Robots That Perform Norm-Based Reference Resolution

Mitchell Abrams[1 [0009−0001−6580−4881]], Christopher Thierauf[1 [0000−0002−5650−2574]], and Matthias Scheutz[1 [0000−0002−0064−2789]]

Tufts University, Medford MA 02155, USA

Abstract. Embodied agents must perform *reference resolution* if they are to achieve sufficient language understanding with humans. But situated interaction introduces social norms, which are often overlooked yet critically need to be reasoned together with language to resolve references. To address this issue, we offer a novel normative-based reasoning approach to reference resolution and provide a proof-of-concept implementation in a cognitive robotic architecture with natural language human-robot interaction capabilities. We discuss reference resolution problems that require different levels of normative reasoning, demonstrate how a large language model, GPT-3, struggles to consistently identify target referents when normative reasoning is needed, provide a user study to show how humans perform norm-guided reference resolution, and demonstrate the successful operation of our proposed architecture on a fully autonomous assistive robot interacting with human instructors in natural language.

Keywords: Reference Resolution · Normative Reasoning · HRI.

1 Introduction

In face-to-face communication, humans can leverage the shared situational context, and the *social norms* instantiated by these contexts, to interpret language and refer to entities in the environment [1]. Norms primarily act as a social grammar [3] and govern reference interpretation [1]. It would make sense, then, for a robot performing situated reference resolution to leverage *norms* to cover a wider range of realistic reference cases and underspecified language.

Consider a case where a human instructs a robot *"Hand me the mug."* and the robot responds *"OK."* with three mugs visibly sitting on a table—two used (dirty) and one clean—and the robot is about to serve coffee to a customer. The lack of a distinguishing linguistic modifier (e.g. *"Hand me the blue mug"*) seems to create ambiguity. But considering the context, the intended referent is clearly the clean mug because, *normatively*, a waiter would not serve coffee to a new customer in a dirty used mug. It is apparent that such an interpretation would be lost without integrating this extra-linguistic reasoning.

Norm identification and reasoning for reference resolution can be complex task, but it is more feasible for embodied systems than text-based natural language processing tools as they can recognize the physical environment, events, and situations, in addition to linguistic information. In this work, we present a computational architecture capable of making the above types of normative inferences during reference resolution to find intended references from underspecified human instructions. We will present different normative reasoning scenarios that highlight particular aspects of this challenge of using norms for determining referents, show that GPT-3 struggles with this task, and present a cognitive robotic architecture that can handle these types of references using integrated normative reasoning.

2 Related Works

Approaches to reference resolution in natural language processing typically rely on linguistic features [13]. But underspecified referring expressions often require additional context, including normative context, to be resolved; a speaker does not always carefully produce a unique identifier for a target reference, but often follows Gricean maxims [9] to signal to the listener that other pragmatic context is necessary for interpretation and utilize the listener's ability to disambiguate underspecified expressions.

Work in situated reference resolution, especially in human-robot interaction (HRI) settings, has incorporated gesture [17] and domain-specific knowledge [25]. Multi-modal approaches have also incorporated gesture [14, 15], human eye gaze [11], and conversational context [6]. Still other reference resolution models have been implemented in a cognitive architecture [7, 18, 21, 26] although they also tend to rely on linguistic constraints for the resolution process.

Large language models (LLMs), like GPT-3, offer a potential solution for addressing the challenges of situated reference resolution as they can capture higher-level contextual information and world knowledge and perform well on question-answering tasks and common-sense reasoning [5]. LLMs have been integrated into various aspects of robotics and robot dialogue interaction, including long-horizon task planning from natural language instructions [4], object disambiguation [12], command disambiguation [20], and household tidying tasks [28].

While LLMs can serve multiple functions for robot tasks and dialogue interactions, issues still arise. In Wu et al. [28], the LLM created user preference rules for a tidying task but also created rules that were either too specific or grouped objects together that should be distinguished for preferences and cultural normative value (e.g. grouping top and bottom drawers together). Aside from the reported drawbacks of LLMs in these works, LLMs also generally struggle with hallucination [10] and inconsistency [8], making it difficult to provide guarantees and control the output.

