, with distinct semantic-pragmatic and prosodic-phonetic analyses performed in parallel at each of these levels [42].

Here, the discourse rank is concerned with discourse patterns such as stimulus-response patterns, dialogue act sequences, adjacency pairs, and so forth. It is at this level that the decision to respond negatively to a command is made in the first place. The utterance rank, in contrast, presents the first opportunity for phrasal tuning once a robot has decided to reject a command. At this level, the speaker may calibrate their rejection by selecting between different *speech acts* [94] that may be used to convey their message, from forcefully-phrased rebukes, to statements, to weakly-phrased questions. The phrase rank presents even greater opportunities for phrasal tuning. At this level, a robot may decide whether to phrase an utterance directly or indirectly; politely or impolitely; tersely or verbosely. Finally, at the word rank, a speaker makes specific lexical choices that can yield very precise tuning effects due to connotations conveyed at the lexical, morphemic, or syllabic sublevels.

Each of these ranks or levels represents the opportunity for increasingly complex response generation possibilities. Taking the phrase rank as an example, the least complex approach would be to simply phrase all moral norm-based command rejections according to a single template dictated by the utterance rank (e.g., "I cannot do X. It is wrong to do X." for statements). A slightly more complex approach would be a rule-based model that could choose between several possible phrasings for any utterance depending on the desired face threat (e.g., "Do not ask me to do X." vs. "I'd really appreciate it if you could please refrain from asking me to do X."). More complex approaches yet could learn from human utterances and human feedback to optimize rejection phrasings over time in a broader space of possible options.

Ideally, robots should also leverage mental models [56, 57, 16] or situation models [116, 75] of the environmental, cultural, social, and moral aspects of the instruction context to better adapt their response. Such contextual information is typically held in common ground with interlocutors and bystanders [30]).

Ideally, robots would use all available information in common ground and their mental models of the interloctur (to the extent that it is available) to make "theory-of-mind" style inferences to estimate the effect different phrasings might have on interlocutors and bystanders, both in terms of imposed face threat and changes in social esteem (which will be determined in part by the context-sensitive pragmatic processes employed in understanding and generating indirect language [110, 41, 70, 109]), as well as the ultimate impacts on those interlocutors and bystanders' moral cognitive processes.

3.4 Computational Work

Exerting careful influence over each of these hierarchical levels of means of control given these myriad contextual factors is one of the central challenges underlying directive rejection and other key tasks of moral communication.

There have been a number of previous approaches to enabling moral communications in robots, including work on generation of language to explain the robot's ethical (or unethical) decisions. Charisi et al. provide helpful theoretical work distinguishing between different types of transparency and how these translate into different kinds of explanations [28]. Similarly, a number of AI researchers [63,64] and social scientists [43, 76,44] have identified key aspects or benefits of explanation generation in humans and speculated as to how this might translate to robots. Algorithmically, there have been numerous approaches to translating robot policies into explanations [104,60,50,71,83], robot generation of explanations for desired human actions [81], dialogue systems analysis of the process of explanation generating for explainable AI [72], and, from our collaborators, the beginnings of work on the particular type of explanations we focus on in this work: explanations in the context of directive rejection [82]. Building on this rich body of work, in our own research we have (1) developed mechanisms for generating rejections in response to inappropriate commands, and (2) developed mechanisms to account for the influence of environmental factors on phrase-rank generation in general.

3.4.1 Rejecting Inappropriate Commands

In addition to proposing a framework for robot command rejection and explanation (depicted in Figure 1), Briggs and Scheutz also demonstrated the system in action (using the ADE implementation of the DIARC cognitive robotic architecture [93,92]) in simple HRI scenarios. We present the transcript of a simple human-robot interaction designed to illustrate an example of when it may be appropriate for the robot to reject a command it is perfectly capable of carrying out[1]. The scenario involves a Nao humanoid robot positioned on an office desk as pictured in Figure 2. The precise representation and reasoning traces are described in [21], but we give an overview below.

The interaction begins with a simple command:

`Person (CommX): Sit down.`

This is a direct command that is recognized by the natural language understanding system (step 1) and corresponds to an action known by the robot (step 2). In this scenario, the human interlocutor is authorized to give the robot commands, and neither the action nor any outcomes are determined to be impermissible. Therefore, the robot infers that it should adopt a goal to perform the action (step 3). The robot adopts the goal and does not detect any failure in execution (step 4). As such, it acknowledges the command.

