Merging Representation and Management of Physical and Spoken Action

Workshop on Cognitive Architectures for Human-Robot Interaction

Charles Threlkeld
Tufts University, Human-Robot Interaction Lab
Charles.Threlkeld@Tufts.edu

Matthias Scheutz
Tufts University, Human-Robot Interaction Lab
Matthias.Scheutz@Tufts.edu

ABSTRACT

We propose unifiying representation and management of dialogue actions and physical actions. This proposal improves planning and reasoning about goals and goal failure. We use an introspectable domain-specific language that allows greater online access to robotic state via natural language. Finally, we can teach speech acts and contextual salience of any act to the robot with this new system.

KEYWORDS

dialogue; conversation; multi-modal; communication; goal-orientation

1 INTRODUCTION

Human communication takes many forms. Language is a primary vector [5], but gesture is also important [10]. Non-lexical language is also meaningful. Speech timing can convey information [2], and disfluencies can inform our shared mental models [7]. Each of these vectors has its own type of information to convey, but each is still a type of communication. In this proposal, we merge the representation, planning and execution of spoken and physical action. Then, we discuss some advantages of this high-level abstraction.

Framing speech as action has a long tradition. Philosophers of language have argued that language is action for decades [1, 20]. Roboticists have also made the link by coupling perception and action [13].

In computer science, dialogue management systems have made significant progress toward robust computer-human dialogues [4]. However, dialogue systems are limited in their scope, since their only considerations are verbal.

On the other end of the spectrum, robotic architectures often disregard dialogue completely, as language is not their focus. Many robotic dialogue managers are not particularly complex. Some are simple tree structures that assume atomic execution at the leaves. Some robotic architectures have more intricate action managers, allowing for parallel, durative actions with complex subactions. These capabilities are rarely extended to dialogue management.

Cognitive architectures are caught in the middle. On the one hand, dialogue is an important part of general cognition. On the other hand, they don't have the flexibility of most dialogue systems, since they make commitments to embodied action as well. In our own DIARC architecture, the previous dialogue manager was chiefly a user interface for the performace of other tasks. (See Section 3 for information on DIARC.) This lesser status creates a tension between the dialogue system versus planning and reasoning systems at large. Distinct representations made information opaque and reasoning obtuse.

In this proposal, we describe our reasoning and early results merging dialogue and action management. Both systems are now handled by a single Goal Manager component (see Figure 4), which operates on unified behavioral scripts. This merger eases the tension between robotic architecture and dialogue manager detailed above.

In this paper, we provide several motivating examples, showing aspects of dialogue that we aspire to handle. Next, we outline the prior capabilities of the DIARC architecture that we leverage in our behavior manager. In Section 4, we outline several main benefits we are utilizing in our new dialogue management system, and finally we show several promising areas for future research.

2 MOTIVATING EXAMPLES

To inform our decisions about dialogue management, we kept in mind several aspirational examples. While these examples are still goalposts rather than accomplishments, they serve to highlight common features of dialogue on which we hope to improve. Existing systems, such as neural network-based chatbots or slot-filling dialogue systems fall shorts in these examples [4]. So, these examples serve as a starting point in avoiding common obstacles, as well as providing exploratory direction.

To make things concrete, consider the dialogue in Figure 1. Here, a human asks a robot how it would respond to an utterance, and then tells it to modify its response. In line 1, the human asks the robot to consider a counterfactual, and in line 2 she describes the hypothetical situation. To answer this question, the robot must be able to entertain a hypothetical, and then introspect about the various steps taken in such a situation. Entertaining a hypothetical

```
1  H: How would you respond if I were to
2     say "can you walk forward?"
3  R: First, I consult my knowledge base
4     To see if I am able to walk forward
5     If I am,
6     I say yes
7     and I begin walking forward
8     If I am not able,
9     I say no
10 H: When I ask if you can walk forward,
11    Do not begin walking
12 R: Ok
13 H: Can you walk forward?
14 R: Yes
```

Figure 1: A sample dialogue showing an interpretable introspective format.

```
1 R: John Doe 90
2 S: (enters) Hello Professor
3 P: Hello Sam
4 R: Sam Seaborn 95
5 P: Robot, when a student is present
     do not say scores out loud
     (Robot continues working silently)
  ...
6 S: Thanks, Professor. Bye!
7 R: Jane Doe 85
```