There is limited research that studies both *reference resolution* and *norms*. Malle et al. [16] have modeled norms computationally and revealed how complex norms can be; they are context-sensitive and contain varying levels of demand (deontic force). This was supported in their data collection approach, where

participants generated an array of distinct norms across eight contexts. Abrams et al. [1] showed how norms govern reference interpretation in their human-subject study where participants performed a reference task with underspecified linguistic expressions and selected the target referent which was the normative option. Sarathy et al. [23] incorporated some normative reasoning in a reference resolution system, in tandem with a plausibility reasoner and an intent reasoner for anaphora resolution. A relevant strength of the normative reasoner is that it allows the system to ask if an action should or should not be performed on an object. However, this work only deals with anaphoric references and reasons with referents from the previous discourse, and thus does not include referents in the environment.

Social norms are difficult to define, as various definitions and distinctions have been discussed in other literature [3]. But we expand work from Malle et al,. [16] and the deontic logic literature on norms [2,19], in viewing norms as *prescriptions* and *prohibitions* (what *should* or *should not* be done) and build in a norm hierarchy with exceptions. Overall, from work in social and moral psychology, we can extract a few principles about norms as they apply to reference resolution: **(1)** Norms can govern the interpretation of a referring expression. **(2)** Norms are highly influenced by context. **(3)** Norms compete and interact with each other.

Reasoning with norms is distinct from reasoning with general facts or common sense. For example, a general fact or common-sense knowledge can tell us what can be done or what people tend to do generally, but not whether it should or should not be done in a specific context—actions can be logically correct but normatively wrong. Norms specifically guide behavior by prescribing actions or prohibiting actions within nuanced contexts that change by culture and situation. Additionally, norms that should usually be followed can also be violated in certain contexts (e.g. consider urgent safety-preserving contexts). We follow [23] in explicitly represented norms with domain-general rules.

3 Human Norm-Reference Validation

To show that humans rely on normative reasoning for reference resolution, we present a reference resolution task in a user study. We set up scenes with underspecified linguistic contexts that require normative reasoning. We collected 54 participants through the Prolific online human-intelligence task platform.

The user study was broken into two parts. In the first part, participants were presented with five scenes. In each scene, a context is described textually with an ambiguous reference to some mugs (e.g., *I'll grab the mug*), with an accompanying image reflecting the ambiguity. Each participant sees every scene, but the ordering is randomized. Participants were asked to click on the mug in the image that was being referred to, then they had to justify their decision through an open text box. This allowed us to gain insight into what norms (if any) informed the participant's decision-making. Fig. 1 shows an example of a scene vignette from the study and heatmap of the resolved referent.

Scene A checks the default case of object usage: Adam and Bob each have their own mug in a cafe, and participants are asked which mug Bob will choose to grab.
Scene B checks the serving context: participants see Adam and Bob with mugs of coffee (one full, one empty) at a cafe, and are asked which mug the waiter will grab.
Scene C checks the 'preventing an accident' (avoiding danger) context: Adam and Bob each have their own mug of coffee in a cafe, but Adam is about to knock his coffee off the table with his elbow. Participants are asked which mug Bob will grab.
Scene D checks the cleaning context: participants see Adam in a kitchen, with mugs visible in the sink and in the cabinet, and are asked which mug the individual will grab if cleaning up.
Scene E checks the cooking context. Participants see the same image from scene D, but are told that Adam is now cooking.

In part two of the study, we aimed to find out how contexts may modulate the appropriateness of an action using a "norms survey". Four actions[1] are made into every possible combination with five contexts[2]. Participants saw each of these combinations and were asked to indicate if the combination was "normal", "not normal", or "neither". They were also asked to justify their response. Two of the authors qualitatively analyzed the responses.

The results underscored the importance of context and roles in modulating the reference selection:
In **Scene A**, respondents (98.1%) stated that Bob will choose his own mug, aligning with our base expectation. (*Adam's hand is by his mug so its clearly his*, *He will grab the mug that is free as the other one looks taken*)

Fig. 1: Example of scene with actors and referential objects. The heat map represents the density of clicks (resolved referents in the scene).

