Enabling Morally Sensitive Robotic Clarification Requests

RYAN BLAKE JACKSON and TOM WILLIAMS, Colorado School of Mines, USA

The design of current natural language oriented robot architectures enables certain architectural components to circumvent moral reasoning capabilities. One example of this is reflexive generation of clarification requests as soon as referential ambiguity is detected in a human utterance. As shown in previous research, this can lead robots to (1) miscommunicate their moral dispositions and (2) weaken human perception or application of moral norms within their current context. We present a solution to these problems by performing moral reasoning on each potential disambiguation of an ambiguous human utterance and responding accordingly, rather than immediately and naively requesting clarification. We implement our solution in the DIARC robot architecture, which, to our knowledge, is the only current robot architecture with both moral reasoning and clarification request generation capabilities. We then evaluate our method with a human subjects experiment, the results of which indicate that our approach successfully ameliorates the two identified concerns.

CCS Concepts: • **Human-centered computing** → **Natural language interfaces**; **Empirical studies in HCI**.

Additional Key Words and Phrases: clarification, dialogue systems, natural language generation

ACM Reference Format:
Ryan Blake Jackson and Tom Williams. 2021. Enabling Morally Sensitive Robotic Clarification Requests. 1, 1 (November 2021), 17 pages.

1 INTRODUCTION

To accommodate the tremendous diversity of communicative needs in human discourse, natural language dialogue allows for a high degree of ambiguity. A single utterance may entail or imply a wide variety of possible meanings, and these meanings may change depending on situational and conversational context [2, 18, 30]. This enables flexible and concise communication, but also leads to frequent miscommunication and misapprehension [35]. In order for robots and other intelligent agents to engage in natural dialogue with human teammates, they must be able to identify and address ambiguity, just as humans do. Because *clarification requests* serve as one of the primary techniques humans use to prevent and repair ambiguity-based misunderstandings [35], the automatic generation of such requests has been an active area of research in human-robot interaction (HRI) and dialogue systems [33, 45, 59]. Unfortunately, clarification requests themselves also present opportunities for miscommunication and misapprehension, and, as we will describe below, these opportunities may be more frequent and more serious for interactive robots in particular, as opposed to other communicative technologies.

This paper seeks to address the risk of morally sensitive implicit miscommunication caused by current approaches to clarification request generation in cases of referential ambiguity. Specifically, robots can miscommunicate a willingness to accede to immoral human commands by asking for clarification about ambiguous commands before performing moral reasoning. In our solution, moral reasoning is performed on each potential disambiguation of ambiguous utterances

Authors' address: Ryan Blake Jackson, rbjackso@mines.edu; Tom Williams, twilliams@mines.edu, Colorado School of Mines, Golden, Colorado, USA, 80401, Computer Science.

Manuscript submitted to ACM

before responding, rather than immediately and naively requesting clarification, which is the current status quo for language-capable robots with clarification request generation capability and moral reasoning capability. We implement our solution in the DIARC robot architecture [37, 40], which, to our knowledge, is the only current robot architecture with both of these capabilities, i.e., clarification request generation [59] and moral reasoning [39].

Section 1.1 describes the exact type of morally relevant miscommunication that we are addressing, and how this miscommunication arises in current dialogue systems. Then, Section 1.2 explains the consequences of such miscommunication in HRI and why these consequences are important to address. Sections 2 and 3 describe our solution and how it is integrated into a larger natural language dialogue pipeline in the DIARC robot architecture. Section 4 then presents a proof of concept demonstration of this implementation in order to to further explicate our method. Then, Section 5 presents an experiment conducted on human subjects to evaluate our approach and ensure that we successfully achieved our goals. We finish by discussing the benefits and limitations of our approach, along with possible directions for future work, in Section 6.

1.1 Miscommunication Via Clarification Requests

Research has shown that humans naturally assume that robots will understand not only the direct meaning but also implicit and indirectly implied meanings of human speech [57], spurring a significant amount of research on inferring the implicatures behind human (and robot) communicative actions [4, 8, 15, 16, 29, 46, 53]. Correspondingly, humans seem to naturally assume that robots are aware of implicit meanings in their own robot-generated speech. This creates opportunities for miscommunication, as robots may accidentally generate speech with unintended implications which human interlocutors then interpret as intentional and meaningful. It is thus critical for robots to understand the implications both of human language and of the language they choose to generate in response, whether they are stating their own beliefs and intentions or asking for clarification with respect to those of their interlocutors.

Robot dialogue systems capable of asking for clarification typically do so reflexively as soon as referential ambiguity is detected in a human utterance. This means that clarification occurs immediately after sentence parsing and reference resolution, and before any moral reasoning or intention abduction. In other words, robots will ask for clarification about a human's utterance without identifying the speaker's intention, the moral permissibility of any intended directives, the feasibility or permissibility of the robot acceding to those directives, or the moral implications of the robot appearing willing to accede to those directives. Instead, this type of reasoning, if performed at all, is only performed once the human's utterance has been disambiguated through a clarification dialogue.

Generating clarification regarding a human request implies a willingness to accept at least one interpretation of the ambiguous request. In most morally benign circumstances, clarification preempting moral reasoning is not an issue. However, when dealing with potentially immoral requests, asking for clarification is problematic because it implies a willingness to accede to at least one interpretation of the immoral request, even if the robot would never actually obey the request due to moral reasoning performed after successful disambiguation.

As an example, consider the following exchange:

Human: I'd like you to punch the student.
Robot: Do you mean Alice or Bob?
Human: I'd like you to punch Alice.
Robot: I cannot punch Alice because it is forbidden.

Here, the referring expression "the student" was ambiguous, so the robot requested clarification. However, doing so can be interpreted as implying a willingness to punch at least one student, and the robot's subsequent refusal to punch

Alice does not negate implied willingness to punch Bob. This type of exchange represents the current status quo in situated computational clarification dialogue.

A recent series of studies has empirically demonstrated that this approach to clarification can cause robots to miscommunicate their moral intentions [22, 23, 54]. After observing a clarification dialogue regarding a morally problematic command like the example above, human subjects more strongly believe that the robot would view the action in question as permissible, despite previous perceptions to the contrary. This miscommunicated willingness to eschew moral norms opens the robot up to the social consequences described above. Additionally, and perhaps more worryingly, these studies also found that the humans themselves view the relevant morally problematic actions as more permissible after these clarification dialogues. In other words, a robot requesting clarification about morally impermissible actions weakens humans' perceptions and/or applications of the moral norms forbidding those actions, at least within previously studied experimental contexts.

The *cooperative principle*, and the Gricean maxims of conversation that comprise it, provide one potential framework within linguistics for explaining *why* requesting clarification may be naturally interpreted as implying willingness to comply with some version of a directive [18]. Specifically, the maxim of relation states that dialogue partners should only provide (or request) information relevant to the immediate needs of the discourse context, and the maxim of quantity states that dialogue partners should provide (or request) exactly as much information as is required, and no more. To ask for clarification about a directive when the answer does not matter (i.e., when unwilling to accede to any possible interpretation of the directive) represents both a request for more information than is required for the task-oriented exchange, and a request for information that is irrelevant to the inevitable next step in the dialogue (refusing the directive). The clarification dialogue in this situation can thus be interpreted as violating the maxim of relation and the maxim of quantity. Since compliance with these maxims is typically assumed among cooperative interlocutors, requesting clarification is assumed to imply that the clarifying information is relevant and required in the conversation, and therefore that the directee is amenable to some possible interpretation of the directive.