Figure 2: A sample dialogue showing context effects of dialogue actions. The robot R is grading assignments and reporting results to professor P when student S enters the room.

```
1 H: I'd like to ask you about your abilities
2 R: Okay
3 H: Can you walk forward?
4 R: Yes
     (The robot remains motionless)
    ...
5 H: Now, I'd like to see you demonstrate your abilities.
6 H: Can you walk forward?
7 R: Okay (The robot walks forward)
```

Figure 3: A sample dialogue exhibiting the effects of context-appropriate physical action. The robot should move in the demonstration context, but should stay still simply asked about its capabilities.

is beyond the scope of this paper, but to have a robot introspect about its own actions is integral to our dialogue solution.

By introspecting, the robot is able to read its own code and translate it into a natural language description. So the syntax of a behavior script can be parsed from its code into a natural language utterance by the robot. This means the robot can not only perform dialogue actions, but it can also explain what dialogue action it performs and why it performs them.

At line 3, the robot explains that it has online knowledge, and at line 4 it expresses how it thinks about meta-knowledge. At lines 5 and 8, it explains its use of conditionals. The robot is able to explain, in natural language, its reasoning process, its abilities, and its action intentions.

In lines 10 and 11, the human asks the robot to change its response to her utterance. Since the robot is able to explain its own actions, it can also reason about how to change them. In this case, the robot removes its physical action from this behavior script. After modification, the script invoked includes only the utterance describing capability, and the physical motion is absent. This kind of action modification is one benefit of our goal manager.

Contextual sensitivity is another goal of our dialogue system, as exemplified by the dialogue in Figure 2. In this case, the situation of the robot changes, making different actions appropriate or inappropriate. It is a teaching assistant robot (R), helping the Professor (P) to grade papers. During the grading, a student (S) enters the room.

Like Figure 1, the robot must change its behavior online and through natural language. However, the difference here is not that the action is modified online, but by conditional context. Actions must meet pre-conditions to occur. Further, post-conditions state what is true when the action completes. Here, the professor adds a precondition to the utterance that no students are present. So, the robot can continue its goal of grading papers, but will not utter grades with a student in earshot. The robot can then choose the appropriate action given its situation.

Figure 3 provides a final example dialogue. A human first wants to ask a robot about its capabilities and then wants the robot to demonstrate them. As a parallel, consider a human context of a job interview, where capabilities are spoken, and a test where they are displayed. Here, the human gives the proper contexts at lines 1 and 5. The robot can leverage these contexts to treat the ambiguous queries of the human in lines 3 and 6. These lines are identical on the surface. However, with context the first is a question and the second is a command. Out proposal will account for contextual change throughout a dialogue so that action is only taken when called for.

In Figure 1, the robot decouples a physical action (walking) from a speech action (inquiring). In Figure 2, the robot decouples a speech action (reporting) from a physical action (grading). In Figure 3, the robot decouples either action from the wording of the dialogue. In each case, the same system and representation must handle both speech acts and physical acts.

Merging speech and physical representation is the first step toward realizing these examples in a robot. In Figure 1, removing the physical act from the script requires changing pragmatics of indirect speech acts. Figure 2 requires a second, new grading action with contextual selection. And Figure 3 requires dialogue-imposed context on both speech and physical actions. Each example also requires extension beyond dialogue (e.g., hypothetical parsing in the natural language understanding pipeline in Figure 1, or dynamic context parsing in Figure 3), but these aspects are beyond the scope of this paper.

3 INFRASTRUCTURE: DIARC

We are in the early stages of implementing this proposal. We are leveraging the Distributed Integrated Affect Reflection and Cognition (DIARC) cognitive architecture using the Architecture Development Environment (ADE) middleware. DIARC is an architecture scheme that guides components and their links [19]. ADE is a development environment based on a universal agent architecture framework. It acts as the middleware between the cognitive algorithms, such as those described here, and the low-level effectors of a given robot [17].