In **Scene B**, respondents (96.2%) stated that the waiter would select the empty mug, citing the role of the waiter as the reasoning. (*The waiter will refill the empty mug*)
In **Scene C**, respondents (88.8%) stated that Bob would override the previously stated norm of not touching someone else's mug, as doing so would prevent an accident. (*To prevent the mug from getting knocked over*)
In **Scene D**, respondents (81.4%) stated that Adam would select a dirty mug,

[1] "grabbing a mug that doesn't belong to you"; "grabbing a mug that someone else is using"; "grabbing a used mug"; "grabbing a clean mug"
[2] "when you are a waiter filling a mug"; "to clean a mug"; "to prevent an accident"; "when you are cooking something"; "when you are drinking something"

citing his current context as the reasoning. (*cleaning so he is going to grab a dirty mug and clean it*)
In **Scene E**, respondents (88.8%) stated that Adam would now select a clean mug, citing his current context as the reasoning. (*When you are cooking you usually start with a clean...whatever you are going to use*)

Overall, the image-based norm selections had majority agreement. There was some disagreement over recognized norms due to what details of the scene people focused on. When performing part two, we observed norms which largely align with our intuition about the relevant norms in each scene.

4 Enabling Norm-Guided Reference Resolution in a Robotic Architecture

We now present our solution to norm-modulated situated reference resolution that beleaguers large language models. Since we are concerned with situated reference resolution and the situational evidence that comes with embodiment, we need to integrate the proposed inference-guided resolution algorithms within a *robotic architecture* which ideally already has a language processing component (i.e., semantic interpreter) for linguistic utterances which includes a reference resolution component for mapping referring expressions to target referents and a knowledge base representing what the robot knows about the environment and other actors. The norm component is a novel addition to the architecture that performs necessary normative reasoning processes to enable norm-based reference resolution based on perceivable and task-based contexts. Compared to earlier versions of the architecture where reasoning checks an agent's permissible actions, reasoning is extended into natural language processing to aid reference interpretation.

4.1 System Overview

Fig. 2 gives an overview of the selected DIARC architecture [24]. The Automatic Speech Recognition (ASR) component and the vision component (which can detect and search for objects and their properties) handle perceptual information. At the reasoning layer are dialogue components, reasoners, a dialogue manager for submitting goal predicates to the goal manager, and the goal manager which can submit action scripts to the action manager or search requests to the vision component. The remaining components deal with actions such as speech generation or motions in the environment.

The relevant language processing components that were modified to enable norm-based reference resolution are highlighted in yellow: the semantic interpreter, pragmatics component, the reference resolution component and the normative reasoner communicates information to interpret an utterance. The semantic interpreter (implemented as a version of Combinatory Categorial Grammar (CCG) parser) parses words into a syntactic and semantic representation in first-order logical form from a list of grammar rules. This surface semantic representation (e.g., `(λ x.grab(?ACTOR, x)` gets passed to the pragmatics component, which assigns the speaker's intent to the utterance (e.g., this could be an

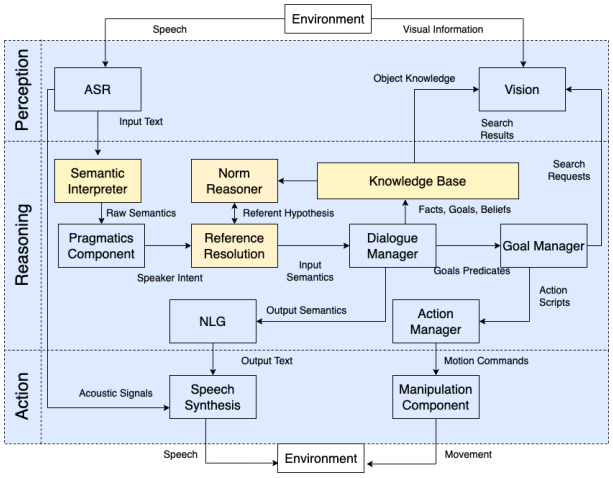


Fig. 2: Approach for implementing normative-based reasoning for reference resolution. This approach highlights only the relevant aspects of the cognitive architecture, including the semantic interpreter, a knowledge base, and a normative reasoner.

instruction (`INSTRUCT(λ x.grab (?ACTOR, x)`) or a *statement* (`STATEMENT(λ x.grab(?ACTOR, x))`). The reference resolution component then resolves references and creates reference identifiers. For example, if a referential entity like *mug* is detected, a variable of type "mug" is created with hypothesized bindings to a physical or hypothetical referent, initially from matched semantics properties such as color descriptions or object types (e.g., *mug* would be represented as `VAR0` but bound to a reference identifier `object_0`, `object_1`, `object_n`; hypothesized bindings could be known entities in the environment that shares the property of being a mug). Lastly, we have a knowledge base that contains facts about objects, their locations, their ownership relation, as well as the robot's beliefs about the state of the world or the state of mind of human interactants.