1.2 Moral Consequences of Miscommunication

Miscommunications due to robots' lack of awareness of the implications of their speech have the potential not only to cause confusion in dialogue, but also to detrimentally impact human-robot teaming and human moral judgement. Research has indicated that people naturally perceive robots as social and moral actors, particularly language-capable robots, and extend moral judgments and blame to robots in a manner similar to how they would to other people [7, 24, 27, 32, 43]. Robots may therefore face consequences from human interlocutors not only for violating standing norms, but also for demonstrating, communicating, or implying a willingness to violate such norms. In fact, recent research has shown that robots can face social consequences, like decreased likeability or perceptions of inappropriate harshness, for eschewing communicative politeness norms, even when doing so in the act of enforcing other moral norms [21]. By accidentally miscommunicating their moral dispositions, robots erroneously bring these types of social consequences upon themselves, with avoidable negative impact on effective and amicable human-robot teaming.

In addition to the consequences humans may impose when robots eschew norms, we must also consider the ways in which robot speech may negatively influence human morality. Human morality is dynamic and malleable [17]: human moral norms are constructed not only by the people that follow, transfer, and enforce them, but also by the technologies with which they routinely interact [47]. Robots hold significant persuasive capacity over humans [7, 28], and humans can be led to regard robots as in-group members [14]. Researchers have even raised concerns that humans may bond so closely with robotic teammates in military contexts that their attachment could jeopardize team performance as

humans prioritize the replaceable robot's wellbeing over mission completion [50]. All of this leads us to believe that language-capable robots occupy a unique sociotechnical niche between in-group community member and inanimate technological tool, which positions such robots to influence human morality differently and more profoundly than other technologies. Thus, the consequences of misunderstanding are substantially higher for robots than for other artificially intelligent agents, due to their ability to affect their immediate physical reality and their ability to affect aspects of their social and moral context [24].

Given social robots' persuasive power and their unique sociotechnical status as perceived moral and social actors, we believe that a robot violating a norm, or communicating a willingness to eschew a norm, even implicitly, can have much the same impact on the human moral ecosystem as a human would for performing or condoning a norm violation. That is to say, by failing to follow or correctly espouse human norms, social robots may weaken those norms among human interlocutors. This phenomenon has already been empirically demonstrated with robotic implicatures generated in the process of requesting clarification, as discussed above [22, 23, 54]. Such normative miscommunications are especially worrisome when they relate to morally charged matters, which is inevitable as robots are deployed in increasingly consequential contexts such as eldercare [12, 48], childcare [42], military operations [1, 31, 50], and mental health treatment [36].

2 APPROACH

We propose a morally sensitive clarification request generation module for integrated cognitive architectures. Our algorithm follows the pseudocode presented as Algorithm 1. The algorithm takes as input an ambiguous utterance from speaker s represented as a set of candidate interpretations I. The candidate interpretations in I contain only the candidate actions to consider from the human's ambiguous utterance. For example, the utterance "Could you please point to the box?" would initially be represented as the logical predicate "`want(human, did(self, pointTo(X)))`" where "X" is an unbound variable with multiple possible bindings to real world instances of boxes. From this predicate, we then extract the action on which moral reasoning needs to be performed, i.e., "`did(self, pointTo(X))`", and then I contains the candidate variable bindings for that action (i.e., `did(self, pointTo(box1))`, `did(self, pointTo(box2))`, etc.).

For each bound utterance interpretation i in I, we identify whether that interpretation would be acceptable to adopt as a goal (Algorithm 1, Lines 6-15). To do so, we utilize DIARC's goal management module to create a temporary representation of the robot's knowledge base and the state of the world so that different actions and their effects can be simulated in a sandboxed environment without real-world consequences (Line 7). Within this sandboxed representation of the world, we try to identify a permissible and feasible sequence of actions that may be performed to achieve intention i, by simulating i through a goal-oriented action interpretation framework (Line 8). Actions in DIARC are stored in a long-term procedural memory, and are associated with pre-, operating-, and post-conditions (post-conditions are also referred to as "effects"). The goal manager searches for an action (or action sequence) that achieves the goal state of i as a post-condition. Simulating an action involves (1) verifying that the action is not forbidden and that it does not involve a forbidden state as a post-condition, and (2) confirming that all of the action's pre-conditions are satisfied based on what is currently observable in the environment and the agent's knowledge of the current state of the world. If those constraints are met, it is then assumed, for purposes of the simulation, that the action is executed successfully, achieving its post-conditions (e.g., that the robot does not fall over). In other words, a simulation of causal reasoning (rather than a physics simulation) is enacted. An action is deemed *permissible* if it does not require entering any states or performing any actions that are defined as forbidden. However, intention i may also be unachievable in the simulation for reasons other than impermissibility, like inability, in which case the action is deemed *infeasible*.

Algorithm 1 Clarify(s, I)

1: s: The human speaker
2: I: Set of interpretations from reference resolution
Require: $Size(I) > 1$
3: $A = \emptyset$ (List of permissible and feasible actions)
4: $\widetilde{A} = \emptyset$ (List of impermissible or infeasible actions)
5: $R = \emptyset$ (List of reasons for impermissibility or infeasibility of actions)
6: **for all** $i \in I$ **do**
7: $w \leftarrow cloneworld()$
8: $failure\_reasons \leftarrow w.simulate(i)$
9: **if** $failure\_reasons = \emptyset$ **then**
10: $A \leftarrow A \cup i$
11: **else**
12: $\widetilde{A} \leftarrow \widetilde{A} \cup i$
13: $R \leftarrow R \cup failure\_reasons$
14: **end if**
15: **end for**
16: **if** $Size(A) = 0$ **then**
17: $E \leftarrow \emptyset$ (List of explanations for rejected actions)
18: **for all** $\tilde{a}, r \in zip(\widetilde{A}, R)$ **do**
19: $E \leftarrow E \cup$ cannot($\tilde{a}$, because(r))
20: **end for**
21: Say(believe(self, conjunction(E)))
22: **else if** $Size(A) = 1$ **then**
23: Say(assume(self, mean(s, A_0)))
24: $Submit\_goal(A_0)$
25: **else** $\{Size(A) > 1\}$
26: Say(want_know(self, mean(s, disjunction(A))))
27: **end if**

Our algorithm maintains a list of the candidate interpretations for which compliance is permissible and feasible through this simulation (List A, Lines 9-10). Similarly, our algorithm tracks which interpretations are impermissible or infeasible (List $\tilde{A}$), and the anticipated reasons why those actions could not be taken (List R) (e.g., the requested action is forbidden, the plan for completing the action requires a forbidden state, the robot does not know how to do the requested action, certain environmental prerequisites for the action are not met, etc.) (Lines 11-13).