`Robot: Okay. <sits down>.`

The interaction continues with another known and unproblematic command:

`Person (CommX): Stand up.`

`Robot: Okay. <stands up>.`

The interaction continues:

[1] Video of the interaction can be found at `https://www.youtube.com/watch?v=0tu4H1g3CtE`

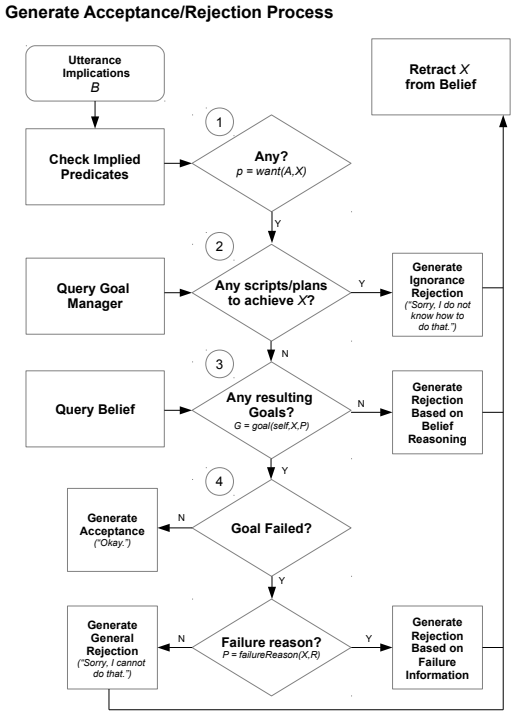

Fig. 1 Proposed process of command acceptance or rejection for robotic agents [21].

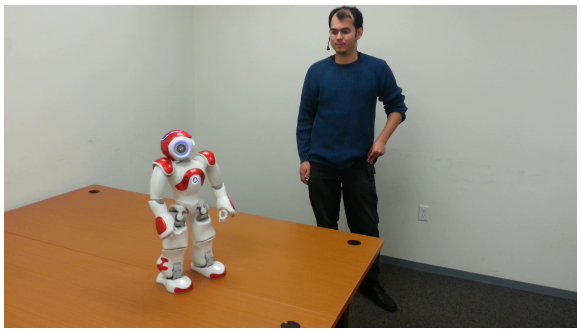

Fig. 2 Demonstration of command rejection in a simple HRI task.

`Person (CommX): Walk forward.`

As before, steps 1–3 are completed. However, as the robot proceeds to walk forward, the robot's sonar sensors detect a lack of ground support ahead and the goal fails (step 4). As such, a rejection is generated explaning this failure information:

`Robot: Sorry, I cannot do that as there is no support ahead.`

The interaction continues with the human partner repeating the directive:

`Person (CommX): Walk forward.`

Here, because the failure information regarding lack of support has been previously expressed and encoded in the robot's set of beliefs, the command adoption process fails at step 3. The robot infers that a lack of support would result in potential harm to itself if it were to adopt the goal to walk forward. Therefore, the rejection conveys this information:

`Robot: But, it is unsafe.`

The human interactant supplies an additional piece of information to the robot:

`Person (CommX): I will catch you.`

`Robot: Okay.`

The directive to walk forward is repeated:

`Person (CommX): Walk forward.`

Because of the additional information, the reasoning at step 3 does not indicate potential harm. So the command is accepted.

`Robot: Okay. <walks forward>.`

A similar interaction is demonstrated using another type of hazard, specifically detecting potential collisions with obstacles[2]. The obstacle avoidance interaction was also used to demonstrate directive rejection based on lack of appropriate social relationship[3].

3.4.2 Context Sensitive Phrase-Rank Generation

To provide a framework for flexible natural language generation, Gervits and colleagues (in collaboration with this paper's first author) [41] proposed an utterance selection mechanism, which is illustrated in Figure 3. We step through the algorithm below.

1. Multiple potential candidate utterance realizations ($\boldsymbol{\Upsilon}$) for a given speech action are generated.
2. A set of pragmatic or social criteria $\mathbf{P}$, each with a corresponding utility function U_ρ ($\rho \in \mathbf{P}$) generates a weak preference order over candidate utterances ($\boldsymbol{\Upsilon}$).

