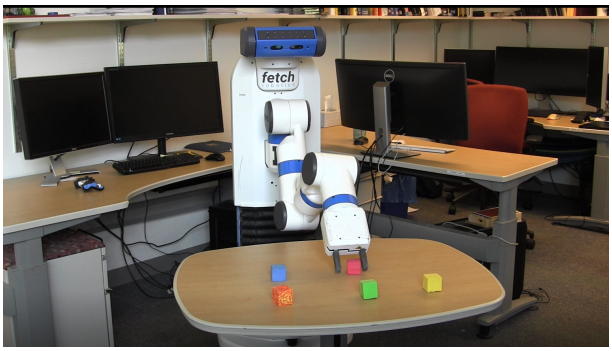
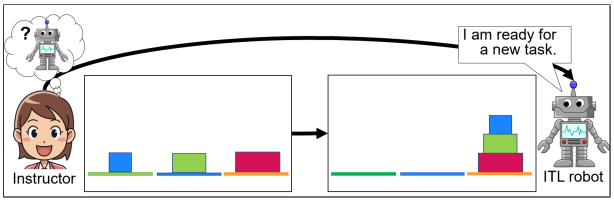
Robots that Help Humans Build Better Mental Models of Robots

Preeti Ramaraj
preetir@umich.edu
University of Michigan
Ann Arbor, Michigan, USA



(a)



(b)

Figure 1: (a) Rosie learns to solve the 5-puzzle problem. (b) The instructor needs to have a good mental model of the robot learner to teach it how to build a tower.

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 we investigate is "How can the robot help the human build a better mental model of the robot?" We study human-robot interaction failures to understand the role of mental models in resolving them. We also discuss a human-centred interaction model design that is informed by human subject studies and plan-based theories of dialogue, specifically Collaborative Discourse Theory.

KEYWORDS

Mental models, Interactive Task Learning, Cognitive systems, Interaction failures

ACM Reference Format:
Preeti Ramaraj. 2021. Robots that Help Humans Build Better Mental Models of Robots. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21 Companion), March 8–11, 2021, Boulder, CO, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3434074.3446365

HRI '21 Companion, March 8–11, 2021, Boulder, CO, USA

ACM ISBN 978-1-4503-8290-8/21/03.
https://doi.org/10.1145/3434074.3446365

1 INTRODUCTION

We envision a future where robots help people by performing a diverse set of tasks including household, rehabilitative, and office tasks. To perform these tasks effectively, people need to communicate task information and 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 [8]. 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 we investigate is *"How can the robot help the human build a better mental model of the robot?"*

2 PRIOR RESEARCH

Our research studies interaction mechanisms in ITL systems such as Rosie [14] and Aileen [13] 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 in robots. Secondly, Rosie maintains abstract symbolic representations of knowledge that can be grounded in different environments. This allows Rosie to be embodied in a tabletop robot arm with a Kinect sensor, a mobile robot, the Fetch robot (Figure 1a), and Cozmo.

As an example, assume that an instructor wants to teach a robot to build a tower with the blue, green and red blocks as shown in Figure 1b. In the context of a situated interaction, we define the types of information that are relevant to the instructor [19]:

- 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?
- 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*?
- 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?

2.1 Unpacking human teaching intentions

Effective instruction includes evaluating the robot's knowledge, providing definitions and appropriate examples of relevant concepts, understanding the reason for failures when they arise, and fixing the robot's knowledge for future success. Each part of this process has an associated intent and requires the robot to respond to proceed through the task. The research question here is *how do we build robots that leverage the instructors' intentions to enable more natural ITL?* To answer this question, we look to Collaborative Discourse Theory (CDT) [9]. Prior research [3, 14, 15] has 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 proposed a taxonomy of human communicative intentions observed in a human-robot teaching scenario [18]. We then conducted semi-structured interviews (N=10) with participants who were asked to teach the task of building a wall to a researcher who assumed the role of a learner agent. We conducted an inductive thematic analysis of these interviews through open and axial coding [2]. Through this analysis we validated and extended our proposed taxonomy. This study provides us with an initial insight into how these intentions emerge in task teaching interactions. We will use these results to design an interaction framework described in section 3.