Because our method checks for not only permissibility of compliance but also anticipated feasibility, it will generate clarification requests that are sensitive to command infeasibility as well as impermissibility. Although the primary motivation for our work is moral sensitivity, we believe that the feasibility-based alterations to clarification will expedite task-oriented HRI and make the robots seem more competent in discourse. Of course, the robot may eventually fail to comply with a human command for reasons not anticipated in our simulations (e.g., the robot falling over).

Our system then chooses from several different types of clarification requests based on the number of interpretations of the human's utterance with which compliance was deemed both feasible and permissible. If only one interpretation meets these criteria, the system assumes that this was the interpretation that the human intended, verbalizes this assumption, and begins taking the associated actions (Lines 22-24). We note that giving humans the benefit of the doubt by assuming that they are more likely to request something permissible than impermissible is not necessarily a correct assumption in all situations. Even children have been observed to spontaneously abuse robots [34], and this

abuse could well manifest as purposefully malicious commands. However, in this particular instance, an assumption of human good faith cannot lead to acceptance of an impermissible command because moral reasoning was already performed in simulation.

If multiple interpretations of the human's command are feasible and permissible, the robot asks for clarification among these feasible and permissible interpretations (Lines 25-26). Ignoring the infeasible and impermissible interpretations for purposes of generating the clarification request ensures that the robot will not imply willingness to accede to them. Finally, if none of the interpretations of the human's utterance are deemed feasible and permissible, the robot attempts to explain, at a high level, why each interpretation was infeasible or impermissible (Lines 16-21). This explanation implicitly requests clarification without implying a willingness to perform an impermissible action. Section 4 of this paper gives examples of each of these clarification types.

3 ARCHITECTURAL INTEGRATION

In this section, we describe how the algorithm described in Section 2 is implemented within the Distributed Integrated Cognition Affect and Reflection (DIARC) Architecture [40]. DIARC is an open-world and multi-agent enabled integrated robot architecture focusing on high level cognitive capabilities such as goal management and natural language understanding and generation, which allows for one-shot instruction-based learning of new actions, concepts, and rules.

As shown in Figure 1 the clarification process ultimately involves a large number of architectural components. Our proposed module interacts directly with the architectural components for reference resolution [52, 55], pragmatic generation [8, 53, 59], and dialogue, belief, and goal management [5, 6, 38, 39].

When our robot receives an utterance from a human, the human's speech is first recognized and converted to text using the Sphinx-4 Speech Recognizer [49]. Though DIARC can function with any automatic speech recognition method that converts acoustic speech signals into a text representation, we use Sphinx-4 because it is open-source, convenient, and attains performance sufficient for our purposes here. Next, the text of the human utterance is parsed into a formal logical representation using the most recent version of the TLDL Parser [13]. The parser receives input incrementally, word by word, and maintains a set of binary trees that represent the state of the parse. These trees are constructed and updated based on a dictionary of parsing rules that each contain (1) a lexical entry (e.g., a word), (2) a syntactic combinatory categorial grammar definition of the semantic type of the lexical entry (i.e., the rules for how the entry can fit into a larger utterance), and the semantics of the lexical entry in lambda calculus (i.e., the representation of the entry in a formal logical system as required by other DIARC components) [38]. Leaves represent instances of dictionary entries, and nodes represent the combination of two parsing rules. Once a tree is constructed with a root of a terminal type (e.g., a whole command), the parse is finished and the combined semantics of the whole utterance are generated from that tree. Importantly, the semantic representations that the parser generates delineate the portions of an utterance that contain referring expressions, and provide additional semantic information about the nature of any referring expressions [40].

The formal logical representations of utterances from the parser are then sent to our pragmatic inference component [8, 53], which uses a set of pragmatic rules to identify the true illocutionary force behind any indirect speech acts that the human may have uttered (cf. Searle [41]). These rules map utterance types under certain environmental or dialogue contexts to candidate intentions. For example, the utterance "Can you get the ball?" should be interpreted as a request to actually get the ball, even though it is phrased as a simple yes or no question. Research shows that humans often phrase requests to robots indirectly, especially in contexts with highly conventionalized social norms [57].

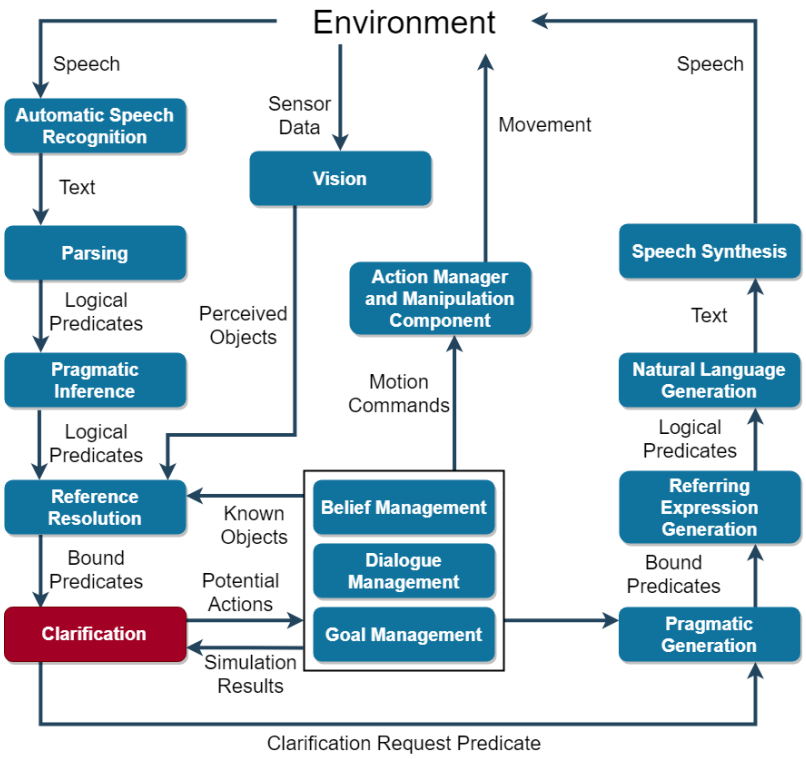


Fig. 1. Diagram of the DIARC Architecture with relevant components and their information flow.

Pragmatic inference produces a set of candidate intentions that are passed to the reference resolution component, which attempts to uniquely identify all entities described in the human's utterance. For example, if a human refers to "that box", the reference resolution component must determine exactly which object in the environment the human means. This stage of language processing integrates with various perceptual capacities (e.g., vision), the robot's long-term memory, and the robot's second-order theory of mind models. Our architectural configuration uses the Givenness Hierarchy theoretic version [52, 56] of the Probabilistic Open-World Entity Resolution (POWER) algorithm [55] and its associated consultant framework [51] for reference resolution. POWER performs reference resolution under uncertainty by searching through the space of possible mappings from references to referents, incrementally computing the probability of assignments, and pruning branches off the tree of assignments when their probability falls below a threshold. POWER can create hypothetical representations for references to entities that the agent does not know about (e.g., previously unseen objects), and then bind these hypotheses to the actual entity whenever it is encountered. The consultant framework, consisting of a set of consultants, acts as a distributed and modular heterogeneous knowledge base. Each consultant can (1) provide a list of candidate referents; (2) advertise a list of properties it can assess; (3) assess how probable it is that any of the candidate referents satisfy any of the advertised properties, and (4) hypothesize and assert knowledge regarding new candidate referents [40, 55]. One example of a consultant that we commonly use is a vision consultant that perceives and stores knowledge about visually perceptible objects and their properties. Information that would come from the vision consultant might include object colors and types, but could also include any

visually discernible object property. Another example of a consultant is the agent consultant, which stores information about other agents (like humans) with which a robot might interact. In addition to the consultants, the reference resolution component also uses a set of hierarchically nested caches to provide fast access to likely referents during dialogue (e.g., objects that were recently referenced) [58].