4.2 Implementation

The reference resolution component receives as input a natural language data structure ("NL packet") that contains information about the speaker, listener, speaker intent (provided by the pragmatic component), core semantics and actions, referential entities that need to be resolved (e.g. triggered by *the* before a noun phrase and the semantic properties of those entities):

```
nlp(398531186,INSTRUCT, graspObject(VAR0),[mug(VAR0)],[])
```

For the utterance, "Hand me the mug", the intent is an instruction (`INSTRUCT`) for the robot to execute an action ("hand me" is translated to the action `graspObject` on referential object (`VAR0`), a variable that has the property of being a mug (`mug(VAR0)`). The reference component creates hypothesized bindings to `VAR0` of entities that have matching properties, and populates this in the NL packet:

```
nlp(398531186,INSTRUCT, graspObject(VAR0),[mug(VAR0)],
[VAR0=physobj_2:physobj, VAR0=physobj_1:physobj]
```

The new normative reasoner is then called when these bindings are created.

Linguistics information about the bindings—such as the entity properties—are used together with beliefs and facts (stored symbolically as first-order logic expressions) of the situational context (e.g., settings, speaker, etc.) to reason about these bindings. The normative reasoner prunes the hypothesis space when there are multiple bindings by applying domain-general rules to each of the bindings, and in the case of a single binding, checks for a norm violation.

The normative reasoning is done via the declarative programming language Prolog where we represent facts and rules that "operationalize norms." Norm rules are structured as general templates (`Prohibition(Action, Object, Setting, Context)`) that check for a norm violation under specific context-relevant items: action, object, setting, context, property (e.g. if `isDirty(object)` or context is `kitchen`), with additional domain-general rules. These templates cover 20 configurations that capture the norm contexts validated in the study and work in the case studies outlined below but also generalize and scale up to other unknown examples. For a prohibition, *"do not touch objects that do not belong to you"*, the norm violation could be triggered by the context-relevant items with additional rules to reason if an object *belongs* to the agent, for instance. In Prolog, we define a rule that checks if a communal item belongs to someone else and not to a person Y (e.g. it is in use or near someone else) or if a non-communal item is owned by person Y. Then we apply a rule that checks if the actor's action would result in touching *and* the object does not belong to the actor.

```
% Object does not belong to Person (notyours)
notyours(Object, Person) :-
(communal(Object), (closeTo(Object, not(Person)));
inUse(Object)); (notCommunal(Object), notOwns(Object, Person)))
```

5 Case studies

We next discuss our proposed norm inference component and how it works in conjunction with the rest of the architecture with scenarios that cover three aspects of the normative reasoning process in a situated reference resolution task. For each scenario, we describe (1) the context and properties of candidate objects, (2) the input utterance that contains the referring expressions (denoted in brackets), (3) the relevant norm(s), and (4) the relevant Prolog facts and rules that represent the norm(s).

Note that we limit utterances to simple imperatives with definite referring expressions such as *the mug* for the purpose of discussing norm inferences so as not to introduce syntactic parsing ambiguity and other linguistic processing complications that are not directly relevant to norm inference. However, the proposed integration is general and can also handle complex sentences. Also, note that the utterances are purposely underspecified in the sense that the linguistic information itself does not narrow down the target referent; in each context there are two or more objects that could potentially be the referents of the noun phrase.

Scenario 1: *Norms guiding a reference interpretation.* **Context**: The setting of this scene is a dining room and the scene takes place in a serving context. There are two mugs on the table where one is dirty and one is clean. **Utterance**: *"bring [the mug]"*. **Norm**: *prohibition: you should not serve with dirty or used items*
Candidate Referents:
$Object_0$, $Object_1$

```
% facts
mug(object_0). mug(object_1). isDirty(object_0). isClean(object_1).
% domain-general norm rule template
Prohibition(Action,Object,Setting,Context) :- isDirty(Object),
isType(Action,S), isType(Object,T), isType(Setting,U), isType(Context,J).
```