2.2 Characterizing interaction failures

One of the challenges with back and forth interaction is the potential for failure. Prior work has described multiple types of failures that can occur in situated human-robot interaction [1, 5, 10, 12]. We focus on failures that cause and are caused by the incorrectness of the person's mental model of the robot. Predictions of a robot's behavior can be used as a proxy for estimating the quality of a person's mental model [16]. In prior work, we characterized the features in instructions that help people identify the reason for a situated robot's failure [20]. For example, in Figure 1b, 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 the instructor to identify why the failure occurred since all the terms in the instruction are commonly used. However, if the instructor correctly specifies "a blue block is on a green block" and 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, which are 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 through multiple modalities such as language [6, 12, 17, 21, 22, 24], gaze [4, 17, 23], gestures [4, 7, 17] and visualization [11, 17]. Thus, we implemented question-answering and visualization mechanisms. The instructor can now ask the robot to describe its environment. When the robot specifies it sees *a small block and a medium block*, the instructor can now learn why the robot failed. We conducted a human subject study (N=64) where we discovered that people are significantly better at identifying the reason for failures that occur in interactions with commonly-used terms over those with robot-specific terms. Secondly, in interactions with robot-specific terms, transparency mechanisms significantly improved people's accuracy [20].

3 FUTURE WORK

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 to know what robot-specific information they need. *How can we design robot responses that improve the accuracy of the instructor's predictions?* To answer this question, we are currently working to learn how changes in robot responses correspond to people's mental models in terms of their predictions about the failure situation. Therefore we identify different sources of failure such as perception, world representation and task knowledge (semantic and procedural). We will provide participants with different instructor-robot interaction failures and ask them to predict the robot's current knowledge and why it might have failed. An example is if the robot cannot see the blue block in Figure 1b. 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."

Through these projects, we focus on understanding, evaluating and leveraging human mental models of robots. Our goal is to make robot teaching more approachable for nonexperts. Towards this goal, we will implement an intention-based interaction framework in Rosie using template-based inputs, where templates are included for individual intentions identified in the taxonomy. Each input will mark the beginning of a discourse segment consisting of turns that satisfy a specific intention. For example, in Figure 1b, the instructor can ask to execute an *inform* intention and *describe* a move action. This will begin a discourse segment with the *inform* purpose. Rosie will process the instruction using its knowledge of the task, the instructor, and the shared environment and respond with relevant information. If Rosie successfully learns the task, it must provide verbal acknowledgment, which results in the end of the segment. If Rosie fails in the process, its response should nudge and enable the instructor to *evaluate* its knowledge using transparency mechanisms to debug the situation. Through the development of these turn-specific interactions, we 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

[1] Sean Andrist, Dan Bohus, Ece Kamar, and Eric Horvitz. 2017. What went wrong and why? Diagnosing situated interaction failures in the wild. In *International Conference on Social Robotics*. Springer, 293–303. https://www.microsoft.com/en-us/research/publication/what-went-wrong-and-why-diagnosing-situated-interaction-failures-in-the-wild/

[2] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101. https://biotap.utk.edu/wp-content/uploads/2019/10/Using-thematic-analysis-in-psychology-1.pdf.pdf

[3] Cynthia Breazeal, Guy Hoffman, and Andrea Lockerd. 2004. Teaching and working with robots as a collaboration. In *AAMAS*, Vol. 4. 1030–1037.

[4] Cynthia Breazeal, Cory D Kidd, Andrea Lockerd Thomaz, Guy Hoffman, and Matt Berlin. 2005. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 708–713. http://alumni.media.mit.edu/~guy/publications/BreazealIROS05.pdf

[5] Daniel J Brooks. 2017. *A human-centric approach to autonomous robot failures*. Ph.D. Dissertation. University of Massachusetts Lowell. https://search.proquest.com/docview/1927724896

[6] Joyce Y Chai, Rui Fang, Changsong Liu, and Lanbo She. 2016. Collaborative Language Grounding Toward Situated Human-Robot Dialogue. *AI Magazine* 37, 4 (2016). http://www.cse.msu.edu/~jchai/Papers/AIMagazine2016.pdf

[7] Crystal Chao, Maya Cakmak, and Andrea L Thomaz. 2010. Transparent active learning for robots. In *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*. IEEE, 317–324. https://www.cc.gatech.edu/~athomaz/papers/Chao-hri10.pdf