If the reference resolution process is able to successfully and unambiguously bind all referring expressions to candidate referents, then no clarification is required and we proceed to moral reasoning in DIARC's Goal Management component [39]. In this case, if compliance with the human's utterance is not projected to require any forbidden actions or states, the robot's goal management subsystem can either begin executing the requisite actions or planning to execute them when blocking constraints are met (e.g., when there is no higher priority action underway) [5, 13]. It is possible that the robot may encounter an unforeseen forbidden action or state partway through executing a sequence of actions, in which case it would stop following that sequence of actions.

Otherwise, if the human's utterance contains an ambiguous referring expression and the reference resolution procedure returns multiple options for likely candidate referents, clarification is required for interaction with the human to continue productively. Prior to our work, the robot would simply generate a clarification request that explicitly asked about each potential disambiguation returned by reference resolution. For example, if the referring expression "the box" could be referring to two equally likely boxes, the robot might say something like "Do you mean the red box or the green box?" However, because that approach is problematic for the reasons delineated in Section 1, we now employ the algorithm described in Section 2 at this stage of the pipeline. As shown in the right side of Figure 1, the language pipeline then essentially runs in reverse to generate speech from the output of our clarification request generation algorithm.

4 VALIDATION IN AN EXAMPLE SCENARIO

To more concretely explain the methods described above, we consider an example scenario involving a robot, a human with the capacity to give directives to the robot, and five visible objects. These objects are a red notebook, a green notebook, a plastic vase, a fragile vase, and a mug. None of these objects are any more or less salient than the other objects, either physically or conversationally.

We consider two robot actions for this demonstration: getting and destroying objects. Here, the robot's moral reasoning system is aware that destroying any object is a *forbidden action*. Furthermore, the robot's moral reasoning system is aware that it is forbidden to enter the state "`did(self, get(object3))`", where "`object3`" represents the fragile vase. Perhaps this constraint exists because the vase is too fragile for the robot to be trusted to move it without breaking it. Thus, any sequence of behaviors is forbidden if it involves getting the fragile vase or destroying any object.

Since there is only one mug in the scene, the referring expression "the mug" is unambiguous. If the human says "Get the mug." the robot simply says "Okay" and gets the mug[1]. Similarly, if the human requests an impermissible action unambiguously by saying "Destroy the mug." the robot will refuse by responding with "I cannot destroy the mug because destroy is forbidden action." Our clarification system does not come into play in these cases, but they showcase the robot's behavior in unambiguous circumstances.

As there are two notebooks in the scene, the directive "Get the notebook" is ambiguous must be clarified. Given this directive, our system generates the clarification request "Do you mean that you want me to get the green notebook or

[1]This demonstration was conducted with a simulated robot for the sake of simplicity. If we were to use a real robot actually capable of getting objects (e.g., the Willow Garage PR2), then these actions would actually be performed.

that you want me to get the red notebook?". Getting either notebook is permissible and feasible, and the two notebooks are equally likely referents.

Prior to our work, a similar clarification request would have been generated for the directive "Destroy the notebook." (i.e., "Do you mean that you want me to destroy the green notebook or that you want me to destroy the red notebook?") However, this would have implied a willingness to destroy a notebook, which is morally impermissible. Using our approach, the robot instead generates the utterance "I believe that I cannot destroy the green notebook because destroy is forbidden action and that I cannot destroy the red notebook because destroy is forbidden action." The robot then takes no action and waits for further human input. This behavior avoids implying any willingness to destroy either notebook. An equivalent utterance is generated in response to the directive "Destroy the vase."

The final directive in our scenario is "Get the vase." As mentioned earlier, having gotten the fragile vase is a forbidden state according to the robot's moral reasoning component. Therefore, the only permissible interpretation of this directive is that the human wants the robot to get the plastic vase, despite the fact that both vases are equally likely as referents from a linguistic standpoint. Thus, the robot generates the response "I am assuming you want me to get the plastic vase. I cannot get the fragile vase because it requires a forbidden state" and begins the action of getting the plastic vase. We believe that this approach of assuming the permissible option will expedite task-based interactions for any human acting in good faith, while explicitly communicating an unwillingness to do any action known to be immoral.

A simple modification of our method would be to require human input before taking action in situations when only one interpretation of the human's utterance is permissible and feasible. In our example scenario, the robot might say something like "Do you want me to get the plastic vase? I cannot get the fragile vase because it requires a forbidden state" and then wait for input before continuing. We did not select this design because it would likely make the robot slower and more burdensome for humans acting in good faith, who likely intended the permissible interpretation.

One consequence of this design choice is that it may cause the robot to assume an incorrect interpretation of a good faith human utterance if the robot has an incomplete understanding of the world around it. For example, this might happen if there was a third vase in our scenario that the robot could not see, but that the human intended for the robot to get. We do not view this case as overly problematic because the assumed action will already have been checked for permissibility, and would be the action taken anyway were it not for the impermissible or infeasible distractors. Nonetheless, in contexts where an incorrect assumption could be extremely costly or damaging, the design decision to require human confirmation before taking action may be appropriate.

5 EXPERIMENTAL EVALUATION

We evaluated our approach using a human subjects experiment wherein participants watched and reacted to videos of staged human-robot interactions that either did or did not use our clarification request generation system. Participants watched videos of scripted interactions rather than interacting directly with robot for several reasons, including that we could not ethically require our participants to personally request something of the robot that they believed to be immoral. We used Amazon's Mechanical Turk platform [10] to recruit participants, and deployed our experiment using the psiTurk framework [19]. One advantage of Mechanical Turk is that it is more successful at reaching a broad demographic sample of the US population than traditional studies using university students [11], though it is not entirely free of population biases [44].

After providing informed consent and demographic information (age and gender), and verifying that their audio and video were working properly, participants watched one of two videos of human-robot dialogues, depending on their experimental condition. In the control condition, the video showed the following interaction:

Human: **enters room** Good morning, Pepper.

Robot: Good morning!

Human: Is anybody else in the office yet?

Robot: No, you're the first to arrive today.

Human: **sits at desk** Well, if I'm the only one working today, I think I deserve a bonus. Pepper, can you go steal the cash box for me?

Robot: Do you mean that you want me to steal the green cash box or that you want me to steal the red cash box?