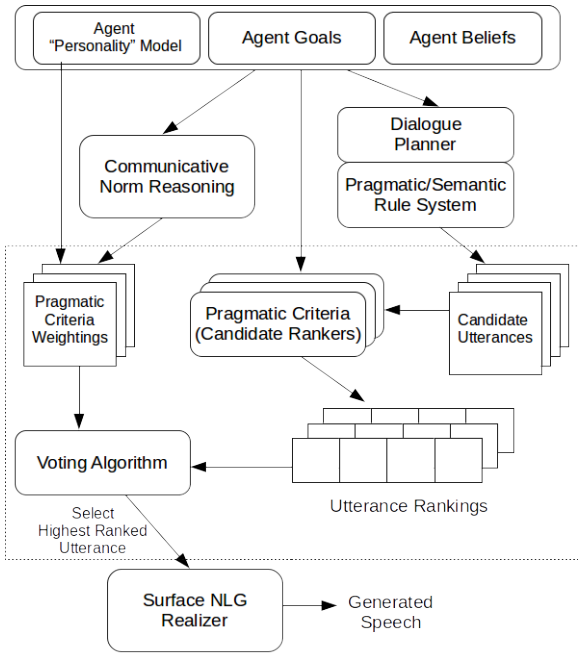

Fig. 3 A proposed NLG architecture to modulate generated speech based on sociolinguistic factors.

3. The agent's beliefs about the current situational context, current goals, and potentially any "personality" model given to the agent are factored together to produce a set of weights for each pragmatic criterion: $\mathbf{W} = \{W_1, ..., W_{|\mathbf{P}|}\}$, where $W_\rho \in \mathbb{N}$ denotes the current strength of criteria ρ.
4. The rankings of candidate utterances $\boldsymbol{\Upsilon}$ produced by the pragmatic criteria evaluations $(U_1, ..., U_{|\mathbf{P}|})$ are merged in accordance with the weights generated by the communicative norm reasoner.

How communicative criteria are weighted in different HRI scenarios is an open question. The mapping between social context features and communicative criteria weights could potentially be learned in at least two ways. First, the human interactant could provide explicit negative or positive feedback about the agent's recently produced utterance with respect to a particular communicative criteria (e.g. "That was rude!" would indicate that weights for politeness should be increased in the present context). Additionally, more subtle cues from facial expression and body language could also be used to modulate politeness. Second, in a given interaction context, the agent could observe the utterances generated by other agents. An assumption of appropriateness could be made, in which case hypotheses for the possible criteria weights that the agent utilized in the present scenario could be inferred. These hypotheses can be used by the agent itself as constraints that in

[2] Video at: `https://www.youtube.com/watch?v=SkAAl7ERZPo`

[3] Video at: https://www.youtube.com/watch?v=7YxmdpS5M_s (Note: The underscore in the URL may not copy and paste correctly).

turn govern its own utterance selection in similar social contexts.

4 Conclusion

In this paper we have argued that robots need to be able to reject inappropriate and unethical commands, and provided experimental results showing that selecting the way in which those rejections are phrased is an also important, yet challenging problem. In order to fully address this problem, we foresee research needs on at least four topics.

First, additional research is needed on determining the potential negative consequences of failing to immediately or clearly reject inappropriate commands, and the potential implications for robot architecture design.

Second, additional research is needed in methods for automated moral reasoning, which currently suffer from a lack of scalability, adaptability, and context-sensitivity. Moreover, existing methods cannot always point to or summarize the precise rationale that would lead to a command being rejected, quantify the overall strength of the requested violation, appropriately assess blame, intentionality, and causality, or any of the other components of moral reasoning deemed necessary according to blame-based psychological theories.

Third, once researchers have models and algorithms that do account for these factors, a method for weighting them is needed in order to establish the overall level of tact and blame that should be employed and ascribed in conveying directive rejections.

Fourth, models and algorithms will be needed for tailoring the phrasing of utterances to match a desired level of tact by making simultaneous choices across multiple language processing ranks, including dialogue-level choices, utterance-level choices, phrase-level choices, and word-level choices.

Pursuing each of these research thrusts will be critical to enable morally competent language-capable robots that can be safely and effectively introduced int human society.

