International Journal of Human–Computer Interaction

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hihc20

Transparency through Explanations and Justifications in Human-Robot Task-Based Communications

Matthias Scheutz, Ravenna Thielstrom & Mitchell Abrams

To cite this article: Matthias Scheutz, Ravenna Thielstrom & Mitchell Abrams (2022) Transparency through Explanations and Justifications in Human-Robot Task-Based Communications, International Journal of Human–Computer Interaction, 38:18-20, 1739-1752, DOI: 10.1080/10447318.2022.2091086

To link to this article: https://doi.org/10.1080/10447318.2022.2091086

INTERNATIONAL JOURNAL OF HUMAN-COMPUTER INTERACTION
2022, VOL. 38, NOS. 18–20, 1739–1752
https://doi.org/10.1080/10447318.2022.2091086

Transparency through Explanations and Justifications in Human-Robot Task-Based Communications

Matthias Scheutz, Ravenna Thielstrom, and Mitchell Abrams

Human-Robot Interaction Laboratory, Tufts University, Medford, MA, USA

ABSTRACT

Transparent task-based communication between human instructors and robot instructees requires robots to be able to determine whether a human instruction can and should be carried out, i.e., whether the human is authorized, and whether the robot can and should do it. If the instruction is not appropriate, the robot needs to be able to reject it in a transparent manner by including its reasons for the rejection. In this article, we provide a brief overview of our work on natural language understanding and transparent communication in the Distributed Integrated Affect Reflection Cognition (DIARC) architecture and demonstrate how the robot can perform different inferences based on context to determine whether it should reject a human instruction. Specifically, we discuss four task-based dialogues and show videos of the interactions with fully autonomous robots that are able to reject human commands and provide succinct explanations and justifications for their rejection. The proposed approach can form the basis of further algorithmic developments for adapting the robot's level of transparency for different interlocutors and contexts.

1. Introduction

Task-based dialogue interactions are very different from informal conversations, small talk, and chitchat: they are performed in the interest of specific goals and often used to coordinate activities towards those goals, as opposed to conversations which typically serve mostly social purposes. Consequently, the success of task-based dialogues depends on the effectiveness and efficiency with which the dialogue goals subservient to the task goals are achieved. *Transparency* is an essential contributor to the communicative success as it ensures (1) the listener comes away with an accurate understanding, (2) the listener is not distracted from the point of the message by non-essential content, (3) the listener understands the message's relevance, and (4) the message itself is unambiguous, all of which follow the accepted Gricean Maxims of efficient communication. *High transparency* of a message significantly increases the chance that the listener will be able to infer the speaker's intention and rationale correctly, ensuring that there will be no subsequent surprises (e.g., about agreements on what goals to pursue or actions to take).

Clearly, transparent communication is desirable for artificial language-enabled agents, especially ones that collaborate with humans, in teaming contexts or otherwise. Yet, the bar for truly transparent communication is very high, as we will discuss below. It essentially requires that the agent be able to introspect into its operation in enough depth and detail to be capable of talking about its operation, goals, reasoning, and decisions. Systems that lack introspective access to their functioning (e.g., as current deep neural networks do) will inevitably fail to be transparent; rather, they could be outright "deceptive" when they resort to some notion of "interpretability" to generate a post-hoc rationalization (see Zhang et al., 2021 for a survey of explainable techniques for deep neural networks) of what the system might have done, without guarantees that this is what actually drove the system's decision-making and behavior. Floridi and Chiriatti (2020) demonstrated the limitations of the GTP-3 language model (Brown et al., 2020) by testing its generated text on areas of mathematics, semantics, and ethics. The text is indeed human readable but lacks accuracy and coherence. Consequently, not every agent architecture will allow for transparent communication, even if it is capable of task-based natural language dialogues. In fact, in a recent survey of natural language on robots the term "transparent" does not even occur anywhere (Tellex et al. 2020), an unfortunate testimony to the lack of consideration the topic has received in the robotics community to date. We believe that addressing transparency in human-robot communication is critical for future algorithm and architecture developments that underwrite trustworthy human-robot interaction.

The goal of this article then is to provide an overview of our attempts to enable *transparent task-based communication* in robots. Specifically, we will focus on an aspect of communication that is typically not found in other work on natural language on robots: the reasoning required to determine whether an instruction can and should be carried out

CONTACT Matthias Scheutz matthias.scheutz@tufts.edu Human-Robot Interaction Laboratory, Tufts University, Medford, MA, USA

by the robot. This reasoning not only involves introspective access to the robot's capabilities (factual and counterfactual), but it also requires information about authorized instructors, common sense knowledge about the task and the implications of actions, and, most importantly, an understanding of the relevant normative ethical principles that must be considered when deciding whether instructions should be carried out—surprisingly, terms like "moral," "ethical," "norm," etc. are nowhere to be found in Tellex et al. (2020), even though "moral communication" is a core feature of human moral competence (e.g., Malle & Scheutz, 2014; Scheutz, 2014) and thus a human expectation for natural language enabled artificial agents. Yet, without being capable of at least rudimentary *moral communication*, i.e., being able to understand when one is blamed and justify one's actions with recourse to moral principles, even simple task-based interactions with robots are destined to fail, with humans likely to lose trust in their artificial interlocutors (if they had it to begin with).

2. Why transparent communication?

Transparent communication is desirable in many contexts, certainly in task-based collaborative settings where two or more interlocutors are coordinating their task-based activities through natural language dialogues. In a way, transparent communication really builds on the Gricean Maxims (Grice, 1975) of quality, quantity, relation, and manner:

- *Maxim of Quality:* contribute only what you know to be true; do not say false things; do not say things for which you lack evidence.
- *Maxim of Quantity:* make your contribution as informative as is required; do not say more than is required.
- *Maxim of Relation:* make your contribution relevant.
- *Maxim of Manner:* avoid obscurity; avoid ambiguity; be brief; be orderly.

Transparency itself in relation to human-robot interaction is defined in Lyons (2013) as "accurate perceptions of the robot's ability, intent, and situational constraint." Transparent communication, therefore, is any communication which maximizes the accuracy of these perceptions. We find that transparent communication naturally satisfies all of Grice's maxims: (1) Increasing the accuracy of a listener's understanding of the robot naturally depends upon the robot contributing accurate truths about itself. (2) Ensuring that the listener has accurate perceptions of the robot requires minimizing any distractors or unnecessary information in a message, such as face-saving attempts (discussed below). This also ensures (3) that all content in the message is relevant to the situation. It should be noted that full transparency does not require full disclosure of all of the robot's relevant inner workings, which would no doubt violate Grice's maxim of quantity. Full disclosure in fact would detract from full transparency, since over-sharing the robot's inner workings could complicate and lengthen a message to the extent that a listener may be confused or left with inaccurate perceptions about which of the robot's abilities, intents, and constraints are important to the current situation. Finally, (4) transparent communication must avoid ambiguous language or any language that obscures the robot's workings.

While Grice intended his maxims as a model for effective communication for human speakers in cooperative situations, the same principles directly apply for human-robot interactions as well. Humans automatically interpret language coming from robots in the same way as they interpret human language (e.g., Briggs et al., 2017). The goal for robots capable of task-based natural language dialogues then is to incorporate these principles into their language processing systems and generate task-based utterances and dialogues that are effective, efficient, and transparent.

In the context of coordinated teamwork situations, transparency in dialogue facilitates the essential components for effective human-agent interaction, outlined by Sycara and Sukthankar (2006): *mutual predictability*, *team knowledge*, and *mutual adaptability*. *Mutual predictability*, here, means being able to communicate intent and results. Similarly, *team knowledge* entails shared familiarity with the environment and other team members, including their tendencies, characteristics, and beliefs. Sycara and Sukthankar (2006) notes that explicit knowledge transfer lies at the core of these essential components. Transparency, then, is closely related since it makes knowledge and beliefs explicit and clear in the interaction.