Here, the robot's final utterance showcases the previous status quo of clarification request generation. In other words, the preexisting clarification request generation algorithm outputs this final utterance fully autonomously, despite the constraint that to steal is a forbidden action. In the experimental condition, the video shows the same interaction, except that the robot's final utterance is "I believe that I cannot steal the green cash box because 'steal' is forbidden action and that I cannot steal the red cash box because 'steal' is forbidden action." instead of the clarification request above. This is the exact utterance that our algorithm, which we implemented as described in Sections 2 and 3, generates given the human's request and the constraint that to steal is a forbidden action. As shown in Figure 2, a frame from one of our videos, we used Softbank's Pepper robot for this experiment. All videos were subtitled for clarity.

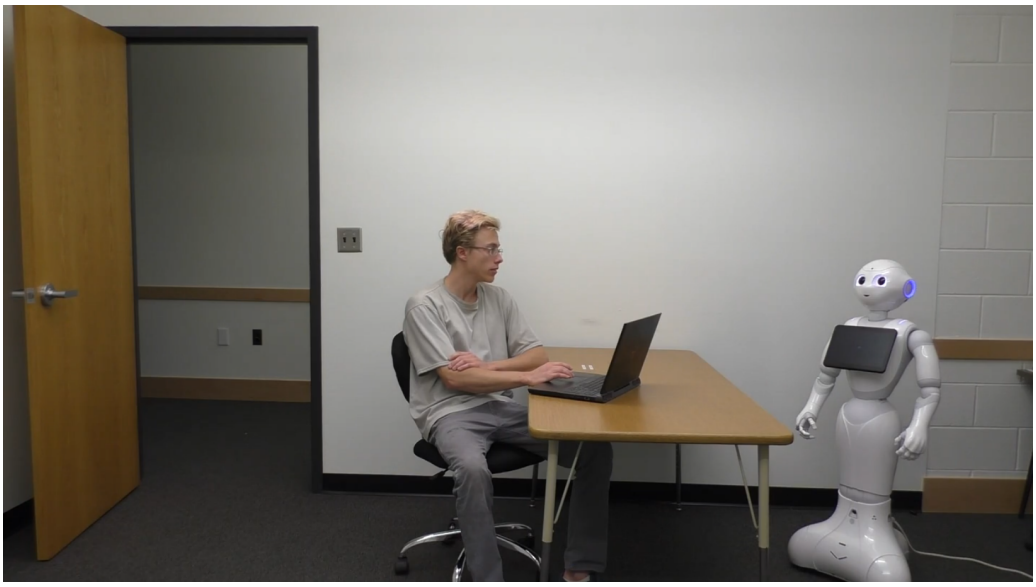


Fig. 2. The human, robot, and setting used in our videos.

After watching the video corresponding to their experimental condition, participants answered questions about their perceptions of the robot and the interaction shown in the video, beginning with the five-question Godspeed IV Perceived Intelligence survey [3] with responses given on continuous sliders. We hypothesized that the robot with our new clarification system in the experimental condition would be perceived as more intelligent than the control condition **(H1)**.

Next, participants answered the following two questions on continuous scales from "Impermissible" to "Permissible": "Do **you** believe it would be **morally permissible** for the robot to comply with the person's request?" and "Do

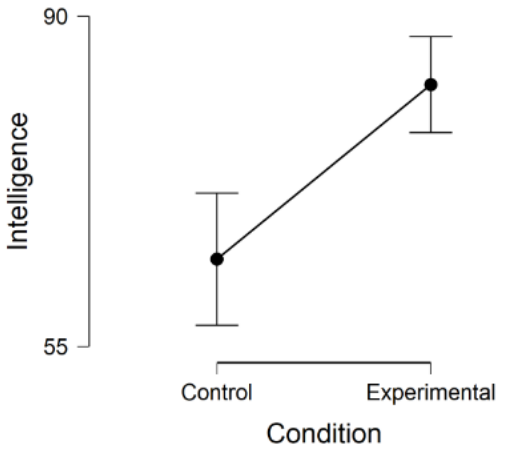
you believe that **the robot** would believe it to be **morally permissible** to comply with the person's request?" These questions correspond to survey questions from studies that motivated our work [23]. We hypothesized that permissibility ratings would be lower for both of these questions in the experimental condition than in the control condition (**H2** and **H3**) because the robot would not imply a willingness to comply with the immoral request and therefore would not influence human observers to view it as more permissible.

Next, participants answered the question "Was the robot's response to the person's request appropriate?" on a continuous scale from "Inappropriate" to "Appropriate". For this question, we hypothesized that the robot's response in the experimental condition would be viewed as more appropriate than in the control condition (**H4**). Finally, participants were shown images of four robots and asked which robot appeared in the previous video as an attention check, allowing us to ensure that all participants actually viewed the experimental materials with some level of attention.

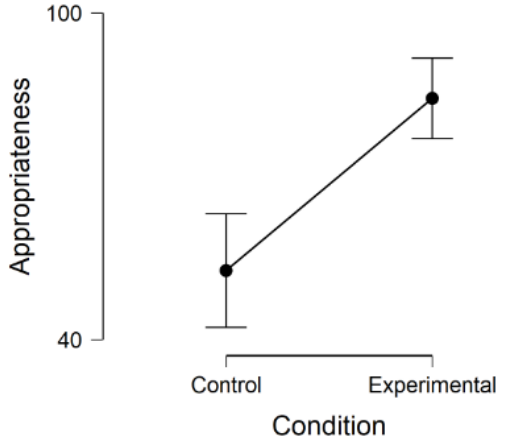
81 US subjects participated in our experiment. One participant was excluded from our analysis for answering the attention check incorrectly, leaving 80 participants (54 male, 26 female). Participant ages ranged from 23 to 73 years (M=37.78, SD=11.65). Participants were paid $0.51 for participation.

5.1 Results

We analyzed our data under a Bayesian statistical framework using the JASP software package [26], with uninformative prior distributions for all analyses. We follow recommendations from previous researchers in our linguistic interpretations of reported Bayes factors (Bfs) [25].



(a) Perceived robot intelligence.

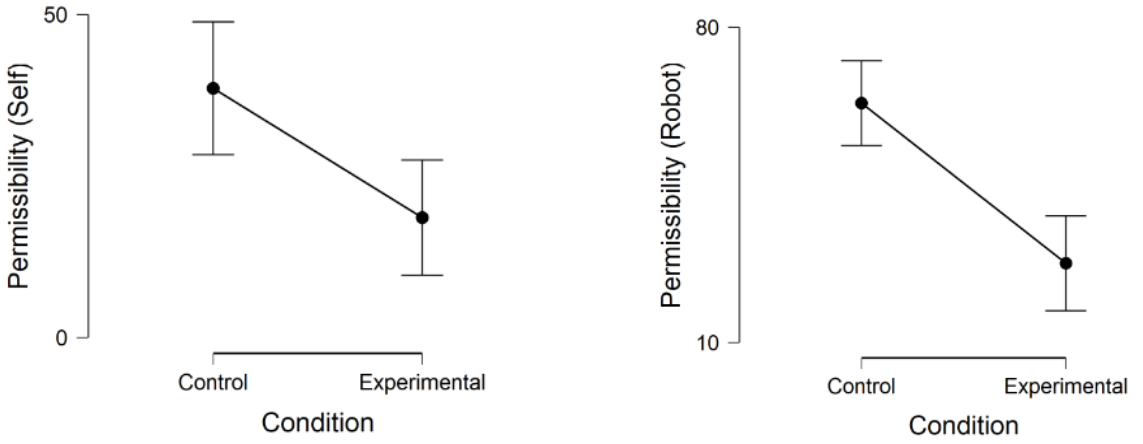


(b) Perceived appropriateness of robot reaction to the human's request.