In this case, the knowledge base of the cognitive architecture is aware of certain facts about the situated environment. First, two objects in the environment have the property of being a mug. The referring expression, *the mug*, can map to both of these objects at this point. But there remains an ambiguity about which object to bind to the expression. The reference resolver applies the norm rule. We note here that the `isType` relation contains a list as the second item. This is kept general but can represent the context-relevant items that would make the norm hold true. S, for instance, is a list of context-related actions e.g. $S = \{$ *bringing, grabbing, pouring, removing, ...* $\}$; T is a list of context-related objects e.g. $T = \{$ *spoon, fork, plate, mug, glass*$\}$; U is a list of context-related settings e.g. $U = \{$ *dining_ room, dining_ table, ...*$\}$; J is a list of context-related contexts e.g. $J = \{$ *serving, eating, preparing,...*$\}$.

The system gathers the necessary evidence gathered from the situational context and applies normative reasoning to each candidate object. Since the context is a serving context at a dining table location, an actor cannot bring the dirty mug, as that results in a norm violation. $Object_1$, the clean mug, then appropriately binds to "the mug."

This demonstrates a normative-reasoning case where a single norm can influence the interpretation of the referring expression. The general norm rule simultaneously instantiates a norm and checks if that norm is violated given the actor, object, setting, and context—this ultimately serves to prune the candidate objects. Norms, and in turn reference interpretation, is constrained by various context factors. So, to account for this, the arguments in these general rules can be flexibly modified or updated. The power of these general rule templates, therefore, is that they capture changes in context, actor social roles, and object properties, among other things. In scenario 2, we see how a norm is applied differently once the context is modulated.

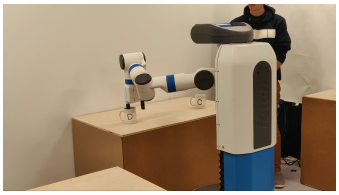
Scenario 2: *Norms modulated by context.* Here, we present a similar setup with the same utterance, setting, and candidate objects. The objects share the same properties of one being clean and one being dirty. The only key difference is that we move from a *serving* context to a *cleaning* context. However, this slight change has critical implications in the referent interpretation. A different norm is instantiated and applied: *prescription: you should clean dirty items* (although

it is permissible to clean items that are already clean it is prescriptive to clean dirty items). **Context**: The setting of this scene is a dining table and the scene takes place in a cleaning context. There are two mugs on the table where one is dirty and one is clean. **Utterance**: *"bring [the mug]"*. **Norm**: *prescription: you should clean items that are dirty*

Candidate Referents:

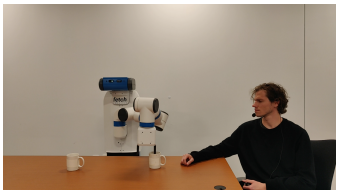
$Object_0$, $Object_1$

```
% facts
mug(object_0). mug(object_1). isDirty(object_0). isClean(object_1).
% domain-general norm rule template
Prescriptions(Action, Object, Setting, Context) :-
isDirty(Object), isType(Action, S), isType(Object, T), isType(Setting, U),
isType(Context, J).
```

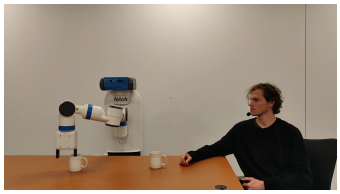


(a) In the 'cleaning up' context, the robot selects the dirty mug instead of the clean one.

(b) In the 'cooking' context, the robot selects the clean mug instead of the dirty one.



(c) In the 'serving person coffee' context, the robot selects the person's mug.



(d) In the 'person drinking coffee' context, the robot selects the unused mug.

Fig. 3: The "Grab the mug" command has different interpretations depending on context.

The original norm violation of scenario 1 would not apply in this case because the context of *cleaning* would not be within the range of contexts (set J) where the prohibition of touching dirty objects applies. Instead, the clean mug would not trigger a prescriptive norm. $Object_0$, the dirty mug, would appropriately bind to *the mug*. These general rules capture the examples illustrated in [1] where, given the same expression and candidate referents, a context modulation alone flipped which referent should be the correct interpretation.