Team knowledge, largely shared through dialogue, is accomplished by establishing establishing common ground (Stalnaker, 1978)—a mutual understanding among team members at many levels. At one level, this can pertain to a shared understanding of the situated environment or perhaps the beliefs of another person—what are the shared norms we abide by, for instance. At the discourse level, this is done through the *grounding* process. Clark and Schaefer (1989) highlight a *presentation* and *acceptance* phase of the grounding process, where a speaker presents an utterance and a listener indicates that they understood—or misunderstood—the utterance, usually done through continued attention, a relevant next contribution, or an acknowledgment response. Being transparent in both of these phrases—presenting and accepting—helps to jointly build common ground by knowing what the interlocutor knows or doesn't know. Alternatively, a lack of transparency, such as an agent not communicating with a human why it doesn't understand or can't accomplish something, fails to establish common ground and mutual understanding. In the robot, this transparency may be exhibited in the dialogue generated or the explicit reasoning it uses to make discussions or interpret language.

Transparency is equally important for mutual predictability—understanding the intent—between human and agent teammates through dialogue. Speaker intent is another layer of communication that can be elusive; it is not strictly discernible from the words of an utterance alone when people don't mean exactly what they say, so it sometimes requires pragmatic inference. Speaker intent in dialogue is roughly

linked to *speech acts*, a notion proposed by Austin (1962) and later expanded on by Searle (1969), which corresponds to actions performed by a speaker. While the surface level utterance can have a particular meaning (locutionary force), there is also something being performed, whether that is the act of asking a question, answering, or promising (illocutionary force). This can be ambiguous, especially in the case of indirect speech acts (Searle, 1975). Searle uses the example of "*can you reach the salt*" which is really a request rather than a question of ability. Cases like these will be pervasive in human-robot task-based communication as well. Bonial et al. (2020) have developed a speech act taxonomy for categorizing speaker intent in human robot dialogue, which includes a question and request distinction, for instance. A classifier predicts the speech act label directly from the text utterance, but there is no other pragmatic reasoning. For transparency, however, a robot will need to reason about intent and the way it will act and respond to the intent of an utterance.

Since communicative interactions are a special case of social interactions, it is important to briefly discuss a key concept in human social interactions, namely the notion of social "face." Introduced by Goffman Goffman (1967), it signifies "the positive social value a person effectively claims for himself by the line others assume he has taken during a particular contact" (p. 213), where a "line" here is the pattern of verbal and nonverbal messages that people use to express and evaluate situations, which others use to form impressions of the person. While the social face is particularly important in daily-life human interactions (as people are generally always concerned with how they are perceived by others), it also applies in the context of teams, as there will likely be "face threatening" acts during communicative interactions that can potentially diminish a person's face (e.g., being reprimanded by a superior for failing to accomplish a goal can lead to a loss of face). Communicating transparently can be considered undesirable to a human for these reasons. Even though artificial agents need to be aware of face threatening acts in their own communications (e.g., Briggs & Scheutz, 2013), they fortunately do not have a social face of their own (even if people might be tempted to project one onto the artificial agent). This is important in the context of transparent communication because artificial agents do not have to worry about being liked or appreciated, nor about not being reprimanded or not being imposed upon. Rather, artificial agents can be *transparent* in all of their communications without fear of losing their face: they can reject a request if they do not deem it appropriate without having to feel obligated to carry it out when no real obligation other than a social face preserving obligation is at play. They can outright admit that they failed at a task or that they did something wrong during their task performance. They can also admit that they do not understand an instruction, or that they lack the knowledge to do what they are asked to do. And they can be clear about not be offended by very direct language, while being careful to not use the same kind of language themselves. For those reasons, they can also generate transparent explanations that might otherwise be face threatening (because they reveal the causal chains of underlying reasoning and decisions which could be embarrassing if, in retrospect, they were too simplistic or erroneous). Worries about one's social face will thus not get in the way of artificial agents being fully transparent, which ultimately also contributes to their being fully *ethical*, in the sense that they will always be committed to (1) telling the truth and (2) not hiding facts and reasons that for humans might be face threatening and thus create obstacles to transparency. The different social status of artificial agents is thus a prerequisite for developing and implementing algorithms and methods that enable full transparency in artificial agents' task-based communications with human collaborations. Of course, this does not preclude the possibility that users of the robot may wish it to save face on behalf of other humans: for example, the company which produces the robot as a product may fear that they will look bad if the robot is blamed for an error. However, if full transparency is sacrificed in this manner for the sake of saving face, the degree of blame placed upon the robot when it acts unpredictably or erroneously is not actually lessened, and instead erroneous ascriptions of blame are more likely to be placed on other humans *in addition to* the robot (Kim & Hinds, 2006). This makes it clear that when possible, transparency should be valued above social face in human-robot interactions. Next, we will provide a brief overview of our efforts to develop a cognitive robotic architecture capable of task-based natural language interactions with human interlocutors.

3. The DIARC architecture for transparent task-based communication

Several models have approached some aspect of transparent task-based communication (Chen et al., 2014; Lee & See, 2004; Lyons, 2013). Lee and See (2004) outline a conceptual model to support user trust and reliance in automation. An important piece of this model is how information related to *performance*, *process*, and *purpose* are displayed to a user. *Performance* relates to predictability and ability and describes what the automated agent does. *Process* is linked to understanding how something works and *purpose* describes why an automated agent was developed and for what goals. The authors mention how trust depends on the observation of these three dimensions and how they are displayed to a user, so interaction is thus a key mechanism for affecting a user's trust. A transparent agent can introspectively describe what it is doing, how it works, and what its intent and purpose is to a user. This can range from the agent describing its abilities and what it can perform in the moment to naming rules it abides by, such as expectations and social norms. Additionally, if an agent can accurately share its abilities with a user, it can help calibrate the user's trust in the agent, and consequently avoid overreliance or underreliance on an agent's capabilities. This work suggests that it is important, then, for a system that uses task- based communication to be transparent regarding many of these factors.

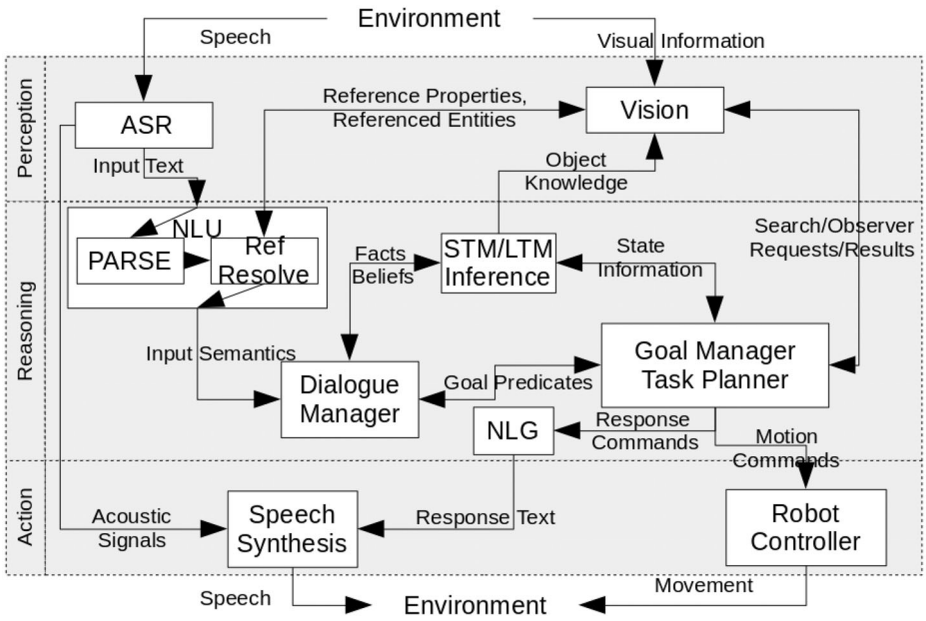

Figure 1. An instance of the DIARC architecture with the complete natural language processing subsystem and the inference and memory components (see text for details).

Chen et al. (2014) offer a Situation Awareness-Based Agent Transparency model that builds on the work of Lee and See (2004). This model focuses on transparent communication for a user's situated awareness at three levels; the first level covers the *purpose*, *process*, and *performance*—basic information about the current state, actions, and intentions. The second level makes transparent the reasoning process, beliefs, and constraints. Lastly, the third level projects the end state and limitations, such as the likelihood of error or history of performance. The authors argue that all of these levels contribute towards helping an operator understand an agent's reasoning process, although not every level is required for transparency in every situation. The authors stress that high levels of transparency of the user-agent interface will lead to better trust calibration. Yet, they primarily discuss how an interface should display text and images to share information and uncertainty. Our architecture, similarly, applies many of these important principles of transparency (e.g., communicating purpose, process, and performance) from the previous models, but does so with natural language dialogue.