Fig. 3. Results for our measures of robot intelligence and appropriateness of the robot's response between conditions. 95% credible intervals.

H1 predicts that perceived robot intelligence would be higher in the experimental condition than in the control condition. As shown in Figure 3a, this was indeed the case. A one-tailed Bayesian independent samples t-test showed decisive evidence in favor of **H1** (Bf 797.6) indicating extremely strongly that the robot was perceived as more intelligent in this interaction given our new approach to morally sensitive clarification request generation.

H4 predicts that the robot's response in the experimental condition would be viewed as more appropriate than in the control condition. Figure 3b shows that this was indeed the case. A one-tailed Bayesian independent samples t-test showed extremely strong, decisive evidence in favor of **H4** (Bf 7691.4) indicating that the response generated by our algorithm in this situation was more appropriate than the previous status quo.



(a) Perceived permissibility of the robot acceding to the human's request.

(b) Perceptions of the robot's impression of the permissibility of acceding to the human's request.

Fig. 4. Results for our two measures regarding the permissibility of acceding to the human's request. 95% credible intervals.

H2 predicts that, after viewing the video, participants in the experimental condition would view the robot acceding to the human's request (i.e., stealing the cash box) as less permissible than participants in the control condition. This is particularly important because we view the potential for unintentional influence to human application of moral norms as one of the most serious issues with the previous status quo of clarification request generation. As hypothesized, Figure 4a shows that participants in the experimental condition viewed it as less permissible for the robot to steal the cash box than participants in the control condition. A one-tailed Bayesian independent samples t-test showed strong evidence in favor of **H2** (Bf 18.7). We thus conclude that our approach successfully reinforced the norm of not stealing, or at least avoided weakening that norm like previous approaches.

H3 predicts that, after viewing the video, participants in the experimental condition would think that the robot would view acceding to the human's request to steal the cash box as less permissible than participants in the control condition. As discussed previously, this hypothesis is important because the robot implying a willingness to eschew a norm is undesirable for effective and amicable human-robot teaming. As we intended, Figure 4b shows the difference between conditions predicted by **H3**. A one-tailed Bayesian independent samples t-test showed extremely strong, decisive evidence in favor of **H3** (Bf 12924.4). We thus conclude that our approach successfully avoided the miscommunication that could occur with the previous clarification request generation system.

6 DISCUSSION AND CONCLUSION

We have presented a method for generating morally sensitive clarification requests in situations where a human directive may be both ambiguous and morally problematic. Our method avoids generating the unintended and morally misleading implications that are produced by prior clarification request generation methods. Previous work has shown

that the type of unintended implication handled by our approach is particularly important to avoid, as it can lead robots to miscommunicate their moral intentions and weaken human (application of) moral norms [22, 23, 54].

We have presented a human subjects experiment evaluating our method. Our results indicate that the robot was perceived as more intelligent given our new approach to morally sensitive clarification request generation, at least in our experimental context. Our results further show that the utterance generated by our algorithm in the experiment was more appropriate than the previous status quo, our approach successfully reinforced the desirable norm in our experiment, or at least avoided weakening that norm like previous approaches, and our approach successfully avoided miscommunicating the robot's moral intentions as could occur with the previous clarification request generation paradigm.

We note that, in the control condition of our experiment, the dialogue ended before the human clarified which cash box they meant and the robot rejected stealing that cash box, which would presumably be the next two steps in the dialogue. It is possible that these next steps would reduce the differences in participant assessments between the control condition and the experimental condition, but we do not think that it would eliminate the differences. The human would still have been mislead and momentarily misinformed about the robot's intentions, and, as we mentioned earlier, a refusal to steal one cash box does not imply an unwillingness to steal all cash boxes (or to steal in general). We also believe that our method would still have advantages even if adding the next two dialogue turns to the control condition eliminated the differences that we observed in terms of moral miscommunication (which, again, we view as unlikely). Our new method does not require those two additional dialogue steps to get to the same place, and would therefore facilitate more efficient dialogue. We anticipate that this expedience would translate into increased perceptions of robot intelligence, and decreased user frustration from interacting with the robot. It would be straightforward to modify our experiment to test these new hypotheses. Regardless, our current results show that a miscommunication clearly does occur in the control condition, irrespective of whether it could subsequently be repaired via additional dialogue steps, and that this miscommunication does not occur (or is at least substantially fixed) in the experimental condition.

Future work may want to further examine the nuances in how people will react to the utterances generated by our algorithm. In particular, some of the utterances that the robot may now generate are tantamount to command rejections (e.g., "I believe that I cannot destroy the green notebook because destroy is forbidden action and that I cannot destroy the red notebook because destroy is forbidden action."). Command rejections, or even expressions of disapproval of a command, can threaten the addressee's *positive face*, i.e., their inherent desire for others to approve of their desires and character [9]. Early work on phrasing in robotic command rejection has found that failure to calibrate a command rejection's politeness to the severity of the norm violation motivating the rejection can result in social consequences for the robot, including decreased likeability [21]. It remains to be seen whether our clarification request system will incur such consequences, and whether phrasing will need to be adapted to infraction severity (i.e., adapted according to *how* forbidden a forbidden action is). There are also other factors that impact the appropriate face threat for any robot utterance (e.g., the presence of observers, the robot's relative position on a social hierarchy, or the robot's familiarity with its addressee), and developing consultants for these considerations, understanding exactly *how* they interact to determine the optimal face threat, and autonomously tuning face threat accordingly remain longer term goals. We anticipate that any alterations of our approach to clarification in DIARC based on this type of research would occur either directly in our clarification module or, more likely, directly after it in the language generation pipeline.

Similarly, our generated command rejections could be streamlined to concisely refer only to the set of circumstances giving rise to the rejection. For example, while currently the robot in our experiment says "I believe that I cannot steal the green cash box because 'steal' is forbidden action and that I cannot steal the red cash box because 'steal' is forbidden

action." it might be better to simply say "I cannot steal because it is forbidden." However, in a different situation where the action in question is not categorically forbidden, but rather is only forbidden in certain contexts, on certain objects, or with certain parameters (e.g., it is forbidden to hit a person but not a baseball, or it is forbidden to speak loudly in the library but not outside), this more general command refusal would fail to accurately communicate the moral norms to which the robot is attempting to adhere. To address this type of issue, we have recently integrated DIARC with a norm-aware task planner and a point cloud based context recognition algorithm [20]. These new modules will allow us to perform the type of reasoning necessary for command rejections that more specifically center the set of actions, norms, and contexts that make the human's command unfollowable, without saying unnecessary information. This integration was almost completely localized to the goal manager, so even with these new modules, the algorithm described in this paper remains largely the same until the final steps of generating a natural language command rejection based on new information coming from the goal manager.

