# Robots that Help Humans Build Better Mental Models of Robots

**Preeti Ramaraj**
Computer Science and Engineering
University of Michigan, Ann Arbor
2260 Hayward Street, Ann Arbor, MI - 48105
preetir@umich.edu

## Abstract

Interactive Task Learning (ITL) is an approach to teaching robots new tasks through language and demonstration. It relies on the fact that people have experience teaching each other. However, this can be challenging if the human instructor does not have an accurate mental model of a robot. This mental model consists of the robot's knowledge, capabilities, shortcomings, goals, and intentions. The research question that I investigate is "How can the robot help the human build a better mental model of the robot?"

## Introduction

I envision a future where robots help people by performing diverse tasks including household, rehabilitative, and office tasks. To perform these tasks effectively, people must communicate task information, requirements as well as environment setup to robots on the fly. Interactive Task Learning (ITL) aims to achieve this goal by creating robots that learn from a human instructor through language and demonstration (Gluck and Laird 2019). Teaching inherently requires that the instructor has an accurate mental model of the robot. Creating, maintaining, and improving one's mental model of the robot requires that the robot can use natural interaction patterns that humans engage in. The research question that I investigate is *"How can the robot help the human build a better mental model of the robot?"*

## Current Progress

My research studies interaction mechanisms in ITL systems such as Rosie (Mohan et al. 2012) and AILEEN (Mohan et al. 2020) that learn new tasks and concepts from natural language instruction. Both these agents are implemented in the Soar cognitive architecture. The symbolic nature of Soar makes it a good candidate for exploring transparency and explanation in robots. Also, the learning mechanisms of Soar are task-general therefore it does not depend on the environment for specific training. This allows Rosie to be embodied in a tabletop robot arm with a Kinect sensor, a mobile robot, the Fetch robot, and Cozmo.

As an example, assume that an instructor wants to teach a robot to build a tower using the blue, green and red blocks shown in Figure 1. In the context of a situated interaction, I define the types of information that are relevant to the instructor (Ramaraj and Laird 2018):

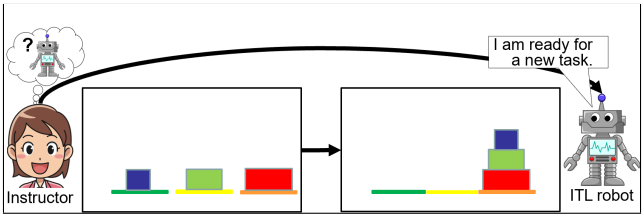


Figure 1: The instructor needs to have a good mental model of the robot learner to teach it how to build a tower.

1. Perception - The instructor needs to know what the robot perceives in its environment. For example, can the robot identify a *blue block* in its environment?

2. Long-term knowledge - This includes the robot's prior learned task knowledge in terms of definitions, task procedures, actions, and goal states. For example, does the robot know the definition of *larger*?

3. Grounded task knowledge - This knowledge is how the robot applies its knowledge to the environment to perform actions or tasks. The instructor needs to know if the robot can successfully build the tower. If it cannot, why not?

Effective instruction includes evaluating the robot's current knowledge, providing definitions and appropriate examples of relevant concepts, understanding the reason for failures when they arise, and fixing the robot's knowledge for its future success. Each part of this process has an intent associated with it and requires the robot to respond to proceed through the task. The research question here is *how do we build robots that leverage the intentionality of the instructors to enable more natural ITL?* To answer this question, we look to Collaborative Discourse Theory (CDT) (Grosz and Sidner 1986). Prior research in Rosie leveraged CDT to enable flexible and mixed-initiative interactive behavior. However, these interactions are largely driven by the robot's learning needs with very little understanding of how humans teach. We propose a taxonomy that organizes human intentions observed in a human-robot teaching scenario (Ramaraj, Klenk, and Mohan 2020). I then conducted semi-structured interviews with nonexpert teachers to validate and

extend this taxonomy. I will use the results to design an interaction framework described later.

One of the challenges with back and forth interactions is the potential for failure. I focus on failures that cause and are caused by the incorrectness of the person's mental model of the robot. Predicting a robot's behavior can be used as a proxy for the quality of a person's mental model (Norman 2014). In prior work, I characterized the features in instructions that help people identify why a situated robot failed (Ramaraj et al. 2019). For example, in Figure 1, let's assume that the instructor mistakenly specifies that "a blue block is on a red block" while describing a part of the tower and the robot responds with "A blue block is not on a red block." It is easy for them to identify why the failure occurred since all the terms in the instruction are commonly used. However, assume that the instructor correctly specifies "a blue block is on a green block" but the robot responds with "I don't see it" because it has only learned the terms *small block* and *medium block*. It would be difficult for the instructor to determine why the robot failed. This is because the robot has only learned these task-specific terms that is unknown to the instructor revealing a gap in their mental model of the robot.