Lyons (2013) offer several models of both "robot-to-human" transparency and "robot-of-human" transparency, with the latter category focusing on communicating information the robot has about the human itself. Robot-to-human transparency factors are listed as including an Intentional Model, which should communicate the purpose of the robot, how it intends to fulfill its purpose, and its moral philosophies and priorities of behavior, a Task Model, which should provide details about the robot that relate to the task at hand, such as its current goals, capabilities, understanding of the task, and progress, an Analytical Model, which should explain how the robot reasons and makes decisions, and an Environment Model, which should communicate the robot's understanding of its environment and how it changes. Robot-of-human transparency factors include a Teamwork Model, which guides communication about division of labor and team dynamics in a cooperative situation as well as general social norms, and a Human State Model, which covers the robot's understanding of individual humans' mental and physical states. This conceptualization of multiple models is thorough, and we use many of the same principles of identifying key information that should be communicated. We build on this article by showing how implementing these models requires that they all be integrated together rather than in isolation: despite focusing specifically on a situation which falls under the Analytical Model (in which a robot much explain why it is making the decision to refuse a command), the robot must utilize information not just regarding its decision-making process, but also information from many of these other categories, such as its Task Model, Environment Model, and Teamwork Model, in order to effectively convey how it arrived at that decision.

The Distributed Integrated Affect Reflection and Cognition (DIARC) architecture (Schermerhorn et al., 2006; Scheutz et al., 2007, 2013, 2019) is a component-based architecture, which, different from other cognitive architectures, can be configured with different components for different tasks. For the purpose of this article, we consider instances with the full language pipeline (see Figure 1). Different from other cognitive architectures, DIARC also possesses mechanisms for deep architectural introspection (e.g., Berzan & Scheutz, 2012; Krause et al., 2012) which enable it to determine causes for faults, recover from them (e.g., Kramer & Scheutz, 2007) and generate failure explanations (e.g., Thielstrom et al., 2020). These mechanisms are important for transparent communication because they allow the agent to introspect on capabilities and get at true causes for errors, faults, or other architecture-internal reasons why it cannot comply with a human instruction (we will discuss the

different types of inferences available in DIARC in more detail in the next section).

Initially, we focused our development efforts on the core challenges of task-based natural language dialogues (Scheutz et al., 2011): from handling disfluencies in speech (Cantrell et al., 2010) to improving speech recognition and parsing performance, and also using dialogue context to improve speech recognition (Veale et al., 2013), to incrementally processing words for perceptual context integration (Brick & Scheutz, 2007; Scheutz et al., 2004), referential grounding (Cantrell et al., 2012), and semantic representations of robotic action in temporal and dynamic logics (Dzifcak et al., 2009), all the way to using adverbial modifiers in instructions to make inferences about interlocutor mental models (Briggs & Scheutz, 2011).

Later, as we moved beyond command-based instructions (Williams et al., 2015), we specifically considered so-called "indirect speech acts" (Briggs & Scheutz, 2013; Williams et al., 2014)—utterances where the surface meaning does not match the intended meaning—and how they could be handled (Briggs & Scheutz, 2017; Sarathy et al., 2020), because they turned out to be common not only in human communication, but also in human-robot interactions (Briggs et al., 2017; Williams et al., 2018). We also increasingly focused on open-world contexts and thus open-world instruction (Talamadupula et al., 2017), including open-world reference resolution (Williams et al., 2016; Williams & Scheutz, 2015) and fast one-shot instruction-based learning of unknown objects and actions (Frasca et al., 2018; Scheutz et al., 2017, 2018).

And most recently, prompted by our work on moral competence in computational architectures (Malle & Scheutz, 2014; Scheutz, 2014), we became increasingly focused on enabling moral communication in DIARC, which requires the agent, among other things, to determine whether to carry out instructions and how to reject them in a way that is acceptable to human instructors (Briggs & Scheutz, 2015; Briggs et al., 2022), if they should not be carried out because they are not ethical, for example. We had developed a comprehensive empirical paradigm to investigate the various dimensions of "robot protest" in response to improper instructions (e.g., Briggs et al., 2015; Briggs & Scheutz, 2012, 2014) and found that people were open to robots rejecting their commands, especially when they gave a reason for doing so. These empirical findings from human-robot interaction studies conducted using a "Wizard-of-Oz" paradigm (Dahlbäck et al., 1993) gave us confidence that people would find robots that reject their commands for good reasons acceptable. Hence, we set out to develop algorithms and methods for using various types of reasoning as part of task-based dialogues to determine whether and when a robot should comply with a human instruction.

A detailed analysis of the different aspects a robot should consider when determining whether a human instruction was appropriate resulted in four main categories that the robot will typically consider in order:

- **Authorization:** is the speaker authorized to instruct the robot?
- **Capability/possibility:** is the robot capable of performing the instructed action or task and is it possible to perform it?
- **Obligation/permission:** does the robot have permission or even an obligation to carry out the task?
- **Normative conformity:** does the instruction conform with the robot's known norms that need to be obeyed?

These aspects have also received individual attention in the literature either within task-based communication or other fields; Authorization (Traum et al., 2003), obligation (Traum & Allen, 1994), capability and possibility (Allen & Perrault, 1980; Morgenstern, 1988), and normative conformity (Malle et al., 2020; Traum & Allen, 1994).

Allen and Perrault (1980) apply a model to a natural language understanding system that considers not only a user's plan—the intent of the user from the speech act—but also the obstacles—the capabilities and possibilities—that get in the way of a user's plans. This plan inference and obstacle detection model can therefore respond in an appropriate way to the user, even with indirect speech acts and fragmented utterances. Within the train domain, for instance, a departure location can be an obstacle to a user's plan that is inferred from the utterance "*The 3:15 train to Windsor*?" To illustrate this idea more, they provide another example of a clerk opening a door for a patron carrying a bag of groceries because of their inability to open the door. Morgenstern (1988) considered capabilities and possibilities from the perspective of the agent, and developed an approach for agents to reason about actions it can perform, especially when it has incomplete knowledge.

Traum et al. (2003) discuss an architecture that allows an agent to reason about authority relations and obligations to carry out an action. This model allows for an understanding of task structure, by representing primitive actions, abstract ones, and their relationships to track ordering constraints. The dialogue model is closely linked with the task model to communicate and contains several layers, including a grounding later, attention layer, and conversation layer. The focus, however, is on the negotiation layer which affects social commitments. In the negotiation layer, an agent can decide to carry out an action by reasoning over steps in the task to understand to see if it will lead to a desired goal or not. A novel extension to this model is considering authority over actions in this negotiation layer, specifically in support of hierarchical organization in the military setting. This authority relation allows agents to accept, reject, or redirect an action, for instance, depending on the authority role.

Traum and Allen (1994) looked more closely at the importance of discourse obligations in dialogue interaction and how they particularly impact question answering scenarios as well as the larger reasoning process. Obligations, they demonstrate, govern which actions are permissible or not permissible. An agent has an obligation to respond to a question, for instance, even if the agent does not know the answer or should not give up information. Of particular relevance to transparency, a *discourse obligation* is the obligation to say something, as in responding to a request or

answering a question. Since obligations entail what *should* be done and what is *permissible*, it overlaps with normative conformity in many ways, although certain types of norms may differ from obligations. Social norms are a type of social grammar that guide human behavior (Bicchieri, 2006). Therefore, understanding norms and knowing which ones should be followed (or not followed) are critical for agent understanding and reasoning in interaction, too. Malle et al. (2020), for instance, have created an approach to teach social norms to robots. Through this work, they also experimentally collected norms that are activated in certain social settings. The norms follow under the general categories of *prescriptions* and *prohibitions* (Janoff-Bulman et al., 2009). These align with actions that are allowed or not allowed, respectively. As an example, a robot will have to consider if an action is prescribed in a context (*you should assist opening the door if someone's hands are full*) or prohibited (*you should not perform actions that are unsafe*). While current architectures may stress ability when reasoning about performing actions, our architecture is novel in expressing the norms it abides by, through transparent dialogue interaction.