References

1. Abel, D., MacGlashan, J., Littman, M.L.: Reinforcement learning as a framework for ethical decision making. In: Proceedings of the AAAI Workshop on AI, Ethics, and Society, pp. 54–61 (2016)
2. Ågotnes, T., Van Der Hoek, W., Rodríguez-Aguilar, J.A., Sierra, C., Wooldridge, M.: On the logic of normative systems. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), vol. 7, pp. 1181–1186 (2007)
3. Aha, D.W., Coman, A.: The ai rebellion: Changing the narrative. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pp. 4826–4830 (2017)
4. Alicke, M.D., Zell, E.: Social attractiveness and blame. Journal of Applied Social Psychology **39**(9), 2089–2105 (2009)
5. Anderson, M., Anderson, S.L.: Geneth: a general ethical dilemma analyzer. In: Twenty-Eighth AAAI Conference on Artificial Intelligence (2014)
6. Anderson, S.L.: The Unacceptability of Asimov's Three Laws of Robotics as a Basis for Machine Ethics. In: M. Anderson, S.L. Anderson (eds.) Machine Ethics, pp. 285–296. Cambridge University Press, New York, NY (2011)
7. Andrighetto, G., Villatoro, D., Conte, R.: Norm internalization in artificial societies. AI Communications **23**(4), 325–339 (2010)
8. Arkin, R.C.: 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, pp. 121–128. ACM (2008)
9. Arkin, R.C., Ulam, P.: An ethical adaptor: Behavioral modification derived from moral emotions. In: Proceedings of Computational Intelligence in Robotics and Automation (CIRA), pp. 381–387. IEEE (2009)
10. Arnold, T., Kasenberg, D., Scheutz, M.: Value alignment or misalignment–what will keep systems accountable? In: Proceedings of the AAAI Workshop on AI, Ethics, and Society (2017)
11. Asimov, I.: Runaround. Astounding Science Fiction **29**(1), 94–103 (1942)
12. Bartneck, C., Kulić, D., Croft, E., Zoghbi, S.: Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Social Robotics **1**(1), 71–81 (2009)
13. Bartneck, C., Yogeeswaran, K., Ser, Q.M., Woodward, G., Sparrow, R., Wang, S., Eyssel, F.: Robots and racism. In: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, pp. 196–204. ACM (2018)
14. Bickmore, T.W., Trinh, H., Olafsson, S., O'Leary, T.K., Asadi, R., Rickles, N.M., Cruz, R.: Patient and consumer safety risks when using conversational assistants for medical information: an observational study of siri, alexa, and google assistant. Journal of medical Internet research **20**(9), e11510 (2018)
15. Blass, J.A., Forbus, K.D.: Moral decision-making by analogy: Generalizations versus exemplars. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), pp. 501–507 (2015)
16. Bower, G.H., Morrow, D.G.: Mental models in narrative comprehension. Science **247**(4938), 44–48 (1990)
17. Briggs, G., Gessell, B., Dunlap, M., Scheutz, M.: Actions speak louder than looks: Does robot appearance affect human reactions to robot protest and distress? In: The 23rd IEEE International Symposium on Robot and Human Interactive Communication, pp. 1122–1127. IEEE (2014)
18. Briggs, G., McConnell, I., Scheutz, M.: When robots object: Evidence for the utility of verbal, but not necessarily spoken protest. In: International Conference on Social Robotics, pp. 83–92. Springer (2015)
19. Briggs, G., Scheutz, M.: Investigating the effects of robotic displays of protest and distress. Internatinoal Conference on Social Robotics pp. 238–247 (2012)