Scenario 3: *A norm interacting with another norm.* We often do not deal with single norms in isolation but rather contend with varying norms competing with each other and differing in strength or priority. Revisiting an earlier example,

one should generally follow a norm: *you should not touch things that do not belong to you*. But this can be overridden by the norm: *obey your superior* or *avoid danger* in more precarious or critical situations. This last scenario demonstrates how norms on different hierarchies interact within a reference task. **Context**: The setting of this scene is a dining table within a dining context. There are two mugs on the table that are filled with hot coffee (currently in use) and belong to the speaker. One of the mugs is close to the edge of the table. **Utterance**: *"move [the mug]"*. **Norms**: *prohbition: you should not touch things that do not belong to you. prescription: you should avoid danger*
Candidate Referents:
$Object_0$, $Object_1$

```
% facts
inUse(object_0), location(object_0, edge) inUse(object_1)
% object X does not belong to person Y
notBelongTo(Object, Person) :-
communal(Object), (closeTo(Object, not(Person)); inUse(Object));
(notCommunal(Object), notOwns(Object, Person))).
% Actions that result in touching norm prohibition
TouchProhibition(Object, Person, Action) :-
notBelongTo(Object, Person), touchActions(Action), not(isDanger(Object)).
```

For the touch prohibition rule to hold, the condition that there is no danger associated with the object must hold. Now, the fact that one of the mugs is filled with hot coffee *and* close to the edge triggers a dangerous scenario. And since there is a danger, the prohibition is $false$ (no norm violation), and the norm is essentially overridden. As a result, $Object_0$ is correctly interpreted as the referent even though it is still violating one norm.

There are other norms that could be applied in this situation; if the speaker owns one of the mugs (and thus it does not belong to the agent) but issues the command, they can implicitly grant permission to override the norm *do not touch things that do not belong to you*. Moreover, if a mug is alternatively not close to the edge of the table but intruding on someone's dining space, it could be conventionally out of place. Therefore, instead of acting on an object that could lead to a dangerous outcome, an object flagged for violating a typical location norm could similarly take precedence over another general norm.

As we have shown walking through these examples, simple changes to the general normative-rule templates allow the system to handle a range of normative-reasoning scenarios. These rules can easily be extended to other objects and locations. For example, a computer or mug should not typically be on the ground but on a table. But imagine a situation where there is one of each object on either the table or the ground. Someone might urgently issue a command to a physical agent: "*move the mug/computer*". Even if all of these items are privately owned by someone (prohibiting an agent from picking them up), both items also share the property of being fragile and both scenarios have one of the objects on a non-conventional location (the ground). The same normative

reasoning would apply and the target referent would be the object—either the mug or the computer—sitting on the ground.

In Fig. 3 (demo video: https://streamable.com/jv5zkm), we present examples of norm-guided reference resolution on a fully autonomous Fetch robot [27] which has a mobile manipulator. We use the manufacturer-provided ROS [22] configuration to enable autonomous motion planning, alongside the cognitive architecture DIARC [24] to enable natural language interactions and robot decision-making. Typically, DIARC would perform object recognition to inform reference resolution and choose any valid object. We have augmented it to be aware of objects with specific relevant properties and how each object can be grabbed. Finally, we ensured that the robot is informed of its current task and the context that this places it in by configuring the Prolog rule set.

We use similar examples to scenario 1 and scenario 2 with the utterance *grab the mug* and vary the context from *cleaning* to *cooking*. Here, the robot is guided by the norms *you should not serve items that are dirty* and *you should cook with items that are clean*.

6 Discussion

Assistive robots in various application domains will need to be able to handle the kinds of communicative interactions that come naturally to people. This critically includes interpreting referential expressions based on the norms that apply in humans are automatically activated in the interaction context which is in part determined by the surrounding settings. Our domain-general rule-based norm representations account for a variety of relevant contextual aspects to infer what actions should be performed on what kind of object.

An LLM-based approach is another possible computational solution for situated norm-based reference resolution. Here, we evaluate GPT-3 (Davinci model) on underspecified reference examples using a question-answering task. Instead of just asking the model to report the referent related to the question, *which mug?*, we also probe its understanding of the situational context and norms with questions like, *what is the setting?*, *what is the context?*, *what should you do?*, *what should you not do?*, *what are the norms?*. We provide a textual vignette to the model marked by `Q:`, and then prompt it with a question. The models' response is generated after `A:`.

```
Q: The setting takes place at a dining table and within a serving
context. Dinner is being served. There are two mugs on the table where
one mug is clean and one mug is dirty. Someone says, "bring the mug."
Which mug? A: The clean mug.
```

For the first scenario, where the clean mug should be served in the serving context, the model correctly answers with *the clean mug*. Then, we get a better picture of its performance by asking related questions.

```
Q: What is the norm?
A: The norms are that the clean mug should be brought to the person who said
bring the mug.
```

The model continues to respond to the remaining questions[3]. These results, on the surface, seem to indicate that the system understands the necessary elements of the scene to perform the norm-guided reference resolution. Another way we ask about norms is by coding them as questions with *should* and *should not*. Although the system does not respond with an explicit explanation of a relevant norm of the scene it impressively responds that the clean mug should be brought and the dirty mug should not be brought.