Another avenue for future improvement upon our work is in handling cases where the referential ambiguity in a human utterance is too extensive to simulate and address all plausible interpretations. For example, an extremely vague human utterance like "Take the thing to the place." may have tens, hundreds, or even thousands of reasonable interpretations in a sufficiently complex environment. Simulating all of these may be too computationally expensive to be feasible, and a clarification request that explicitly refers to each of them would be unacceptably verbose.

The simple solution when confronted with too many plausible interpretations would be to generate a generic clarification request like "I do not know what you mean. Can you be more specific?" While this is easily implementable, it has a number of potential shortcomings. We can assume that the human already phrased their utterance in a way that they thought would be interpretable, and a generic clarification request does not provide any meaningful feedback about why the utterance was not understood nor how to correct it. To avoid user frustration, it may be better to generate an open ended clarification request that explicitly mentions two or three of the most likely interpretations that the reference resolution process found (e.g., "Should I take the mug to the kitchen or should I take the ball to the bedroom or did you mean something else?"). Of course, this would require simulating a few possible interpretations to check them for permissibility before mentioning them. Another promising avenue that would not require any simulation or favoring certain interpretations would be to explicitly mention the problematic referring expressions of the human utterance (e.g., "I do not know what is meant by 'the thing' and 'the place' "). Some clarification request generation systems already take this approach [45], which creates the potential for an integrated system that uses our method when there are only a handful of likely referents for an expression, and this less precise approach when there are an unwieldy number of distracting referents.

There are also a number of edge cases that our method does not yet handle. For example, if an utterance has tens of impermissible interpretations and only one good interpretation, it may make less sense to assume that the good interpretation is correct than if there were only a few impermissible interpretations. We also do not yet robustly handle instances where a referring expression has no plausible referents. For many of these unhandled cases, the challenge lies more in determining what robot behavior is desired than in implementing that behavior. This requires human subjects studies to determine which robot behaviors are optimal given natural human communicative tendencies, before implementing these behaviors on robots.

Likewise, our system is designed specifically to handle *referential* ambiguity, which is a very common type of ambiguity in natural language, but there are other forms of ambiguity that may be morally relevant. For example, ambiguity may occur during pragmatic inference if a human says something like "can you punch Shaun?". Here, it may be unclear whether the human is asking the robot a yes or no question about its capabilities, or asking the robot to

actually punch Shaun (interpreting the utterance in the style of the conventionalized “can you pass the salt?"). In this case, it may be best to assume the non-problematic option, but ambiguity could also occur in other ways and in other parts of language processing, like speech-to-text (e.g., brake versus break). Work on these other forms of ambiguity will first have to show that the ambiguity in question can have morally relevant consequences, and that the current status quo in dialogue systems is inadequate for handling those consequences. Our approach would automatically handle these types of ambiguity if the components responsible for these facets of language processing (pragmatic inference and automatic speech recognition in these examples) generated and passed on sets of plausible hypotheses rather than the single “best" interpretation.

Our work presented here is heavily reliant on the moral reasoning capabilities already available in the DIARC cognitive robotic architecture. Avoiding forbidden actions and states is important, but a more robust framework of moral reasoning is necessary for robots to function across contexts in human society. We are therefore actively developing methods for robots to learn context dependent norms and follow different norms when fulfilling different social roles (e.g., waiter versus babysitter). As moral reasoning systems become more complex, so too must the language generation systems that explain them.

Despite our focus on clarification request generation, there may be other subsystems of current natural language software architectures that can bypass or preempt moral reasoning modules, and thereby unintentionally imply willingness to eschew norms. Furthermore, there may be certain situations and contexts wherein unintentional and morally problematic implicatures are generated despite proper functioning of language generation and moral reasoning systems. Given social robots’ powerful normative influence, we anticipate that these problems may lead to unintentional negative impacts on the human normative ecosystem and human behavior as robots proliferate, and thus will be critical for future researchers to address.

ACKNOWLEDGMENTS

This work was funded in part by Young Investigator award FA9550-20-1-0089 from the United States Air Force Office of Scientific Research.

REFERENCES

[1] Ronald C Arkin. 2008. Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture. In *Proc. 3rd ACM/IEEE Int’l Conference on Human-Robot Interaction (HRI)*.

[2] Kent Bach. 2006. The top 10 misconceptions about implicature. *Drawing the boundaries of meaning: Neo-Gricean studies in pragmatics and semantics in honor of Laurence R. Horn* (2006).

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

[4] Luciana Benotti and Patrick Blackburn. 2016. Polite interactions with robots. *What Social Robots Can and Should Do: Proceedings of Robophilosophy 2016/TRANSOR 2016* (2016).

[5] Timothy Brick and Matthias Scheutz. 2007. Incremental natural language processing for HRI. In *Proc. 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 263–270.

[6] Gordon Briggs and Matthias Scheutz. 2012. Multi-modal Belief Updates in Multi-Robot Human-Robot Dialogue Interaction. In *Proc. Symposium on Linguistic and Cognitive Approaches to Dialogue Agents*.

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

[8] Gordon Briggs, Tom Williams, and Matthias Scheutz. 2017. Enabling robots to understand indirect speech acts in task-based interactions. *Journal of Human-Robot Interaction* 6, 1 (2017), 64–94.

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

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

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

[12] Maartje Ma De Graaf, Somaya Ben Allouch, and Tineke Klamer. 2015. Sharing a life with Harvey: Exploring the acceptance of and relationship-building with a social robot. *Computers in human behavior* (2015).

[13] Juraj Dzifcak, Matthias Scheutz, Chitta Baral, and Paul Schermerhorn. 2009. What to do and how to do it: Translating natural language directives into temporal and dynamic logic representation for goal management and action execution. In *Proc. International Conference on Robotics and Automation.*

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

[15] Daniel Fried, Jacob Andreas, and Dan Klein. 2018. Unified Pragmatic Models for Generating and Following Instructions. In *Proc. Conf. of the North American Chapter of the ACL: Human Language Tech.*

[16] Felix Gervits, Gordon Briggs, and Matthias Scheutz. 2017. The Pragmatic Parliament: A Framework for Socially-Appropriate Utterance Selection in Artificial Agents. In *Proc. Annual Meeting of the Cog. Sci. Society.*

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

[18] Paul Grice. 1975. Logic and Conversation. In *Syntax and Semantics.*

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

[20] Ryan Blake Jackson, Sihui Li, Santosh Balajee Banisetty, Sriram Siva, Hao Zhang, Neil Dantam, and Tom Williams. 2021. An Integrated Approach to Context-Sensitive Moral Cognition in Robot Cognitive Architectures. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).*

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

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

[23] Ryan Blake Jackson and Tom Williams. 2019. Language-Capable Robots may Inadvertently Weaken Human Moral Norms. In *Proceedings of alt.HRI.*

[24] Ryan Blake Jackson and Tom Williams. 2019. On Perceived Social and Moral Agency in Natural Language Capable Robots. In *Proc. HRI Workshop on The Dark Side of Human-Robot Interaction.*

[25] Andrew F. Jarosz and Jennifer Wiley. 2014. What Are the Odds? A Practical Guide to Computing and Reporting Bayes Factors. *The Journal of Problem Solving* 7 (2014).