Transparency mechanisms allow people to access the robot's knowledge and improve their mental model (Hayes and Shah 2017; Perlmutter, Kernfeld, and Cakmak 2016). I implemented question-answering and visualization transparency mechanisms in Rosie. Using these, an instructor can ask the robot to describe its environment. When the robot says it sees *a small block and a medium block*, the instructor can understand why the robot failed. I conducted a human subject study where I discovered that people are significantly better at identifying the reason for failures that occur when common terms are used compared to when robot-specific or hidden terms are used. Secondly, in situations that involve robot-specific or hidden terms, transparency mechanisms significantly improved people's accuracy.

When failures occur, a robot's response is crucial because it directly influences the instructor's follow-on instruction and their next steps. In a complex environment where there are many possible reasons for a robot's failure, it can be challenging for an instructor to predict why it failed or know what robot-specific information they need. *How do we design robot responses that improve the accuracy of the instructor's predictions?* To answer this question, I am currently working to learn how changes in robot responses correspond to people's predictions about the robot's knowledge and the failure situation. I will provide participants with different instructor-robot interaction failures and ask them to predict the robot's knowledge and why it might have failed. An example is if the robot cannot see the blue block in Figure 1. If the instructor describes a part of the tower as "a blue block is on a green block," we would present each participant with different robot responses such as "I don't see a blue block," "I don't see that," or "I don't know what a blue block is." I plan to complete this project by February 2021.

## Future Work

Through these projects, I focus on understanding and evaluating nonexpert mental models of robots. My thesis goal is to use this to contribute to human-robot teaching that is more approachable for nonexperts. Towards this goal, I plan to implement an intention-based interaction framework in Rosie using template-based inputs, where templates are included for the intentions identified in the taxonomy. Each intentional input will correspond to an individual turn. For example in Figure 1, the instructor can ask to execute an *inform* intention and demonstrate a move action. In this scenario, Rosie should learn the *move* action from demonstration and provide acknowledgment once it has completed. If the instructor wants to *evaluate* Rosie's competence, they can use Rosie's transparency mechanisms to confirm whether it has learned this action. In a situation where it fails, I will build in mechanisms for Rosie to respond appropriately to help the instructor debug the situation. Through the development of these turn-specific interactions, I hope to make progress towards an end-to-end complete task interaction where the robot helps the human build a better mental model of itself.

## References


Gluck, K. A.; and Laird, J. E. 2019. *Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions*, volume 26. MIT Press.

Grosz, B.; and Sidner, C. L. 1986. Attention, intentions, and the structure of discourse. *Computational linguistics* .

Hayes, B.; and Shah, J. A. 2017. Improving robot controller transparency through autonomous policy explanation. In *Proceedings of the 2017 acm/ieee international conference on human-robot interaction*, 303–312. ACM.

Mohan, S.; Klenk, M.; Shreve, M.; Evans, K.; Ang, A.; and Maxwell, J. 2020. Characterizing an Analogical Concept Memory for Newellian Cognitive Architectures. *arXiv preprint arXiv:2006.01962* .

Mohan, S.; Mininger, A. H.; Kirk, J. R.; and Laird, J. E. 2012. Acquiring Grounded Representations of Words with Situated Interactive Instruction. In *Advances in Cognitive Systems*, volume 2, 113–130. Citeseer.

Norman, D. A. 2014. Some observations on mental models. In *Mental models*, 15–22. Psychology Press.

Perlmutter, L.; Kernfeld, E.; and Cakmak, M. 2016. Situated Language Understanding with Human-like and Visualization-Based Transparency. In *Robotics: Science and Systems*.

Ramaraj, P.; Klenk, M.; and Mohan, S. 2020. Understanding Intentions in Human Teaching to Design Interactive Task Learning Robots. In *RSS 2020 Workshop: AI & Its Alternatives in Assistive & Collaborative Robotics: Decoding Intent*.

Ramaraj, P.; and Laird, J. E. 2018. Establishing Common Ground for Learning Robots. In *RSS 2018: Workshop on Models and Representations for Natural Human-Robot Communication*.

Ramaraj, P.; Sahay, S.; Kumar, S. H.; Lasecki, W. S.; and Laird, J. E. 2019. Towards using transparency mechanisms to build better mental models. In *Advances in Cognitive Systems: 7th Goal Reasoning Workshop*, volume 7, 1–6.