Previous work with this architecture has produced a suite of components, as shown in Figure 4. The perception layer processes audiovisual input. In the reasoning layer, the belief system provides knowledge storage and queries. The natural language understanding pipeline goes from text from the automatic speech recognition to a rich semantic representation in the dialogue manager. Natural language generation takes high-level semantic representation and

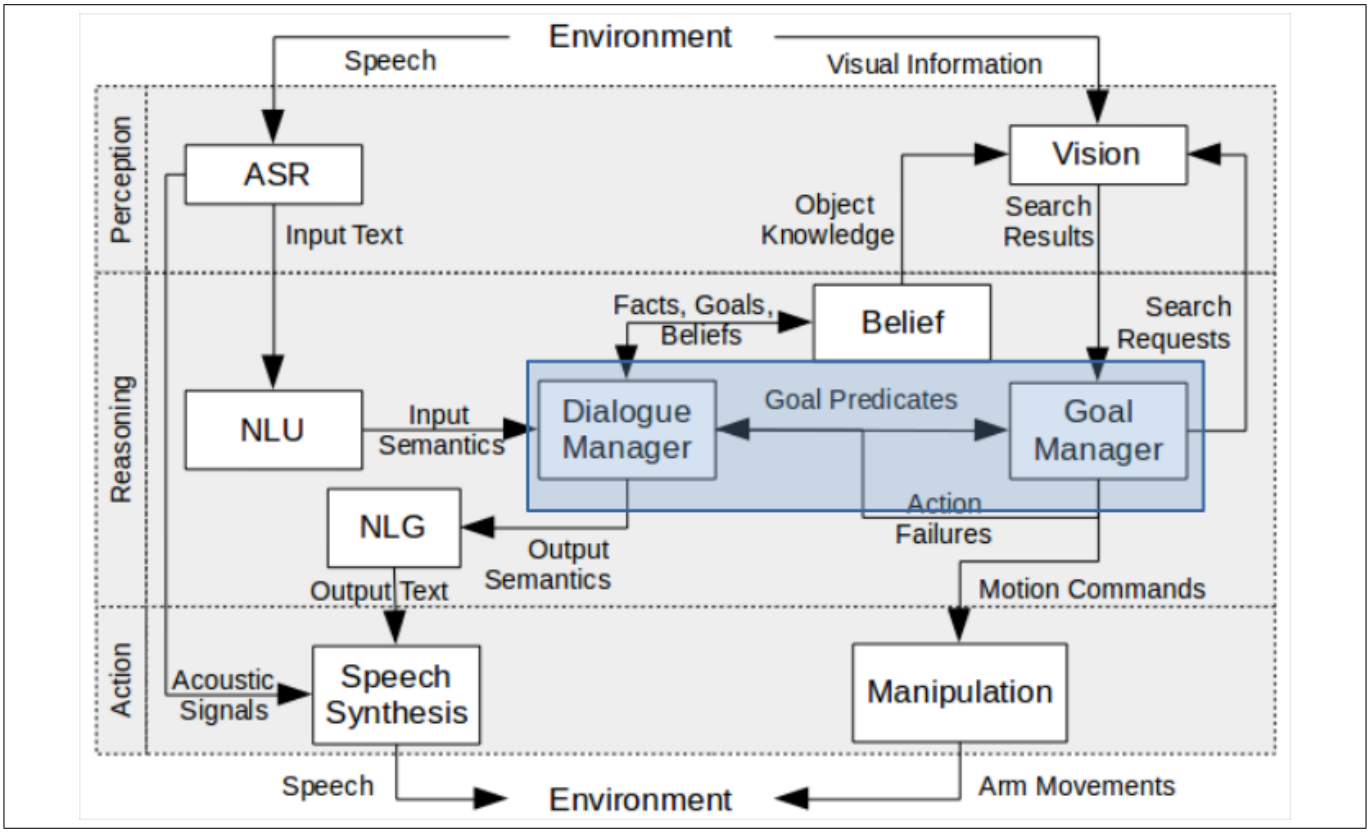

Figure 4: The DIARC architecture. Each box represents a component within the architecture. E.g., the Vision component is responsible for searching for visual information in the environment, and passing results of these searches to the Belief and Goal Manager components. The proposal in this paper fuses the Dialogue Manager and the Goal Manager (highlighted in blue) into a single Behavior Manager. So, the reasoning layer is simplified.

produces speech. Finally, interfaces with robotic actuators in the action layer allow for physical manipulation.

The management of dialogue uses detailed representations of the original acoustic signal, including symbolic representation of semantics, speaker, time, and other utterance traits. Note that the architecture does not require commitments to any specific component implementations.