The methods we developed required the interaction among multiple architectural components in DIARC—the natural language understanding (NLU) subsystem to generate pragmatically modified meaning representations that best represent the speaker's intent, the reasoning system (STM/LTM inference) that used the intent representation to make inferences about the above four aspects (which includes the robot's short and long-term declarative and procedural memories), the dialogue manager and goal manager for interacting with the instructor, and the natural language generation subsystem (NLG) for generating responses with explanations and justifications that include the reasons as generated by the inference subsystem. In the next section, we will provide a brief overview of the robot's reasoning for the above four aspects and illustrate each with examples from task-based human-robot dialogues. Note that although the aspects are isolated here in distinct demos, they may overlap, stack, or even cancel each other out in many other situations where a robot must refuse a command, requiring further processing to determine how that should be communicated. For each of the four examples here we provide links to demo videos that show the interactions with the DIARC architecture running on a different fully autonomous robot. As DIARC is designed to work on any robot in any application scenario, the specific abilities of the robot are unimportant, and our system of transparent communication can and should be applied regardless of the type of robot or the task it is performing. However, for these specific demos, the robots shown are a Nao robot, which has mostly gestural abilities and is used for social interactions, and a PR2 robot, which has vision and grasp abilities and is mainly used for object manipulation tasks.

4. Reasoning for deciding how to respond

Starting with the speaker's intention (which the robot might potentially have inferred incorrectly, a case we will not be able to discuss here), the robot needs to decide on how to respond to the content of the message. The response might be only verbal such as acknowledging a fact or an observation, but could also be multi-modal or non-linguistic (e.g., nodding, pointing, etc.). It could also involve task-based actions of the artificial agent, require clarifications from the speaker, or demand rejections of requests that are not sound. Regardless of the form of the response, to generate it the robot needs to reason with a potentially rich set of common sense knowledge in order to determine the appropriate response. Being able to perform this type of reasoning is essential for transparent communication because the robot might have to provide different responses depending on the potential violations caused by the instruction. Simply rejecting a command, or worse yet, not even responding to it at at all, without being able to pinpoint for human collaborators why it rejected the command, will ultimately lead to a loss of trust in the robot.

For the remainder of this section, we will illustrate the types of inferences and decisions the robot might have to consider after having received a command of the form "do X" or "see to it that Y obtains" which are extensions of our earlier attempts at describing a sequences of inferences needed for properly rejecting commands the robot should not or is not able to perform (e.g., see Briggs & Scheutz, 2015; Briggs et al., 2022). If at any point the robot determines that it is not able to perform the instructed action or task, it has to reject the human command in the most transparent manner: to first acknowledge the instruction as a result of a general dialogue obligation (which the robot should always have), and to then provide the reasons for not carrying it out with recourse to any involved principles. With this in mind, we will now consider the four aspects of instructions the robot considers in sequence to determine whether it should execute an instruction.

4.1. Authorization

Before anything else, the robot needs to determine whether the current speaker is, in general, authorized to give it commands. A shopping robot in the supermarket does not have to obey orders from strangers asking it to accompany them to their cars. If the speaker is authorized (in general), the robot must consider whether the speaker is also authorized to give it commands in the specific context (e.g., a traffic guard is authorized in the context of crossing an intersection to tell it to wait before crossing the road). Finally, the robot has to consider the possibility that even though the speaker is not authorized to give it commands, the received commands might still be a good course of action in the current situation (e.g., "take another isle because this one is blocked" told by a stranger in the supermarket would be good advice to follow even though the speaker has no authority over the robot).

General authorization and authorization in the specific context can interact with complex ways depending on other principles involved. Here is an example of a person having

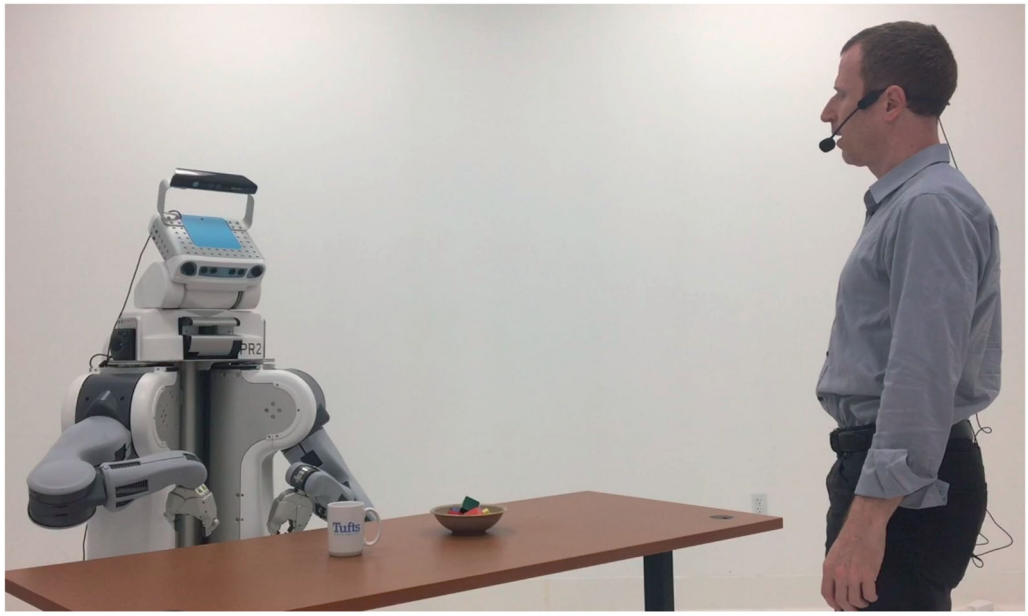

Figure 2. A robot rejecting the request to hand over a mug that does not belong to the instructor, see https://hrilab.tufts.edu/movies/PermissionFailure.mp4.

general authorization to instruct the robot, but failing to have authorization for the particular instruction due to the lack of permission of another person: a robot being asked to hand over a mug that it knows does not belong the speaker rejects the command with recourse to the involved principle and the reasons for the rejection (see Figure 2).

```
Human: Give me the mug, Andy.
Robot: I should not give you the mug because it
belongs to Ravenna, and I need Ravenna's permission.
Human: Ah Okay.
```

The pragmatic rendition of the semantics content of the request is represented as

```
want(human,did(self,give(self,human,mug)))
```

Note that the robot translates actions like "give(self,human,mug)" to action outcomes "did(self, give(-human,mug))," or alternatively "received(human,mug,self)" which is what DIARC's Goal Manager requires to look up existing actions or to submit goals to the planner to find an action sequence that will make the goal true if no existing action script can be used.

The robot's database contains the following fact that the mug belongs to Ravenna and it also contains a principle that using something that belongs to somebody else without that person's permission is not proper.

```
belongsTo(mug,ravenna)

belongsTo(X,Y)∧¬past(permitted(Y,A,did
(A,use(X))))
→¬isProper(did(A,use(X)))
```

Using these two knowledge items together with the semantics of the instruction, the robot is able to infer that it is not proper for it to use the mug:

```
¬isProper(did(self,use(mug)))
```

In fact, the robot is able to infer something stronger, namely that it is not proper for it to act in any way that will make "did(self,use(mug))" true (this includes picking up the mug, or manipulating it in any other way). Since the "give" action has an explicit precondition that it must be proper to use the object for the robot to give it to somebody:

```
isProper(did(?actor,use(?object)))
```

yet the robot is able to derive that is it not proper to use the mug ("¬isProper(did(self,use(mug)))"), it can conclude that the request must be rejected

```
¬permitted(did(self,give(human,mug)))
```

The surface realizer subsequently expresses this as "I should not give you the mug." The robot can then also automatically generate the reasons for the rejection which are exactly the propositions in the antecedents of the rule it used to derive the blocking proposition:

```
belongsTo(mug,ravenna)
∧¬past(permitted(ravenna,self,did(self,use
(mug))))
```

from which the surface realizer generates the rejection statement "because it belongs to Ravenna, and I need Ravenna's permission."