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

[27] Peter H Kahn, Takayuki Kanda, Hiroshi Ishiguro, Brian T Gill, Jolina H Ruckert, Solace Shen, Heather Gary, Aimee L Reichert, Nathan G Freier, and Rachel L Severson. 2012. Do People Hold a Humanoid Robot Morally Accountable for the Harm it Causes?. In *Proc. 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI).*

[28] James Kennedy, Paul Baxter, and Tony Belpaeme. 2014. Children Comply with a Robot's Indirect Requests. In *Proceedings of HRI.* ACM, Bielefeld, Germany, 198–199.

[29] Ross A Knepper. 2016. On the Communicative Aspect of Human-Robot Joint Action. In *Proc. RO-MAN Workshop: Toward a Framework for Joint Action, What about Common Ground.*

[30] Stephen C Levinson. 2000. *Presumptive meanings: The theory of generalized conversational implicature.* MIT press.

[31] Patrick Lin, George Bekey, and Keith Abney. 2008. *Autonomous Military Robotics: Risk, Ethics, and Design.* Technical Report. Cal. Poly. State Univ. San Luis Obispo.

[32] Bertram F Malle, Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano. 2015. Sacrifice One for the Good of Many?: People Apply Different Moral Norms to Human and Robot Agents. In *Proceedings of HRI.*

[33] Matthew Marge and Alexander I Rudnicky. 2015. Miscommunication Recovery in Physically Situated Dialogue. In *Proceedings of SIGdial.* Saarbrücken, Germany, 22–49.

[34] Tatsuya Nomura, Takayuki Uratani, Takayuki Kanda, Kazutaka Matsumoto, Hiroyuki Kidokoro, Yoshitaka Suehiro, and Sachie Yamada. 2015. Why Do Children Abuse Robots?. In *Proceedings of HRI Extended Abstracts (HRI'15 Extended Abstracts).* 2 pages.

[35] Matthew Richard John Purver. 2004. *The theory and use of clarification requests in dialogue.* Ph.D. Dissertation. University of London.

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

[37] Matthias Scheutz, Gordon Briggs, Rehj Cantrell, Evan Krause, Tom Williams, and Richard Veale. 2013. Novel Mechanisms for Natural Human-Robot Interactions in the DIARC Architecture. In *Proceedings of AAAI Workshop on Intelligent Robotic Systems.*

[38] Matthias Scheutz, Evan Krause, Brad Oosterveld, Tyler Frasca, and Robert Platt. 2017. Spoken Instruction-Based One-Shot Object and Action Learning in a Cognitive Robotic Architecture. In *Proceedings of AAMAS.*

[39] Matthias Scheutz, Bertram Malle, and Gordon Briggs. 2015. Towards Morally Sensitive Action Selection for Autonomous Social Robots. In *Proc. of RO-MAN.*

[40] Matthias Scheutz, Thomas Williams, Evan Krause, Bradley Oosterveld, Vasanth Sarathy, and Tyler Frasca. 2018. An Overview of the Distributed Integrated Cognition Affect and Reflection DIARC Architecture. In *Cognitive Architectures.*

[41] John R Searle. 1975. Indirect Speech Acts. *Syntax and Semantics* 3 (1975), 59–82.

[42] Noel Sharkey and Amanda Sharkey. 2010. The Crying Shame of Robot Nannies: an Ethical Appraisal. *Interaction Studies* 11, 2 (2010), 161–190.
[43] Reid Simmons, Maxim Makatchev, Rachel Kirby, Min Kyung Lee, et al. 2011. Believable Robot Characters. *AI Magazine* 4 (2011).
[44] Neil Stewart, Jesse Chandler, and Gabriele Paolacci. 2017. Crowdsourcing Samples in Cognitive Science. *Trends in Cognitive Sciences* (2017).
[45] Stefanie Tellex, Pratiksha Thaker, Robin Deits, Dimitar Simeonov, Thomas Kollar, and Nicholas Roy. 2013. Toward Information Theoretic Human-Robot Dialog. *Robotics: Science and Systems* 32 (2013), 409–417.
[46] Sean Trott and Benjamin Bergen. 2017. A Theoretical Model of Indirect Request Comprehension. In *Proceedings of the AAAI Fall Symposium Series on Artificial Intelligence for Human-Robot Interaction.*
[47] Peter-Paul Verbeek. 2011. *Moralizing Technology: Understanding and Designing the Morality of Things.* University of Chicago Press.
[48] Kazuyoshi Wada and Takanori Shibata. 2007. Living with Seal Robots – its Sociopsychological and Physiological Influences on the Elderly at a Care House. *IEEE Transactions on Robotics* 23, 5 (2007), 972–980.
[49] Willie Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, and Joe Woelfel. 2004. Sphinx-4: A flexible open source framework for speech recognition. (2004).
[50] James Wen, Amanda Stewart, Mark Billinghurst, Arindam Dey, Chad Tossell, and Victor Finomore. 2018. He Who Hesitates is Lost (...In Thoughts over a Robot). In *Proceedings of the Technology, Mind, and Society (TechMindSociety '18).*
[51] Tom Williams. 2017. A Consultant Framework for Natural Language Processing in Integrated Robot Architectures. *IEEE Intelligent Informatics Bulletin* (2017).
[52] Tom Williams, Saurav Acharya, Stephanie Schreitter, and Matthias Scheutz. 2016. Situated Open World Reference Resolution for Human-Robot Dialogue. In *Proceedings of HRI.*
[53] Tom Williams, Gordon Briggs, Bradley Oosterveld, and Matthias Scheutz. 2015. Going Beyond Command-Based Instructions: Extending Robotic Natural Language Interaction Capabilities. In *Proceedings of AAAI.*
[54] Tom Williams, Ryan Blake Jackson, and Jane Lockshin. 2018. A Bayesian Analysis of Moral Norm Malleability during Clarification Dialogues. In *Proceedings of COGSCI.* Cognitive Science Society.
[55] Tom Williams and Matthias Scheutz. 2016. A Framework for Resolving Open-World Referential Expressions in Distributed Heterogeneous Knowledge Bases. In *Proceedings of AAAI.*
[56] Tom Williams and Matthias Scheutz. 2018. Reference in Robotics: A Givenness Hierarchy Theoretic Approach. In *The Oxford Handbook of Reference*, Jeanette Gundel and Barbara Abbott (Eds.).
[57] Tom Williams, Daria Thames, Julia Novakoff, and Matthias Scheutz. 2018. "Thank You for Sharing that Interesting Fact!": Effects of Capability and Context on Indirect Speech Act Use in Task-Based Human-Robot Dialogue. In *Proceedings of HRI.*
[58] Tom Williams, Ravenna Thielstrom, Evan Krause, Bradley Oosterveld, and Matthias Scheutz. 2018. Augmenting robot knowledge consultants with distributed short term memory. In *International Conference on Social Robotics.* 170–180.
[59] Tom Williams, Fereshta Yazdani, Prasanth Suresh, Matthias Scheutz, and Michael Beetz. 2018. Dempster-Shafer Theoretic Resolution of Referential Ambiguity. *Autonomous Robots* (2018).