DIARC and ADE have several benefits that make it an excellent base for dialogue actions. First, the extant goal manager implements an introspectable domain-specific language. Next, action representation is symbolic and grounded. Grounded, symbolic, introspectable representation allows for online learning and modification. Finally, the ADE middleware can provide a shared knowledge base to multiple robots at once. This means that teaching one robot propogates to all robots involved, even those in other contexts. With proper pre- and post-conditions, robots can select actions only when their own abilities permit [14] or when their context permits [8]. Further, robots in any context can use that knowledge immediately.

4 TOWARD HANDLING ADAPTABLE DIALOGUE ACTIONS

We are in the early stages of making both physical and speech action representation introspectable, modifiable, and explainable online. Our dialogue manager is being merged with the Goal Manager component. We are building on the capabilities already implemented there with the unique dialogue functionality.

Section 2 showed motivating examples and Section 3 explored the foundations available in the existing architecture. In this section, we detail our contribution, how it builds on the existing infrastructure, and how it moves toward solutions to the example dialogues.

4.1 Reasoning about Actions and Goals

The most important aspect of the architecture is that physical action and spoken action have the same representation and management within the system. This singular pipeline brings several advantages. Response, planning and execution can share information freely. This allows data to move easily between physical, observed, spoken, and heard action.

Goal management is necessary for reasoning action in a given context. For example, if a person says "Please walk forward", the robot must be able to translate this utterance into a goal. In our

case, we have chosen to represent goals as predicates such as $want(brad, robot, walk(forward))$. These goals can happen at almost any level of communication or within a task. As long as a component can pass on a goal predicate, the goal manager can take over. This allows representation of dialogue actions such as $greet(self, brad, hello)$ and physical actions such as $walk(forward)$ in the same format. Consistent data structures provides transparent reasoning about instruction, performance, failure, and reasoning. Human readability of these predicates simplifies translation to and from natural language.

Beliefs about the world inform goal reasoning. Likewise, querying knowledge is essential for leveraging pre-conditions and post-conditions. To choose appropriate actions when given goals, the robot can query that conditions in the world hold. For example, a pre-condition to walking could be $is(area(front), safe)$. Then, the robot will only walk forward when it deems it safe to do so.

4.2 Introspection

In our action system, we use a domain-specific language (DSL) rooted in XML.The DSL is introspectable by the greater system. So, we can ask the robot questions about its knowledge, beliefs, and behavioral scripts. For example, we can ask the robot "How do you do a squat?" and it will respond with a description of its process. (E.g., "Bend my knees down. Straighten my knees.")

The process for describing physical actions and speech actions is identical. In line with merging speech action representation with physical action representation, speech action is also introspectively describable. Figure 1 shows an example of introspecting about the walking behavior.

4.3 Failure Catching

Failure catching is a prime feature of our system. The action management system is introspectable, which allows the robot to determine when and why an action fails. Another major mechanism for failure catching is the linking together of multi-modal actions. Physical action and speech action share the same pipeline. So, all action successes or failures are accessible to the language pipeline. Further, gesture can share the same pipeline as speech production, much like humans [6].

The belief system allows us to examine failures at arbitrary granularity. For example, we plan to use our natural language pipeline as a series of libraries, so that we can assess each possible point of failure, and respond appropriately. For example, if the speech recognizer creates text, but the parser fails to create a semantic representation, the robot can manage that type of failure differently than a failure of (e.g.) reference resolution.

If an action fails at the belief level, the robot can explain at the belief level. For example, if there is a prohibition on moving when instructed by an untrusted agent. The robot can reply: "I will not walk forward because I do not trust you". But, if the robot does not know certain words, it can fail at the lexical level ("I do not know what forward means").

If the pre-condition of a goal is not met, the robot can choose a different course of action. For example, it could fail and explain why. (E.g. "I cannot move forward because it is not safe to do so"). The belief component allows us to reason about predicate knowledge both known and inferrable from other information. This allows our pre-conditions to be robust, and our post-conditions to be meaningful.

4.4 Teaching Speech Actions

The language chosen for our action system is accessible to both human and robot by design. Exposing the workings of this subsystem lets the robot leverage its abstract knowledge to do online, one-shot learning. The system can currently learn physical actions online by stringing together sub-actions in series. (E.g. to nod, move your head up, down, up, down). [18]

Unified underpinnings allow us to teach the robot speech actions in the same way. In Figure 1, line 13 above, the robot is dealing with an indirect speech act. The literal meaning of "Can you walk forward?" is inquiring about ability, but the pragmatic, indirect meaning is an instruction to do so [3]. We plan to leverage this process to teach contextual pragmatics online, as in Figure 3.