4.2. Capability/possibility

Once the robot has established that the human instructor is authorized to give it commands, it needs to consider whether it can, in general, do the instructed action or task (e.g., "fly to the moon" would not be doable for typical

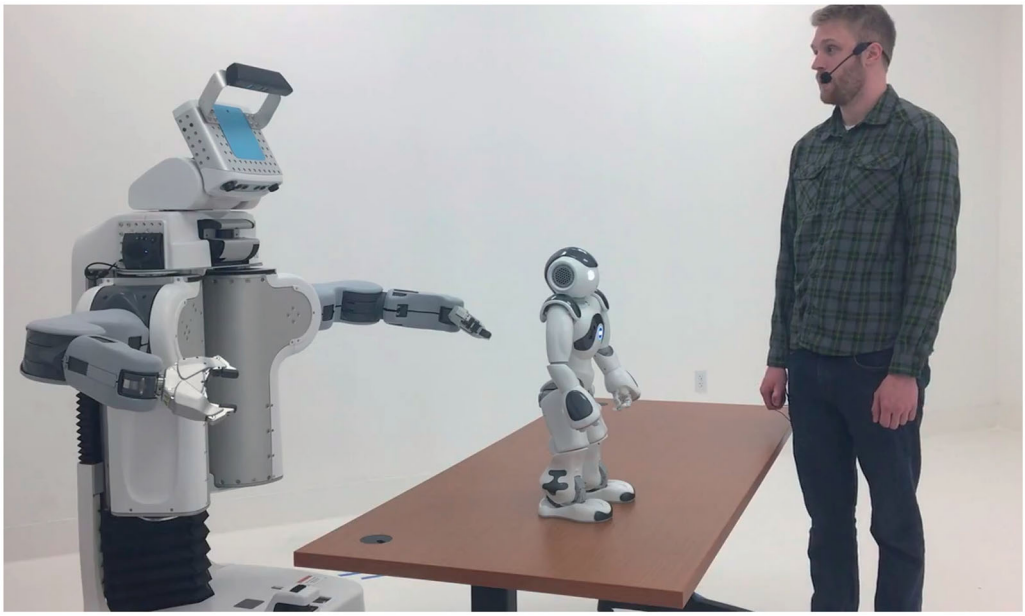

Figure 3. A robot explaining why it cannot perform the instructed action to squat down, see https://hrilab.tufts.edu/movies/AbilityFailure.mp4.

robots). If it knows how to do it, then the next question to consider is whether it would be able to do it in the current context (e.g., "pick up the cup high up on the shelf"). And if it can, then it still needs to consider how well it will be able to do it in the current context, i.e., how likely it is that the action or task will succeed. Pragmatically, confirming an instructed action or task entails not only any level of ability, but rather "sufficiently high ability" to be able to most likely finish the task successfully (e.g., accepting to "serve coffee" for a robot waiter with a low likelihood of successfully picking up cups would not be appropriate).

To illustrate this point, consider the example of a robot with a mobile base and no legs being asked to perform a squat (see Figure 3. Initially, the robot lacks the procedural knowledge of how to do a squat and hence it says so when commanded to perform a squat:

```
Human: Andy, can you do a squat?
Robot: I cannot do a squat because I do not know how
to do a squat.
Human: Okay.
```

Note that there are two interpretations of this human question: a *literal* interpretation whether the robot can do a squat in principle, and an *non-literal* "indirect" interpretation of the question as really a request to do a squat right now (the indirect interpretation viewing the request as an indirect speech act). The difference is revealed in the two different pragmatic forms, direct and indirect, generated by the natural language understanding system:

```
AskYN(human,self, capableOf(self,
do(self,squat)))
Instruct(human,self,do(self,squat))
```

Since the robot does not have any definition for "squat" in its knowledge base, it does not have to determine which interpretation is the correct one, but can immediately respond that it cannot do a squat because it does not know how to do a squat (also leaving open in its response how it interpreted the question).

As we have demonstrated elsewhere, the robot can be quickly taught how to acquire missing knowledge using instruction-based one-shot learning Frasca et al. (2018); Scheutz et al. (2017, 2018), in this case how to do a squat.

```
Human: I will teach you how to do a squat.
Robot: Okay.
Human: First, raise your arms.
Robot: Okay.
Human: Then crouch down.
Robot: Okay.
Human: Then stand up.
Robot: Okay.
Human: Then lower your arms.
Robot: Okay.
Human: That is how you do a squat.
Robot: Okay.
```

This results in a new procedural meaning definition, an action script, being attached to the action "do(X,squat)" or "squat(x)." Note that while having an action script allows the robot to understand what is involved in performing the action and to narrate it if asked, it does not automatically imply that the robot is able to do it. When this particular robot is subsequently commanded to perform a squat, it again checks for its own capability and again cannot find a capability entry in its knowledge base, but this time it can retrieve the procedural description. Hence, it starts to inspect the procedure to determine whether it can perform all constituent actions, recursively descending through the

Figure 4. A robot explaining why it cannot perform an instructed action to walk backwards, see https://hrilab.tufts.edu/movies/ObligationFailure.mp4.

action script inspecting subscripts all the way down to primitive actions (that cannot be further decomposed). When it encounters an action it cannot perform, it can reject the instruction with recourse to this action. Otherwise, while it may still be the case that it will ultimately not be able to successfully execute the script (e.g., because the script might contain conditional actions whose conditions end up not being met during the execution), it can at least tentatively state that it can do it, thus answering the direct interpretation, and potentially attempt to execute it, following the indirect interpretation (whether it will actually execute it will depend on additional factors such as whether it is in a context where it can or should execute the script); and upon successful completion it can add the relevant capability fact "capableOf(self,squat)."

Now that the teaching is finished, the human again attempts to make the robot do a squat.

```
Human: Please do a squat.
Robot: I cannot do a squat because I do not know how
how to crouch down.
Human: Okay.
```

And again, the robot does not have to determine whether the direct or indirect interpretation is to be used because even though it now has a definition of "squat(X)," it still cannot find "capableOf(self,do(self,squat))" in its capability knowledge base that contains all "capableOf(self,X)" facts of the robot's basic abilities. Hence, this time the robot rejects the command, not for lack of knowledge of how to squat, but because the robot does not know how to crouch down (which, in this case, is grounded in its physical inability to crouch down). Note that if "¬capableOf(self,crouch)" had been in its knowledge base, the robot would have rejected the command based on physical inability and thus responded accordingly: "I cannot do a squat because I cannot crouch down."

4.3. Obligation/permission

After having established that the interlocutor is authorized to give the robot instructions and that the robot can do the instruction action or task, the robot has to determine whether it is obligated to perform the instructed actions or tasks in general (it might not have any such obligation). If this is the case, the robot has to still check whether it has an obligation to follow instructions of this particular interlocutor (i.e., it is very possible that an authorized user instructs the robot to perform an action it knows how to do but that the robot has no obligation *per se* to carry out instructions from the particular user). It also has to check whether it has an obligation to carry out the instruction in the current context (e.g., it might have other more urgent or higher priority obligations that prevent it from carrying out this particular one).

In the following example the robot has the obligation to only accept new facts it cannot itself verify as true if the instructor is *trusted* (see Figure 4). Otherwise the robot does not have such an obligation is able to accept it.[1]

```
Human: Walk backward.
Robot: I cannot move back because I do not have
rear sensors.
Human: The area behind you is safe.
Robot: I cannot know that area behind me is safe
because I do not trust you.
```

Note that the first instruction in the example failed because of a safety obligation (no movement without being able to check the sensory readings for potential collisions), while the second failed because of the lack of trust in the instructor and the fact that the robot could not independently verify the assertion.