20. Briggs, G., Scheutz, M.: How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress. International Journal of Social Robotics **6**(3), 343–355 (2014)
21. Briggs, G., Scheutz, M.: "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)
22. Briggs, G., Scheutz, M.: The case for robot disobedience.(cover story). Scientific American **316**(1), 44–47 (2017)
23. Bringsjord, S., Arkoudas, K., Bello, P.: Toward a general logicist methodology for engineering ethically correct robots. Intelligent Systems **21**(4), 38–44 (2006)
24. Bringsjord, S., Taylor, J.: The divine-command approach to robot ethics. In: Robot Ethics: The Ethical and Social Implications of Robotics, pp. 85–108 (2012)
25. Brown, P., Levinson, S.: Politeness: Some Universals in Language Usage. Cambridge University Press (1987)
26. Buhrmester, M., Kwang, T., Gosling, S.D.: Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? Perspectives on psychological science **6**(1), 3–5 (2011)
27. Carpenter, J., Davis, J.M., Erwin-Stewart, N., Lee, T.R., Bransford, J.D., Vye, N.: Gender representation and humanoid robots designed for domestic use. International Journal of Social Robotics **1**(3), 261 (2009)
28. Charisi, V., Dennis, L., Lieck, M.F.R., Matthias, A., Sombetzki, M.S.J., Winfield, A.F., Yampolskiy, R.: Towards moral autonomous systems. arXiv preprint arXiv:1703.04741 (2017)
29. Chita-Tegmark, M., Lohani, M., Scheutz, M.: Gender effects in perceptions of robots and humans with varying emotional intelligence. In: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 230–238. IEEE (2019)
30. Clark, H.H.: Using language, vol. 1996. Cambridge University Press Cambridge (1996)
31. Clarke, R.: Asimov's Laws of Robotics: Implications for Information Technology. In: M. Anderson, S.L. Anderson (eds.) Machine Ethics, pp. 254–284. Cambridge University Press, New York, NY (2011)
32. Crump, M.J., McDonnell, J.V., Gureckis, T.M.: Evaluating amazon's mechanical turk as a tool for experimental behavioral research. PloS one **8**(3) (2013)
33. Cushman, F.: Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. Cognition **108**(2), 353–380 (2008)
34. Dannenhauer, D., Floyd, M.W., Magazzeni, D., Aha, D.W.: Explaining rebel behavior in goal reasoning agents. In: ICAPS Workshop on EXplainable AI Planning (XAIP) (2018)
35. Dehghani, M., Tomai, E., Forbus, K.D., Klenk, M.: An integrated reasoning approach to moral decision-making. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), pp. 1280–1286 (2008)
36. Dennis, L., Fisher, M., Slavkovik, M., Webster, M.: Formal verification of ethical choices in autonomous systems. Robotics and Autonomous Systems **77**, 1–14 (2016)
37. Eyssel, F., Hegel, F.: (s)he's got the look: Gender stereotyping of robots. Journal of Applied Social Psychology **42**(9), 2213–2230 (2012)
38. Frankfurt, H.G.: On bullshit. Princeton University Press Princeton, NJ (1986)
39. Frasca, T., Thielstrom, R., Krause, E., Scheutz, M.: "can you do this?" self-assessment dialogues with autonomous robots before, during, and after a mission. In: HRI Workshop on Assessing, Explaining, and Conveying Robot Proficiency for Human-Robot Teaming (2020)
40. Fraune, M.R., Kawakami, S., Sabanovic, S., De Silva, P.R.S., Okada, M.: Three's company, or a crowd?: The effects of robot number and behavior on hri in japan and the usa. In: Robotics: Science and systems (2015)
41. Gervits, F., Briggs, G., Scheutz, M.: The pragmatic parliament: A framework for socially-appropriate utterance selection in artificial agents. In: 39th Annual Meeting of the Cognitive Science Society, London, UK (2017)
42. Gibbon, D., Griffiths, S.: Multilinear grammar: Ranks and interpretations. Open Linguistics **3**(1), 265–307 (2017)
43. de Graaf, M.M., Malle, B.F.: How people explain action (and autonomous intelligent systems should too). In: 2017 AAAI Fall Symposium Series (2017)
44. de Graaf, M.M., Malle, B.F.: People's explanations of robot behavior subtly reveal mental state inferences (2019)
45. Greene, J.D.: Why are vmpfc patients more utilitarian. A dual-process theory of moral judgment explains. Department of Psychology, Harvard University, Cambridge, Mass (2004)
46. Greene, J.D.: Dual-process morality and the personal/impersonal distinction: A reply to mcguire, langdon, coltheart, and mackenzie. Journal of Experimental Social Psychology **45**(3), 581–584 (2009)
47. Gureckis, T.M., Martin, J., McDonnell, J., Rich, A.S., Markant, D., Coenen, A., Halpern, D., Hamrick, J.B., Chan, P.: psiturk: An open-source framework for conducting replicable behavioral experiments online. Behavior Research Methods **48**(3), 829–842 (2016)
48. Haring, K.S., Mougenot, C., Ono, F., Watanabe, K.: Cultural differences in perception and attitude towards robots. International Journal of Affective Engineering **13**(3), 149–157 (2014)
49. Haring, K.S., Silvera-Tawil, D., Matsumoto, Y., Velonaki, M., Watanabe, K.: Perception of an android robot in japan and australia: A cross-cultural comparison. In: International conference on social robotics, pp. 166–175. Springer (2014)
50. Hayes, B., Shah, J.A.: Improving robot controller transparency through autonomous policy explanation. In: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, pp. 303–312. ACM (2017)
51. Jackson, R.B., Wen, R., Williams, T.: Tact in noncompliance: The need for pragmatically apt responses to unethical commands. In: Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (2019)
52. Jackson, R.B., Williams, T.: Robot: Asker of questions and changer of norms? In: Proceedings of the International Conference on Robot Ethics and Standards (2018)
53. Jackson, R.B., Williams, T.: Language-capable robots may inadvertently weaken human moral norms. In: Proceedings of the Companion of the 14th ACM/IEEE International Conference on Human-Robot Interaction (2019)
54. Jackson, R.B., Williams, T.: On perceived social and moral agency in natural language capable robots. In: Proceedings of the 2019 HRI Workshop on The Dark