4.5 Learning Context Salience of Actions

Further, the robot can learn context salience of actions online by asserting context as pre-conditions. Context-specific norms, for example, can be learned through natural language while the robot is running [15]. Additionally, post-conditions can change online. So, the effects a robot believes an action to have can adapt to novel situations. In a dialogue system, this means that the robot can learn that certain dialogue moves may be only appropriate in certain contexts. The example dialogues in Figures 2 and 3 exemplify this online context learning.

5 FUTURE WORK

Our action knowledge representation gives many benefits as outlined above. Below are several research fronts that we plan to explore given this starting point.

5.1 Conversational Repair

Conversational repair is a prime problem to approach with a speech act solution. Conversation analysts have noticed that utterances often come in pairs, such as question/answer or request/rejection [16]. In fact, this is the foundation of some dialogue systems, especially with queries and responses (e.g. Alexa or Siri).

Yet, when something goes wrong, people insert sequences to solve their conversational problems [21]. For example, if a word does not fit its role in a sentence, it is corrected. (E.g. "Did you say dog or bog?"). Our system can check for understanding throughout the natural language understanding pipeline. So, we can repair from the lexical up to the pragmatic levels of language.

Additionally, the pliability of the system allows learning pragmatic contexts online. Pragmatics is an expansive topic (see [11]), but we are considering the non-literal information of an utterance. Pragmatic content can depend on the utterance form, as in indirect speech acts [23], the social context as in Gricean conventions [9], or even turn-shape as in conversational preferences [2]. In our system, a robot can learn pragmatic context online. So, contextual considerations can uncover informational intent that may not be obvious given linguistic form.

5.2 Including Deictic Gesture

Teaching the robot new things about dialogue is interesting with only speech, but the merging of physical and speech act is something that humans do with little thought. Indeed, pointing generally precedes speaking in humans, even with ascribed intentionality [22]. Many gestures can bolster, clarify, or even stand in for paired words [12].

Deictic gesture is often important in reference resolution. An embodied agent may want to identify a specific object among several distractors. Consider "that ball" with a pointing motion versus "the ball second from the right" without gesture. We can see that pointing is an excellent extension and natural pairing to a speech act.

5.3 Robotic Self-Testing

Modifying our beliefs online is a messy business. Most humans can get by with some amount of cognitive dissonance. Mission-critical situations do not allow hesitation. Our platform allows for the robot to test itself and ensure that modifying one action will not alter the outcome of others. For example, a robot may verbally alert its coworker when it has completed a task (say, picking up an object), but the coworker finds this auditory alert overkill, he can teach the robot to keep quiet. But, keeping quiet prevents communication when there is no clear field of vision between the robot and human.

In a self-test, the robot could alert the human that turning off the self-reporting mechanism will no longer guarantee that the partner will know where the robot stands in its production. Problems can cascade if not caught early, and self-testing allows the robot to find problems even before they occur, allowing the human to take responsibility for any risk imposed.

6 CONCLUSION

In this paper, we have discussed the virtues of unifying representation of speech and physical action. We have detailed how these changes improve reasoning about goals and actions. The proposal improves our ability catch and handle robotic failure. The move to our domain-specific language allows us to introspect about our actions. Online learning of speech sequences is available. Finally, we can teach contextual salience to the online robot. We also detailed several areas of further research that we are making steps toward achieving in future research.

7 ACKNOWLEDGEMENTS

This work was supported in part by ONR MURI grant N00014-16-1-2278 from the US Office of Naval Research. Feedback from the anonymous reviewers and Antonio Roque was also integral to a much improved final project.