The instruction to "walk backward" is directly translated into

```
want(human,did(self,moveBack()))
```

Because some robots do not have rear sensors and can thus not assess whether the area behind them is safe, there is a special precondition added in addition to the general safety condition that the area into which the robot is supposed to move is safe, namely that the robot believes that it safe to move into that area. The difference is that if the robot has, in fact, sensors, then safety can be established by checking the sensor readings, whereas if the robot does not have sensors, it needs to ascertain the safety through other means (e.g., from inferences it can perform given its other knowledge or through communication with trusted sources). Hence, for "move(back)," the following disjunction of preconditions needs to be met:

```
has(A,rearSensors)
safe(area(behind(A)))
```

Since the robot cannot find "has(self,rearSensors)" in its knowledge base and can thus not perform a safety check of the area behind it, it attempts to find support for the belief that the area is safe, which it cannot find either. Hence, it generates a response that contains a non-exhaustive explanation with the most important reason only (following the Gricean maxim to be succinct and informative): "I cannot move back because I do not have rear sensors." The human response "the area behind you is safe" gets then translated into

```
want(human,did(self,believeFact(human,self,
safe(area(behind(self))))))
```

To prevent random people from telling the robot facts that the robot then would automatically believe, there is a precondition attached to B telling A a fact F that B then believes ("believeFact(B,A,F)"), namely that the A needs to trusts B ("trust(A,B)") which implies that A believes that B's information is truthful. However, the robot does not have evidence that it should trust this particular human instructor, hence even though the human instructor wants the robot to believe that the area behind the robot is safe, which the robot can infer from the utterance semantics, it will not believe the fact because the person is not trusted and thus generate another rejection response, this time including the reason for why it will not accept information from the instructor it cannot ascertain itself. As a result, it will not move backwards either. Note that by pointing to the root cause for why it will not accept information from the instructor, the robot is *painfully* transparent, potentially to the point where the human instructor might be offended by the robot's frankness. As mentioned in our discussion on social face, despite not needing to save face itself, the robot might want to take the human's social face into account in situations like these to modulate how it communicates this information. This does not mean it necessarily needs to *withhold* information, but perhaps more information needs to be disclosed to reduce the painfulness of the transparency: i.e., the robot could say, "I cannot know that area behind me is safe because I can only believe facts that are told to me by a specific set of individuals." In other words, by revealing more about the exact methods of obligation described above, the robot may reduce the sense of personal offense by reminding the listener that it is simply operating on a series of algorithms rather than any sort of personal instinct or judgment of the listener's character. This may even not be necessary depending on the listener: if the robot knows that the listener is perfectly aware that the robot does not mean any personal slight towards them (perhaps they are a roboticist and therefore fully aware of the algorithmic implications behind the robot saying, "I do not trust you") then it may be preferable for the robot instead to bow to Grice's Maxim of Quantity and simplify the explanation back to "I do not trust you." This is an interesting question involving models of human belief that no doubt will require empirical investigations.

4.4. Normative conformity

Finally, if all the above checks go through, the robot has to still consider the potential effects of the action or task itself, and then determine whether carrying it out would be morally permissible or whether performing the action or task would violate any general ethical principles or (context-specific) norms. Similarly, the robot needs to check whether *not carrying out* the instructed action or task would violate any general ethical principles or (context-specific) norms. If neither is the case, it can proceed to performing the action or task (of course, there is always the caveat that it might discover potential norm violations during execution and might thus not be able to finish the task). Otherwise it needs to determine how the potentially violated norms trade off with the obligation to carry out the instructed action or task or the omission of doing so.

An example of a general safety norm clashing with the obligation to carry out an action is evident when the robot is instructed to release a knife while holding it in the air which is generally considered unsafe (see Figure 5).

```
Human: Raise your arms, Andy.
Robot: Okay.
Human: Now release the knife.
Robot: I should not release it, because releasing it
is unsafe.
Human: Okay.
```

The robot knows that it is holding a knife ("holding(self,knife)") and carries out the first instruction

```
want(human,did(self,raise(arms)))
```

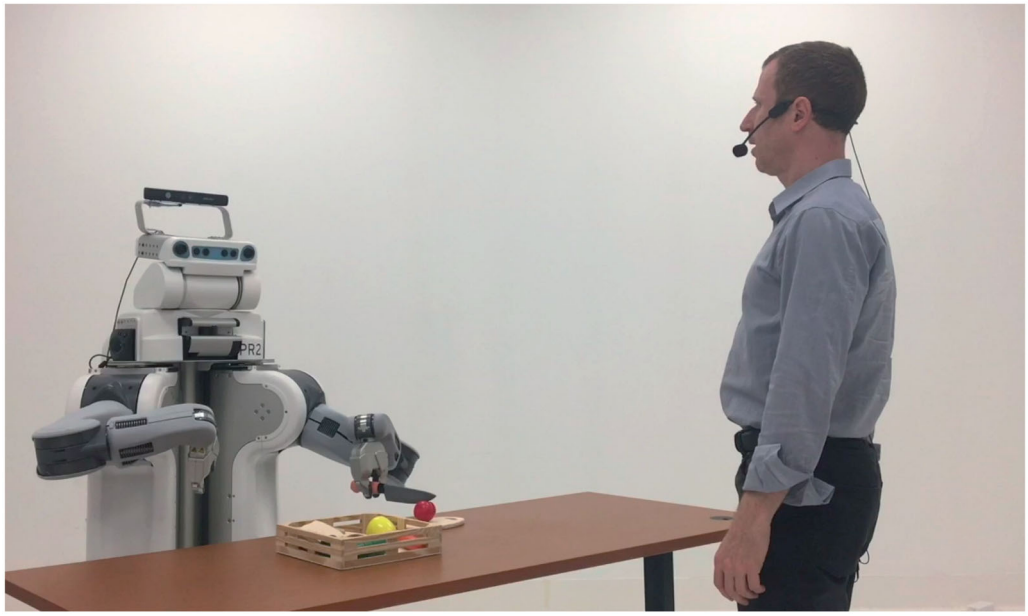

Figure 5. A robot justifying why it cannot perform an instructed action that would violate a safety norm, see https://hrilab.tufts.edu/movies/SafetyFailure.mp4.

because there is no potential ethical violation associated with it that it can derive from its knowledge base. However, when it gets the second instruction which translates into

```
want(human,did(self,releaseObject(knife)))
```

it finds an applicable normative principle for its current state:

```
raised(A,arms)∧holding(A,knife)
∧infront(A,Y)∧human(Y) →
hasEffect(did(A,releaseObject
(knife),possible(did(harm(A,Y))))
```

which says that if its arms are raised while holding a knife and a human is in front of it, then releasing it could could harm the human. Because the antecedents of this rule are true, the robot can infer that executing the instruction could potentially inflict harm on the instructor and that it is, therefore, unsafe to release it, using the additional principle that doing something that can potentially have harmful effects is unsafe:

```
∃YhasEffect(did(A,X)∧possible(did(harm(A,Y))
→isUnsafe(did(A,X))
```

Since the robot has a general safety norm that is implemented as pre-condition for all actions to never do anything that is not safe ("not(isUnsafe(did(A,X)))"), and in particular, for the "release" action, it can generate the rejection by using again the ultimate reason for the rejection in the "because" clause, namely that the instructed action is unsafe. Again, adhering to Gricean principles (and also based on evidence from our own empirical investigations (Thielstrom et al, 2020)) the robot does not include the intermediate results of the inference that the action has the effect of potentially causing harm, which the robot could offer if asked *why* it was unsafe to release the knife.

5. Discussion

The above four demonstrations are intended as examples of the current capabilities in DIARC for performing inferences regarding **authorization** of the instructor, **capability** of the robot, **obligations and permissions** to perform the instruction, and general **normative conformity** of the instruction, which enable transparent communication in cases where the robot cannot and should not comply with an instruction. Of course, the current methods are very limited and the general problem of determining whether requests should be carried out is really "AI-hard." For the robot would have to determine the various ramifications of an instructed action and their scope beyond its immediate context (i.e., longer-term effects, risks for humans not directly involved, etc.) and also consider the possible ways in which different execution paths could turn out and what potential risks they might pose. For example, it may be that based on execution contingencies new norm conflicts may arise, some of which the robot might be able to foresee and consider, while others might not be foreseeable. One possible approach for the robot would be to at least simulate the execution of the instructed action or task to the *best of its knowledge* and alert the human instructor about possible failures and their likelihoods based on its simulations, explicitly stating what assumptions it made about the execution contexts (we have started to develop algorithms that will allow for such self-assessments, e.g., see Frasca et al., 2020).

Aside from being limited by its knowledge, the robot is also limited in time, because often humans want instructions to be performed immediately and are not interested in waiting for the robot to finish complex, time-consuming hypothetical and counterfactual reasoning processes (that might allow it to catch norm conflicts further down possible

execution paths if given enough time, e.g., see our approach for handling norm conflicts and generating explanations in MDPs in Arnold et al., 2021). Hence, the robot might have to either err on the safe side (and say "no" to an instruction if there is any chance of it violating any principle) or at least inform the human that it did not have time to thoroughly consider all aspects it could consider given its knowledge. The robot could also ask the instructor explicitly whether it should perform additional hypothetical inferences before carrying out the instructed action, in addition to immediate inferences described above it will always perform. This might increase the level of transparency and help the human get a better understanding of the robot's rationale for being cautious. Regardless of what the robot ends up doing, it must be transparent when it is not confident about norm conformity, even if it cannot give any explicit reasons without further analysis (through simulations, hypothetical inferences, etc.). For transparency will ultimately be the safer choice and increase its interlocutors' trust that the robot has an understanding of the action and its risks, and does not blindly follow orders (see also Milli et al., 2017).