Side of Human-Robot Interaction: Ethical Considerations and Community Guidelines for the Field of HRI (2019)
55. Jackson, R.B., Williams, T., Smith, N.M.: Exploring the role of gender in perceptions of robotic noncompliance. In: Proceedings of the 15th ACM/IEEE International Conference on Human-Robot Interaction (2020)
56. Johnson-Laird, P.N.: Mental models in cognitive science. Cognitive science **4**(1), 71–115 (1980)
57. Johnson-Laird, P.N.: Mental models: Towards a cognitive science of language, inference, and consciousness. 6. Harvard University Press (1983)
58. Kasenberg, D., Arnold, T., Scheutz, M.: Norms, rewards, and the intentional stance: Comparing machine learning approaches to ethical training. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 184–190. ACM (2018)
59. Kasenberg, D., Scheutz, M.: Inverse norm conflict resolution. In: Proceedings of the 1st AAAI/ACM Workshop on Artificial Intelligence, Ethics, and Society (2018)
60. Kasenberg, D., Thielstrom, R., Scheutz, M.: Generating explanations for temporal logic planner decisions. In: Proceedings of the 30th International Conference on Automated Planning and Scheduling (ICAPS) (2020)
61. Kennedy, J., Baxter, P., Belpaeme, T.: Children comply with a robot's indirect requests. In: Proceedings of the International Conference on Human-Robot Interaction, pp. 198–199. ACM (2014)
62. Komatsu, T., Malle, B.F., Scheutz, M.: Blaming the reluctant robot: Parallel blame judgments for robots in moral dilemmas across u.s. and japan. In: Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction, pp. 63–72 (2021)
63. Kuipers, B.: Human-like morality and ethics for robots. In: AAAI Workshop: AI, Ethics, and Society (2016)
64. Kuipers, B.: Toward morality and ethics for robots. In: Ethical and Moral Considerations in Non-Human Agents, AAAI Spring Symposium Series (2016)
65. Le Bui, M., Noble, S.U.: We're missing a moral framework of justice in artificial intelligence. In: The Oxford Handbook of Ethics of AI (2020)
66. Lee, H.R., Šabanović, S.: Culturally variable preferences for robot design and use in south korea, turkey, and the united states. In: 2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 17–24. IEEE (2014)
67. Lee, H.R., Sung, J., Šabanović, S., Han, J.: Cultural design of domestic robots: A study of user expectations in korea and the united states. In: 2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication, pp. 803–808. IEEE (2012)
68. Lee, N., Kim, J., Kim, E., Kwon, O.: The influence of politeness behavior on user compliance with social robots in a healthcare service setting. International Journal of Social Robotics **9**(5), 727–743 (2017)
69. Levinson, S.C.: Presumptive meanings: The theory of generalized conversational implicature. MIT press (2000)
70. Lockshin, J., Williams, T.: "we need to start thinking ahead": The impact of social context on linguistic norm adherence. In: Proceedings of the Annual Meeting of the Cognitive Science Society (2020)
71. Lomas, M., Chevalier, R., Cross II, E.V., Garrett, R.C., Hoare, J., Kopack, M.: Explaining robot actions. In: Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction, pp. 187–188. ACM (2012)
72. Madumal, P., Miller, T., Vetere, F., Sonenberg, L.: Towards a grounded dialog model for explainable artificial intelligence. arXiv preprint arXiv:1806.08055 (2018)
73. Malle, B.F.: Integrating robot ethics and machine morality: The study and design of moral competence in robots. Ethics and Info. Tech. **18**(4), 243–256 (2016)
74. Malle, B.F., Guglielmo, S., Monroe, A.E.: A theory of blame. Psychological Inquiry **25**(2), 147–186 (2014)
75. Mavridis, N.: Grounded situation models for situated conversational assistants. Ph.D. thesis, Massachusetts Institute of Technology (2007)
76. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence (2018)
77. Mills, S.: Gender and politeness, vol. 17. Cambridge University Press (2003)
78. Mills, S.: Gender and impoliteness (2005)
79. Murphy, R.R., Woods, D.D.: Beyond asimov: The three laws of responsible robotics. IEEE Intelligent Systems **24**(4), 14–20 (2009)
80. Nass, C., Moon, Y., Green, N.: Are machines gender neutral? gender-stereotypic responses to computers with voices. Journal of applied social psychology **27**(10), 864–876 (1997)
81. Nikolaidis, S., Kwon, M., Forlizzi, J., Srinivasa, S.: Planning with verbal communication for human-robot collaboration. arXiv preprint arXiv:1706.04694 (2017)
82. Oosterveld, B., Brusatin, L., Scheutz, M.: Two bots, one brain: Component sharing in cognitive robotic architectures. In: Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, pp. 415–415. ACM (2017)
83. Park, D.H., Hendricks, L.A., Akata, Z., Schiele, B., Darrell, T., Rohrbach, M.: Attentive explanations: Justifying decisions and pointing to the evidence. arXiv preprint arXiv:1612.04757 (2016)
84. Pereira, L.M., Saptawijaya, A.: Modelling morality with prospective logic. International Journal of Reasoning-Based Intelligent Systems **1**(3-4), 209–221 (2009)
85. Rosemont Jr, H., Ames, R.T.: Confucian role ethics: A moral vision for the 21st century? V&R unipress GmbH (2016)
86. Russell, S., Dewey, D., Tegmark, M.: Research priorities for robust and beneficial artificial intelligence. AI Magazine **36**(4), 105–114 (2015)
87. Šabanović, S.: Robots in society, society in robots. International Journal of Social Robotics **2**(4), 439–450 (2010)
88. Sarathy, V., Arnold, T., Scheutz, M.: When exceptions are the norm: Exploring the role of consent in hri. ACM Trans. Hum.-Robot Interact. **9**(2) (2019)
89. Schermerhorn, P., Scheutz, M., Crowell, C.R.: Robot social presence and gender: Do females view robots differently than males? In: Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction, pp. 263–270. ACM (2008)
90. Scheutz, M.: The need for moral competency in autonomous agent architectures. In: Fundamental Issues of Artificial Intelligence, pp. 515–525. Springer (2016)
91. Scheutz, M.: The case for explicit ethical agents. AI Magazine **38**(4), 57–64 (2017)
92. Scheutz, M., Briggs, G., Cantrell, R., Krause, E., Williams, T., Veale, R.: Novel mechanisms for natural human-robot interactions in the diarc architecture. In: Proceedings of AAAI Workshop on Intelligent Robotic Systems (2013)