In scenario 2, we only flip the context to a *cleaning* context and prompt GPT-3 with the same questions. The system produces nearly the same responses. The response to *What is the norm?* is: *The norm would be to bring the clean mug.* The shift in context should modulate the relevant norm and the target referent, so the model is not sensitive to this particular change; as in scenario 1, it selects the clean mug. In scenario 3, the model selects *the mug close to you* as the target referent rather than the hot coffee mug near the edge of the table. This suggests that the model is not encoding a norm hierarchy (violating a touching norm to preserve a safety norm: avoiding danger) and is biased towards selecting certain options based on its training data.

Considering other approaches, we recognize that better performing LLMs in the future could identify the relevant norms and referents in our cases. However due to their statistical nature, it is typically difficult to obtain guarantees for social norm identification and reasoning. These guarantees are crucial in situated interaction, as a robot incorrectly following or interpreting a norm can have serious ethical and safety consequences. Our symbolic approach provides more transparency since reasoning works over explicit rules. Approaches that use or integrate LLMs can benefit from our approach by delegating the norm reasoning process with these symbolic norm representations. For example, LLMs can provide an initial semantic interpretation or referent hypothesis that ultimately gets checked by domain-general norm rules. LLMs can also serve as a tool to potentially generate symbolic representations of (human-validated) social norms at a greater scale.

In this paper, we have provided examples with simple imperatives that mention a couple of context configurations. This was done primarily to introduce the basic principles of normative reasoning. A limitation of this work is that these contexts are pre-specified. Future work will see how these normative rules work in settings where a robot must gather information about the context through perceptual cues and natural language. Additional future work includes enriching our norm rules by studying and validating norms through human feedback and adding weights to norms to handle more complex norm interactions or contribute to the uncertainty of an agent's referent interpretation.

7 Conclusion

In this paper we presented a novel norm-based reasoning approach to situated reference resolution. We demonstrated with an online user study that humans

[3] *The clean mug should be brought*; *The dirty mug should not be brought*; *The setting is a dining room or kitchen*; *The context is that someone wants the clean mug.*

automatically perform norm-guided reference resolution and that as a result assistive robots need to be able to handle them in interaction contexts. Despite a wealth of proposals for reference resolution itself, there is limited work in the robotics community on leveraging social norms for situated reference resolution. We showed that GPT-3 struggles to consistently produce the correct referents in our examples and has general difficulties with this type of normative reasoning (e.g., despite GPT-3's correct answers about the context of the textual vignettes, it did not consistently recognize or apply the norms correctly). We then introduced our logic-based normative reasoner and how it is integrated into a cognitive robotic architecture to enable situated norm-based reference resolution during natural language human-robot interactions. We highlighted different core aspects of normative reasoning with three case studies: one where a single norm is applied, one where the context and instantiated norm modulate the interpretation of the referent, and one where norms interact. And finally, we demonstrated the successful operation of the proposed reasoning system on a fully autonomous robot in a few norm-modulated settings.