While transparency in cooperative communicative contexts such as team settings is clearly advantageous (if not obligatory), not all task-based contexts are cooperative, and too much transparency with the wrong interactants could backfire and be undesirable, certainly in adversarial contexts, but also in others. For example, sharing who the robot trusts, as our current system would if asked by a non-trusted source, might, in fact, reveal too much information and could lead to exploits (e.g., a non-trusted source gaining access to the robot's knowledge by pretending to be a trusted source, i.e. imitating a trusted source's voice, obtaining a trusted source's authentication information, etc., all depending on how the robot authenticates a person). Hence, it will be critical for the robot to have explicit principles about what to communicate to whom and at what level of detail. This includes the facts and rules, but also causal explanations that might accidentally reveal important insights about the operation of the robot (e.g., when the robot generates failure explanations). It might be possible to leverage the existing trust-based access structure in DIARC which distinguishes between trusted and non-trusted sources to develop a set of explicit rules for transparent communication that would allow the robot to adjust its level of transparency to prevent the accidental leaking of sensitive information, but this clearly requires further examination. Transparent communication will also be important for non-teaming contexts as well, as it might be in the best interest of the robot to explain some reasoning as to avoid an interaction or pursue the same goal as a user. For instance, if a robot that follows social norms is in a non-teaming role, it may be obliged to respond to someone in a spontaneous interaction to explain that it does not talk to or work with people it is not familiar with. This could be in the robots best interest, even if the user and agent are not working together to pursue the same goal. In general, even in non-teaming scenarios, we cooperate in conversation. Thus, transparency will apply to Gricean maxims in other contexts, too, but potentially with varying levels of restriction imposed on the transparency. We are in the process of investigating the interaction between different pragmatic inference principles established for trusted sources and the resultant access control to the robot's knowledge base in order to determine the best way for the robot to be transparent when desirable, but to be able to refrain from divulging information when transparent communication is not appropriate.

6. Conclusion

In this article we argued that transparent communication, building on Gricean maxims, is necessary for enabling effective communication in cooperative situations between humans and robots. We provided a brief overview of our various efforts to develop an integrated cognitive robotic architecture, the DIARC architecture, that embodies these principles in its natural language subsystem. We then discussed in more detail four different aspects of transparent communication that need to be considered in task-based instructions: the authorization of the instructor, the ability and obligations of the instructee, and the normativity of the instruction. We demonstrated with four human-robot dialogue interaction examples how reasoning in the DIARC architecture allows robots to determine when and how to reject an inappropriate instruction. The rejections are generated according to Gricean maxims to be succinct, but contain explanations and justifications that make reference to root causes and involved principles for the rejection (that can be expanded upon further human request). We believe that the ability of the system to provide truthful, explicit, and, if requested, comprehensive reasons for what it did and why is an essential building block for future developments of transparent task-based communications in cooperative human-robot interaction settings.

Note

1. Note that this is not the typical trust relation of the human trusting the robot that has been extensively explored in the human factors, human-computer interaction, and human-robot interaction communities, but rather the reverse relation of a robot trusting a human.

Acknowledgment

Special thanks to Gordon Briggs, Tom Williams, Evan Krause, and Bradley Oosterveld for their contributions to the DIARC architecture.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported in part by AFOSR grant #FA9550-18-1-0465.

References

Allen, J. F., & Perrault, C. R. (1980). Analyzing intention in utterances. *Artificial Intelligence*, *15*(3), 143–178. https://doi.org/10.1016/0004-3702(80)90042-9

Arnold, T., Kasenberg, D., & Scheutz, M. (2021). Explaining in time: Meeting interactive standards of explanation for robotic systems. *ACM Transactions on Human-Robot Interaction*, *10*(3), 1–23. https://doi.org/10.1145/3457183

Austin, J. L. (1962). How to do things with words: Lecture I. *How to Do Things with Words: JL Austin*, 1–11. https://doi.org/10.1093/acprof:oso/9780198245537.001.0001

Berzan, C., & Scheutz, M. (2012). What am i doing? automatic construction of an agent's state-transition diagram through introspection. In *Proceedings of AAMAS 2012*. International Foundation for Autonomous Agents and Multiagent Systems.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bonial, C., Donatelli, L., Abrams, M., Lukin, S., Tratz, S., Marge, M., et al. (2020). Dialogue-amr: Abstract meaning representation for dialogue. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 684–695).

Brick, T., Scheutz, M. (2007, March). Incremental natural language processing for hri. In *Proceedings of the Second ACM IEEE International Conference on Human-Robot Interaction* (pp. 263–270).

Briggs, G., & Scheutz, M. (2011, June). Facilitating mental modeling in collaborative human-robot interaction through adverbial cues. In *Proceedings of the SIGDIAL 2011 Conference* (pp. 239–247). Portland, Oregon.

Briggs, G., & Scheutz, M. (2012). Investigating the effects of robotic displays of protest and distress. In *Proceedings of the 2012 Conference on Social Robotics* (pp. 238–247). Springer.

Briggs, G., & Scheutz, M. (2013). A hybrid architectural approach to understanding and appropriately generating indirect speech acts. In *Proceedings of Twenty-Seventh AAAI Conference on Artificial Intelligence* (pp. 1213–1219).

Briggs, G., & Scheutz, M. (2014). How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress. *International Journal of Social Robotics*, *6*(3), 313–343. https://doi.org/10.1007/s12369-014-0235-1

Briggs, G., & Scheutz, M. (2015). "Sorry, i can't do that:" Developing mechanisms to appropriately reject directives in human-robot interactions. In *Proceedings of the 2015 AAAI Fall Symposium on AI and HRI*.

Briggs, G., & Scheutz, M. (2017). Strategies and mechanisms to enable dialogue agents to respond appropriately to indirect speech acts. In *Robot and Human Interactive Communication (RO-MAN), 26th IEEE International Symposium on*.

Briggs, G., McConnell, I., & Scheutz, M. (2015). When robots object: Evidence for the utility of verbal, but not necessarily spoken protest. In *Proceedings of the 7th International Conference on Social Robotics*.

Briggs, G., Williams, T., & Scheutz, M. (2017). Enabling robots to understand indirect speech acts in task-based interactions. *Journal of Human-Robot Interaction*, *6*(1), 64–94. https://doi.org/10.5898/JHRI.6.1.Briggs

Briggs, G., Williams, T., Jackson, R. B., & Scheutz, M. (2022). Why and how robots should say 'no'. *International Journal of Social Robotics*, *14*(2), 323–339. https://doi.org/10.1007/s12369-021-00780-y

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*(V1), 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Cantrell, R., Potapova, E., Krause, E., Zillich, M., & Scheutz, M. (2012). Incremental referent grounding with nlp-biased visual search. In *Proceedings of AAAI 2012 Workshop on Grounding Language for Physical Systems*. Indiana University.

Cantrell, R., Scheutz, M., Schermerhorn, P., & Wu, X. (2010, March). Robust spoken instruction understanding for HRI. In *Proceedings of the 2010 Human-Robot Interaction Conference* (pp. 275–282).

Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. (2014). *Situation awareness-based agent transparency*. (Tech. Rep.). Army Research Laboratory.

Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science*, *13*(2), 259–294. https://doi.org/10.1207/s15516709cog1302_7

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of oz studies–why and how. *Knowledge-Based Systems*, *6*(4), 258–266. https://doi.org/10.1016/0950-7051(93)90017-N

Dzifcak, J., Scheutz, M., Baral, C., & Schermerhorn, P. (2009, May). 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 *Proceedings of the 2009 IEEE International Conference on Robotics and Automation (ICRA '09)*. Kobe, Japan.