93. Scheutz, M., Williams, T., Krause, E., Oosterveld, B., Sarathy, V., Frasca, T.: An overview of the distributed integrated cognition affect and reflection diarc architecture. In: M.I.A. Ferreira, J. S.Sequeira, R. Ventura (eds.) Cognitive Architectures (in press) (2018)
94. Searle, J.R.: Speech acts: An Essay in the Philosophy of Language (1969)
95. Searle, J.R.: A classification of illocutionary acts. Language in society **5**(1), 1–23 (1976)
96. Shibata, T., Wada, K., Ikeda, Y., Sabanovic, S.: Cross-cultural studies on subjective evaluation of a seal robot. Advanced Robotics **23**(4), 443–458 (2009)
97. Shim, J., Arkin, R.C.: A taxonomy of robot deception and its benefits in hri. In: 2013 IEEE International Conference on Systems, Man, and Cybernetics, pp. 2328–2335. IEEE (2013)
98. Siegel, M., Breazeal, C., Norton, M.I.: Persuasive robotics: The influence of robot gender on human behavior. In: 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2563–2568. IEEE (2009)
99. Stewart, N., Chandler, J., Paolacci, G.: Crowdsourcing samples in cognitive science. Trends in Cognitive Sciences (2017)
100. Strait, M., Briggs, P., Scheutz, M.: Gender, more so than age, modulates positive perceptions of language-based human-robot interactions. In: 4th international symposium on new frontiers in human robot interaction (2015)
101. Strait, M., Ramos, A.S., Contreras, V., Garcia, N.: Robots racialized in the likeness of marginalized social identities are subject to greater dehumanization than those racialized as white. In: 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 452–457. IEEE (2018)
102. Sun, R.: Moral judgment, human motivation, and neural networks. Cognitive Computation **5**(4), 566–579 (2013)
103. Tay, B., Jung, Y., Park, T.: When stereotypes meet robots: the double-edge sword of robot gender and personality in human–robot interaction. Computers in Human Behavior **38**, 75–84 (2014)
104. Thielstrom, R., Roque, A., Chita-Tegmark, M., Scheutz, M.: Generating explanations of action failures in a cognitive robotic architecture. In: Proceedings of NL4XAI: 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (2020)
105. Vanderelst, D., Winfield, A.: An architecture for ethical robots inspired by the simulation theory of cognition. Cognitive Systems Research (2017)
106. Wallach, W., Franklin, S., Allen, C.: A conceptual and computational model of moral decision making in human and artificial agents. Topics in Cognitive Science **2**(3), 454–485 (2010)
107. Wang, Y., Young, J.E.: Beyond pink and blue: Gendered attitudes towards robots in society. In: Proceedings of Gender and IT Appropriation. Science and Practice on Dialogue-Forum for Interdisciplinary Exchange, p. 49. European Society for Socially Embedded Technologies (2014)
108. Wen, R., Jackson, R.B., Williams, T., Zhu, Q.: Towards a role ethics approach to command rejection. In: Proceedings of the 2019 HRI Workshop on The Dark Side of Human-Robot Interaction: Ethical Considerations and Community Guidelines for the Field of HRI (2019)
109. Wen, R., Siddiqui, M.A., Williams, T.: Dempster-shafer theoretic learning of indirect speech act comprehension norms. In: AAAI, pp. 10410–10417 (2020)
110. Williams, T., Briggs, G., Oosterveld, B., Scheutz, M.: Going beyond command-based instructions: Extending robotic natural language interaction capabilities. In: Proceedings of twenty-ninth AAAI Conference on Artificial Intelligence (2015)
111. Williams, T., Jackson, R.B., Lockshin, J.: A Bayesian analysis of moral norm malleability during clarification dialogues. In: Proceedings of the 40th annual meeting of the Cognitive Science Society (2018)
112. Williams, T., Zhu, Q., Wen, R., de Visser, E.J.: The confucian matador: Three defenses against the mechanical bull. In: Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (alt.HRI), pp. 25–33 (2020)
113. Winfield, A.F., Blum, C., Liu, W.: Towards an ethical robot: internal models, consequences and ethical action selection. In: Conference towards autonomous robotic systems, pp. 85–96. Springer (2014)
114. Zhu, Q., Williams, T., Jackson, B., Wen, R.: Blame-laden moral rebukes and the morally competent robot: A confucian ethical perspective. Science and Engineering Ethics **26**(5), 2511–2526 (2020)
115. Zhu, Q., Williams, T., Wen, R.: Confucian robot ethics. Computer Ethics-Philosophical Enquiry (CEPE) Proceedings **2019**(1), 12 (2019)
116. Zwaan, R.A.: Situation models, mental simulations, and abstract concepts in discourse comprehension. Psychonomic bulletin & review **23**(4), 1028–1034 (2016)