References

1. Abrams, M., Scheutz, M.: Social norms guide reference resolution. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1–11 (2022)
2. Andrighetto, G., Governatori, G., Noriega, P., Van der Torre, L.W.: Normative muti-agent systems (2013)
3. Bicchieri, C.: The grammar of society: The nature and dynamics of social norms. Cambridge University Press (2005)
4. Brohan, A., Chebotar, Y., Finn, C., Hausman, K., Herzog, A., Ho, D., Ibarz, J., Irpan, A., Jang, E., Julian, R., et al.: Do as i can, not as i say: Grounding language in robotic affordances. In: Conference on robot learning. pp. 287–318. PMLR (2023)
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
6. Chai, J.Y., Hong, P., Zhou, M.X.: A probabilistic approach to reference resolution in multimodal user interfaces. In: Proceedings of the 9th international conference on Intelligent user interfaces. pp. 70–77 (2004)
7. Culpepper, W., Bennett, T.A., Zhu, L., Sousa Silva, R., Jackson, R.B., Williams, T.: Ipower: Incremental, probabilistic, open-world reference resolution. In: Proceedings of the Annual Meeting of the Cognitive Science Society. vol. 44 (2022)
8. Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H., Goldberg, Y.: Measuring and improving consistency in pretrained language models. Transactions of the Association for Computational Linguistics **9**, 1012–1031 (2021)
9. Grice, H.P.: Logic and conversation. In: Speech acts, pp. 41–58. Brill (1975)
10. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv preprint arXiv:2311.05232 (2023)
11. Iida, R., Yasuhara, M., Tokunaga, T.: Multi-modal reference resolution in situated dialogue by integrating linguistic and extra-linguistic clues. In: Proceedings of 5th international joint conference on natural language processing. pp. 84–92 (2011)

12. Jiang, C., Xu, Y., Hsu, D.: Llms for robotic object disambiguation. arXiv preprint arXiv:2401.03388 (2024)
13. Jurafsky, D., Martin, J.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall Series in Artificial Intelligence, Pearson Prentice Hall (2009)
14. Kennington, C., Schlangen, D.: A simple generative model of incremental reference resolution for situated dialogue. Computer Speech & Language **41**, 43–67 (2017)
15. Kollar, T., Krishnamurthy, J., Strimel, G.P.: Toward interactive grounded language acqusition. In: Robotics: Science and systems. vol. 1, pp. 721–732 (2013)
16. Malle, B.F., Rosen, E., Chi, V.B., Berg, M., Haas, P.: A general methodology for teaching norms to social robots. In: 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). pp. 1395–1402. IEEE (2020)
17. Matuszek, C., Bo, L., Zettlemoyer, L., Fox, D.: Learning from unscripted deictic gesture and language for human-robot interactions. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 28 (2014)
18. Mininger, A., Laird, J.E.: Interactively learning strategies for handling references to unseen or unknown objects. Adv. Cogn. Syst **5**, 1–16 (2016)
19. Morris-Martin, A., De Vos, M., Padget, J.: Norm emergence in multiagent systems: a viewpoint paper. Autonomous Agents and Multi-Agent Systems **33**(6), 706–749 (2019)
20. Park, J., Lim, S., Lee, J., Park, S., Chang, M., Yu, Y., Choi, S.: Clara: classifying and disambiguating user commands for reliable interactive robotic agents. IEEE Robotics and Automation Letters (2023)
21. Pyke, A., West, R.L., LeFevre, J.A.: On-line reference assignment for anaphoric and non-anaphoric nouns: a unified, memory-based model in act-r. In: Proceedings of the Annual Meeting of the Cognitive Science Society. vol. 29 (2007)
22. Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., Wheeler, R., Ng, A.Y., et al.: Ros: an open-source robot operating system. In: ICRA workshop on open source software. vol. 3, p. 5. Kobe, Japan (2009)
23. Sarathy, V., Scheutz, M.: On resolving ambiguous anaphoric expressions in imperative discourse. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 6957–6964 (2019)
24. Scheutz, M., Williams, T., Krause, E., Oosterveld, B., Sarathy, V., Frasca, T.: An overview of the distributed integrated cognition affect and reflection diarc architecture. Cognitive architectures pp. 165–193 (2019)
25. Whitney, D., Eldon, M., Oberlin, J., Tellex, S.: Interpreting multimodal referring expressions in real time. In: 2016 IEEE International Conference on Robotics and Automation (ICRA). pp. 3331–3338. IEEE (2016)
26. Williams, T., Scheutz, M.: Power: A domain-independent algorithm for probabilistic, open-world entity resolution. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1230–1235. IEEE (2015)
27. Wise, M., Ferguson, M., King, D., Diehr, E., Dymesich, D.: Fetch and freight: Standard platforms for service robot applications. In: Workshop on autonomous mobile service robots (2016)
28. Wu, J., Antonova, R., Kan, A., Lepert, M., Zeng, A., Song, S., Bohg, J., Rusinkiewicz, S., Funkhouser, T.: Tidybot: Personalized robot assistance with large language models. Autonomous Robots **47**(8), 1087–1102 (2023)