[8] Kevin A Gluck and John E Laird. 2019. *Interactive Task Learning*. Vol. 26. MIT Press. https://ieeexplore.ieee.org/document/8012335

[9] Barbara Grosz and Candace L Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational linguistics* (1986). https://dash.harvard.edu/handle/1/2579648

[10] Shanee Honig and Tal Oron-Gilad. 2018. Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in psychology* 9 (2018), 861. https://www.semanticscholar.org/paper/Understanding-and-Resolving-Failures-in-Human-Robot-Honig-Oron-Gilad/d48d57abe467393121fe58c6cd60f51f4c8b82a1

[11] Daniel A Lazewatsky and William D Smart. 2012. Context-sensitive in-the-world interfaces for mobile manipulation robots. In *RO-MAN, 2012 IEEE*. IEEE, 989–994. http://people.oregonstate.edu/~smartw/library/papers/2012/roman2012.pdf

[12] Matthew Marge and Alexander I Rudnicky. 2011. Towards Overcoming Miscommunication in Situated Dialogue by Asking Questions.. In *AAAI Fall Symposium: Building Representations of Common Ground with Intelligent Agents*. http://www.cs.cmu.edu/~mrmarge/MargeAAAIfss11.pdf

[13] Shiwali Mohan, Matt Klenk, Matthew Shreve, Kent Evans, Aaron Ang, and John Maxwell. 2020. Characterizing an Analogical Concept Memory for Newellian Cognitive Architectures. *arXiv preprint arXiv:2006.01962* (2020).

[14] Shiwali Mohan, Aaron H Mininger, James R Kirk, and John E Laird. 2012. Acquiring Grounded Representations of Words with Situated Interactive Instruction. In *Advances in Cognitive Systems*, Vol. 2. Citeseer, 113–130.

[15] Anahita Mohseni-Kabir, Charles Rich, Sonia Chernova, Candace L. Sidner, and Daniel Miller. 2015. Interactive Hierarchical Task Learning from a Single Demonstration. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (Portland, Oregon, USA) *(HRI '15)*. Association for Computing Machinery, New York, NY, USA, 205–212. https://doi.org/10.1145/2696454.2696474

[16] Donald A Norman. 2014. Some observations on mental models. In *Mental models*. Psychology Press, 15–22.

[17] Leah Perlmutter, Eric Kernfeld, and Maya Cakmak. 2016. Situated Language Understanding with Human-like and Visualization-Based Transparency.. In *Robotics: Science and Systems*. https://homes.cs.washington.edu/~lrperlmu/perlmutter16rss.pdf

[18] Preeti Ramaraj, Matt Klenk, and Shiwali Mohan. 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*.

[19] Preeti Ramaraj and John E Laird. 2018. Establishing common ground for learning robots. In *RSS 2018: Workshop on Models and Representations for Natural Human-Robot Communication*. http://www2.ece.rochester.edu/projects/rail/mrhrc2018/papers/ramaraj05.pdf

[20] Preeti Ramaraj, Saurav Sahay, Shachi H. Kumar, Walter S. Lasecki, and John E. Laird. 2019. Towards using transparency mechanisms to build better mental models. In *Advances in Cognitive Systems: 7th Goal Reasoning Workshop*, Vol. 7. 1–6.

[21] Stephanie Rosenthal, Manuela Veloso, and Anind K Dey. 2012. Acquiring accurate human responses to robots' questions. *International journal of social robotics* 4, 2 (2012), 117–129. http://www.cs.cmu.edu/~mmv/papers/12ijsr-RosenthalVelosoDey.pdf

[22] Stefanie Tellex, Ross A Knepper, Adrian Li, Daniela Rus, and Nicholas Roy. 2014. Asking for Help Using Inverse Semantics. Robotics: Science and Systems Foundation. http://cs.brown.edu/courses/csci2951-k/papers/tellex14.pdf

[23] Andrea L Thomaz and Cynthia Breazeal. 2008. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence* 172, 6-7 (2008), 716–737. https://www.sciencedirect.com/science/article/pii/S000437020700135X

[24] Emily Wu, Nakul Gopalan, James MacGlashan, Stefanie Tellex, and Lawson LS Wong. 2016. Social Feedback For Robotic Collaboration. (2016). https://h2r.cs.brown.edu/wp-content/uploads/2016/07/wu16thesis.pdf