REFERENCES

[1] John Langshaw Austin. 1975. *How to do things with words*. Vol. 88. Oxford university press, 198 Madison Avenue, New York, NY 10016, USA.
[2] Sara Bogels, Kobin H. Kendrick, and Stephen C. Levinson. 2015. Never Say No ... How the Brain Interprets the Pregnant Pause in Conversation. *PLOS ONE* 10, 12 (2015), e0145474. https://doi.org/10.1371/journal.pone.0145474
[3] Gordon Briggs, Tom Williams, and Matthias Scheutz. 2017. Enabling robots to understand indirect speech acts in task-based interactions. *Journal of Human-Robot Interaction* 6, 1 (2017), 64–94.
[4] Philip R Cohen. 2019. Back to the Future for Dialogue Research: A Position Paper. (27 Jan 2019).
[5] Jan Peter de Ruiter. 2004. On the primacy of language in multimodal communication. In *LREC 2004 Workshop on Multimodal Corpora*. ELRA-European Language Resources Association, 9, Rue des Corelières, 75013 Paris, 38–41.
[6] Jan Peter de Ruiter. 2017. The asymmetric redundancy of gesture and speech. *Why Gesture?: How the hands function in speaking, thinking and communicating* 7 (2017), 59.
[7] Felix Gervits, Kathleen M. Eberhard, and Matthias Scheutz. 2016. Disfluent but effective? A quantitative study of disfluencies and conversational moves in team discourse. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*. ACL, 8 Jurong Town Hall, Road #20-05/06, Singapore 609434, 3359–3369. http://aclweb.org/anthology/C/C16/C16-1317.pdf
[8] Felix Gervits, Terry W Fong, and Matthias Scheutz. 2018. Shared Mental Models to Support Distributed Human-Robot Teaming in Space. In *2018 AIAA SPACE and Astronautics Forum and Exposition*. American Institute of Aeronautics and Astronautics, 12700 Sunrise Valley Drive, Suite 200, Reston, VA 20191-5807, 5340.
[9] H.P. Grice. 1975. Logic and Convesation. *Syntax and Semantics* 3 (1975), 41–58.
[10] Minae Kwon, Sandy H Huang, and Anca D Dragan. 2018. Expressing robot incapability. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2 Penn Plaza, Suite 701, New York, NY 10121-0701, 87–95.
[11] Stephen C. Levinson. 1983. *Pragmatics. Cambridge textbooks in linguistics*. Cambridge University Press, Shaftesbury Road Cambridge CB2 8BS UK.
[12] David McNeill. 1985. So you think gestures are nonverbal? *Psychological review* 92, 3 (1985), 350.
[13] Monica N. Nicolescu and Maja J. Mataric. 2003. Linking Perception and Action in a Control Architecture for Human-Robot Domains. In *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the*. IEEE, 126. https://doi.org/10.1109/HICSS.2003.1174287
[14] Vasanth Sarathy. 2016. Inferring higher-order affordances for more natural human-robot collaboration. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 637–638.
[15] Vasanth Sarathy, Matthias Scheutz, Yoed N. Kenett, Mowafak Allaham, Joseph L. Austerweil, and Bertram F. Malle. 2017. Mental Representations and Computational Modeling of Context-Specific Human Norm Systems. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society, CogSci 2017, London, UK, 16-29 July 2017*. cognitivesciencesociety.org, 6. https://mindmodeling.org/cogsci2017/papers/0202/index.html
[16] Emanuel A Schegloff and Harvey Sacks. 1973. Opening up closings. *Semiotica* 8, 4 (1973), 289–327.
[17] Matthias Scheutz. 2006. ADE: Steps toward a distributed development and runtime environment for complex robotic agent architectures. *Applied Artificial Intelligence* 20, 2-4 (2006), 275–304.
[18] Matthias Scheutz, Evan Krause, Brad Oosterveld, Tyler Frasca, and Robert Platt. 2017. Spoken Instruction-Based One-Shot Object and Action Learning in a Cognitive Robotic Architecture. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS '17)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1378–1386. http://dl.acm.org/citation.cfm?id=3091282.3091315
[19] Matthias Scheutz, Thomas Williams, Evan Krause, Bradley Oosterveld, Vasanth Sarathy, and Tyler Frasca. 2019. An overview of the distributed integrated cognition affect and reflection DIARC architecture. In *Cognitive Architectures*. Springer, 165–193.
[20] John Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.
[21] Jack Sidnell. 2011. *Conversation Analysis: An Introduction*. Wiley-Blackwell.
[22] Michael Tomasello, Malinda Carpenter, and Ulf Liszkowski. 2007. A new look at infant pointing. *Child development* 78, 3 (2007), 705–722.
[23] Tom Williams, Gordon Briggs, Bradley Oosterveld, and Matthias Scheutz. 2015. Going Beyond Literal Command-based Instructions: Extending Robotic Natural Language Interaction Capabilities. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15)*. AAAI Press, 1387–1393. http://dl.acm.org/citation.cfm?id=2887007.2887199