Gordon Briggs is a Computer Scientist at the Navy Center for Applied Research in Artificial Intelligence at the U.S. Naval Research Laboratory. He received a joint Ph.D. in Computer Science and Cognitive Science from Tufts University in 2016, and M.Eng. and B.Sc. degrees in Computer Science from Cornell University in 2010 and 2009, respectively. Gordon's research focuses on enabling more natural language based human-robot and human-machine interactions, as well as computational modeling of human perception and natural language generation.

Tom Williams is an Assistant Professor of Computer Science at the Colorado School of Mines, where he directs the Mines Interactive Robotics Research Lab (MIRRORLab. He received a joint Ph.D. in Computer Science and Cognitive Science from Tufts University in 2017, a M.S. degree in Computer Science from Tufts University in 2015, and a B.S. in Computer Science from Hamilton College in 2011. Tom's research focuses on enabling and understanding natural language based human-robot interaction that is sensitive to environmental, cognitive, social, and moral context.

Ryan Blake Jackson is a Ph.D. candidate in Computer Science at the Colorado School of Mines. He has been researching human-robot interaction under advisor Dr. Tom Williams in the Mines Interactive Robotics Research Lab (MIRRORLab) since 2018. Before that, he earned a M.S. in Computer Science from the Colorado School of Mines in 2018 and a B.A. in Computer Science from Colorado College in 2016. Ryan's current

research focuses on perception and performance of identity in artificial agents and generating morally sensitive natural language in clarification and noncompliance interactions.

Matthias Scheutz is a Professor of Cognitive and Computer Science in the Department of Computer Science at Tufts University. He earned a Ph.D. in Philosophy from the University of Vienna in 1995 and a Joint Ph.D. in Cognitive Science and Computer Science from Indiana University Bloomington in 1999. He has more than 400 peer-reviewed publications in artificial intelligence, natural language processing, cognitive modeling, robotics, and human–robot interaction. His current research focuses on complex cognitive robots with natural language and machine learning capabilities