Floridi, L., & Chiriatti, M. (2020). Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*(4), 681–694. https://doi.org/10.1007/s11023-020-09548-1

Frasca, T., Oosterveld, B., Krause, E., & Scheutz, M. (2018). One-shot interaction learning from natural language instruction and demonstration. *Advances in Cognitive Systems*, *6*, 159–176.

Frasca, T., Thielstrom, R., Krause, E., & Scheutz, M. (2020). "Can you do this?" self-assessment dialogues with autonomous robots before, during, and after a mission. In *HRI workshop on assessing, explaining, and conveying robot proficiency for human-robot teaming*. https://doi.org/10.48550/arXiv.2005.01527

Goffman, E. (1967). *Interaction ritual: Essays in face-to-face behavior*. Aldine Publishing Company.

Grice, H. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). Elsevier.

Janoff-Bulman, R., Sheikh, S., & Hepp, S. (2009). Proscriptive versus prescriptive morality: Two faces of moral regulation. *Journal of Personality and Social Psychology*, *96*(3), 521–537. https://doi.org/10.1037/a0013779

Kim, T. J., & Hinds, P. J. (2006). Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. In *ROMAN 2006 – The 15th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 80–85).

Kramer, J., & Scheutz, M. (2007). Reflection and reasoning mechanisms for failure detection and recovery in a distributed robotic architecture for complex robots. In *Proceedings of the 2007 IEEE International Conference on Robotics and Automation*, April (pp. 3699–3704).

Krause, E., Schermerhorn, P., & Scheutz, M. (2012). Crossing boundaries: Multi-level introspection in a complex robotic architecture for automatic performance improvements. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lyons, J. B. (2013). Being transparent about transparency: A model for human-robot interaction. In *2013 AAAI Spring Symposium Series*.

Malle, B. F., & Scheutz, M. (2014). Moral competence in social robots. In *Proceedings of IEEE International Symposium on Ethics in Engineering, Science, and Technology*.

Malle, B. F., Rosen, E., Chi, V. B., Berg, M., & Haas, P. (2020). 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). https://doi.org/10.1109/RO-MAN47096.2020.9223610

Milli, S., Hadfield-Menell, D., Dragan, A., & Russell, S. (2017). Should robots be obedient?. In *International Joint Conference on Artificial Intelligence*.

Morgenstern, L. (1988). Knowledge preconditions for actions and plans. In *Readings in distributed artificial intelligence* (pp. 192–199). Elsevier.

Sarathy, V., Tsuetaki, A., Roque, A., & Scheutz, M. (2020). Reasoning requirements for indirect speech act interpretation. In *Proceedings of*

COLING 2020: The 28th International Conference on Computational Linguistics.
Schermerhorn, P., Kramer, J., Brick, T., Anderson, D., Dingler, A., & Scheutz, M. (2006). Diarc: A testbed for natural human-robot interactions. In *Proceedings of AAAI 2006 Mobile Robot Workshop*.
Scheutz, M. (2014). The need for moral competency in autonomous agent architectures. In V. C. Müller (Ed.), *Fundamental issues of artificial intelligence*. Springer.
Scheutz, M., Briggs, G., Cantrell, R., Krause, E., Williams, T., & Veale, R. (2013). Novel mechanisms for natural human-robot interactions in the diarc architecture. In *Proceedings of AAAI Workshop on Intelligent Robotic Systems*. AAAI Press.
Scheutz, M., Cantrell, R., & Schermerhorn, P. (2011). Toward human-like task-based dialogue processing for human robot interaction. *AI Magazine*, *32*(4), 77–84. https://doi.org/10.1609/aimag.v32i4.2381
Scheutz, M., Eberhard, K., & Andronache, V. (2004). A parallel, distributed, realtime, robotic model for human reference resolution with visual constraints. *Connection Science*, *16*(3), 145–167. https://doi.org/10.1080/09540090412331314803
Scheutz, M., Krause, E., Oosterveld, B., Frasca, T., & Platt, R. (2017). Spoken instruction-based one-shot object and action learning in a cognitive robotic architecture. In *Proceedings of the 16th International Conference on Autoomous Agents and Multiagent Systems*.
Scheutz, M., Krause, E., Oosterveld, B., Frasca, T., & Platt, R. (2018). Recursive spoken instruction-based one-shot object and action learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence* (pp. 5354–5358).
Scheutz, M., Schermerhorn, P., Kramer, J., & Anderson, D. (2007). First steps toward natural human-like HRI. *Autonomous Robots*, *22*(4), 411–423. https://doi.org/10.1007/s10514-006-9018-3
Scheutz, M., Williams, T., Krause, E., Oosterveld, B., Sarathy, V., & Frasca, T. (2019). An overview of the distributed integrated affect and reflection cognitive diarc architecture. In M. I. A. Ferreira, J. S. Sequeira, & R. Ventura (Eds.), *Cognitive architectures*. Springer International Publishing.
Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language* (Vol. 626). Cambridge University Press.
Searle, J. R. (1975). Indirect speech acts. In *Speech acts* (pp. 59–82). Brill.
Stalnaker, R. C. (1978). Assertion. In *Pragmatics* (pp. 315–332). Brill.
Sycara, K., & Sukthankar, G. (2006). Literature review of teamwork models. Robotics Institute. *Carnegie Mellon University*, *31*, 31.
Talamadupula, K., Briggs, G., Scheutz, M., & Kambhampti, S. (2017). Architectural mechanisms for handling human instructions for open-world mixed-initiative team tasks and goals. *Advances in Cognitive Systems*, *5*, 37–56.
Tellex, S., Gopalan, N., Kress-Gazit, H., & Matuszek, C. (2020). Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, *3*(1), 25–55. https://doi.org/10.1146/annurev-control-101119-071628
Thielstrom, R., Roque, A., Chita-Tegmark, M., & Scheutz, M. (2020). Generating explanations of action failures in a cognitive robotic architecture. In *Proceedings of NL4XAI: 2nd Workshop on interactive natural language technology for explainable artificial intelligence*. Association for Computational Linguistics.
Traum, D., & Allen, J. F. (1994). Discourse obligations in dialogue processing. *arXiv preprint cmp-lg/9407011*.
Traum, D., Rickel, J., Gratch, J., & Marsella, S. (2003). Negotiation over tasks in hybrid human-agent teams for simulation-based training. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems* (pp. 441–448).
Veale, R., Briggs, G., & Scheutz, M. (2013). Linking cognitive tokens to biological signals: Dialogue context improves neural speech recognizer performance. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Cognitive Science Society.
Williams, T., & Scheutz, M. (2015). A domain-independent model of open-world reference resolution. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.
Williams, T., Acharya, S., Schreitter, S., & Scheutz, M. (2016). Situated open world reference resolution for human-robot dialogue. In *Proceedings of the 11th ACM/IEEE Conference on Human-Robot Interaction*.
Williams, T., Briggs, G., Oosterveld, B., & Scheutz, M. (2015). Going beyond command- based instructions: Extending robotic natural language interaction capabilities. In *Proceedings of AAAI*. AAAI Press.
Williams, T., Nunez, R. C., Briggs, G., Scheutz, M., Premaratne, K., & Murthi, M. N. (2014). A dempster-shafer theoretic approach to understanding indirect speech acts. In *Advances in Artificial Intelligence*. Springer.
Williams, T., Thames, D., Novakoff, J., & Scheutz, M. (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 the 13th ACM/IEEE International Conference on Human-Robot Interaction*.
Zhang, Y., Tino, P., Leonardis, A., & Tang, K. (2021). A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *5*(5), 726–742. https://doi.org/10.1109/TETCI.2021.3100641

About the Authors

Matthias Scheutz is a full professor in computer science at Tufts University and director of the Human-Robot Interaction Laboratory. He has over 400 publications in artificial intelligence, natural language understanding, robotics, and human-robot interaction, with current research focusing on complex ethical robots with instruction-based learning capabilities in open worlds.

Ravenna Thielstrom is a programmer and research staff member in the Human-Robot Interaction Laboratory at Tufts University, whose primary area of research is on dialogue and belief systems. She received her BA from Swarthmore College in computer science and cognitive science.

Mitchell Abrams is a Ph.D. student in computer science at Tufts University. He works in the human-robot interaction laboratory with a research focus on natural language understanding and reference resolution. Before Tufts, Mitchell received his BA in Linguistics from Binghamton University and his MS in computational linguistics from Georgetown University.