Advances in Cognitive Systems 7 (2019) 1-6 Submitted 5/2019; published 8/2019

Towards using transparency mechanisms to build better mental models

Preeti Ramaraj PREETIR@UMICH.EDU
Saurav Sahay SAURAV.SAHAY@INTEL.COM
Shachi H. Kumar SHACHI.H.KUMAR@INTEL.COM
Intel Labs, Santa Clara, CA 95054 USA

Walter S. Lasecki WLASECKI@UMICH.EDU
John E. Laird LAIRD@UMICH.EDU
Computer Science and Engineering, University of Michigan, Ann Arbor, MI 48109 USA

Abstract

Non-expert users often do not have accurate mental models of robots, leading to failures during interactive task learning. We first observe that the information required to identify the cause of interaction failures, can be classified into commonly-defined, uncommonly-defined and hidden features. We then implemented two transparency mechanisms, question-answering and visual explanation capabilities, through which a non-expert user can access the robot's internal reasoning. We conducted a user study (N=64) where we measured the ability of the user to identify the reason for 12 distinct interaction failures. We first characterize the impact of the identified features of interaction failures, and then test the effect of the two transparency mechanisms on the user's performance. Results show that transparency mechanisms have the most influence on interaction failures with uncommonly-defined and hidden features. These results confirm that transparency mechanisms are effective in improving the user's performance, leading to a higher accuracy in the user's mental model of the robot.

1. Introduction

Interactive Task Learning (ITL, Laird et al. 2017) is a method through which a robot can be taught complete tasks using a combination of natural language instructions and demonstrations. The ITL paradigm is motivated by the need to enable non-expert users to teach robots new tasks through interaction alone. Two issues contribute to failures in non-expert user and robot interaction. First, current robots do not have sufficient skills to estimate and contribute relevant information to the user based on relevant context. This issue is commonly avoided because robots interact with developers or experts who have an accurate understanding (mental model) of the robot's capabilities, knowledge, intentions and shortcomings. Here we run into our second issue - a non-expert user cannot be expected to have an accurate mental model of the robot. However, without such knowledge, non-experts will often be unable to establish common ground with the robot. If we can characterize the knowledge gaps in the non expert's mental model, we could potentially tackle the problem of ad-

dressing these failure better. Thus, we characterize the information required to fill these knowledge gaps into three possible classes - commonly-defined, uncommonly-defined and hidden features, that constitute the interaction failures that occur in a typical human robot interaction.

Previous work has established that successful interaction requires clear and effective communication from all participants to establish and maintain *common ground* (Clark & Wilkes-Gibbs, 1986). We explore effective communication on the robot's end in the form of transparency mechanisms. In order to learn how transparency mechanisms can help bridge gaps in the user's mental model of the robot, we implemented two types of transparency mechanisms in an existing ITL robot. The first allows the robot to answer questions in English about its perception, long-term knowledge, and instantiation of its knowledge to its environment. The second allows the user to request the properties, relations and concepts pertaining to each individual object being presented, to be updated visually on the screen.

We conducted a user study, where we presented a set of 12 interaction failures that occur in a human-robot setting to the user. We measured accuracy and confidence of the user in identifying the reason for the interaction failure. Our contributions are as follows:

- We characterize the impact of features inherent to these interaction failures on the users' performance in terms of accuracy and confidence.
- We present results that show the transparency mechanisms we implemented help improve non-expert's mental models, particularly in interaction failures with hidden or uncommonly-defined features. We measure this improvement through the user's accuracy in identifying the reason for the interaction failure.

2. Related Work

Previous research has shown how a robot providing its uncertainty about its decisions, perception or its environment lets the user understand (Wu et al., 2016; Chai et al., 2016) and provide feedback that helps the robot's learning process (Gouravajhala et al., 2018; Tellex et al., 2014; Thomaz & Breazeal, 2008; Chao et al., 2010; Rosenthal et al., 2012). Since our robot is capable of learning through language, we decided to use task-general question-answering mechanisms as a way to leverage its language capabilities, similar to the research presented in (Hayes & Shah, 2017).

Other research work (Perlmutter et al., 2016; Lazewatsky & Smart, 2012) has used visualization-based transparency mechanisms such as screen, VR as well as highlighting objects in the environment to provide users with additional information. Given this precedent, we wanted to study how visual explanation methods would work alongside language mechanisms of the robot.Our goal was to look at how these mechanisms influence users' understanding of the robot.

Donald Norman states that prediction is a major indicator of the quality of the user's mental model. Even if the users cannot be expected to describe their mental model accurately, their successful prediction of a robot's behavior would indicate a higher accuracy in their mental model of the robot (Norman, 2014). Thus, we evaluate the user's mental model across multiple interaction failures by measuring their performance in terms of accuracy and confidence.

3. Instructable Robot

We use an existing ITL robot Rosie (Mohan, 2015) that is implemented in the Soar cognitive architecture (Laird, 2012) and is embodied in three different robots: a tabletop robot arm with a Kinect sensor, a mobile robot, and a Fetch robot. Rosie can learn over 50 games and puzzles (Kirk & Laird, 2016), mobile delivery and navigation tasks (Mininger & Laird, 2016; Kirk & Laird, 2016) and procedural kitchen-specific tasks (Mohan et al., 2012) using a combination of situated natural language commands and demonstration. In previous work (Ramaraj & Laird, 2018), we identified knowledge that is relevant to establishing common ground with an ITL robot as follows:

- *Perception*: Perception refers to the internal model of the environment that the robot builds up using its observations of the environment. Access to the robot's perception allows the user to maintain common ground with the robot, about its labels of objects in their shared environment.
- *Long Term Knowledge*: The robot's long-term knowledge includes goal definitions, actions, failure conditions and task predicates. Access to the robot's long term knowledge helps the user understand task-specific terms for future use.
- *Instantiated Task Knowledge*: In order for the robot to perform a task, it needs to be able to apply its learned task knowledge to its environment. This is so that the robot can determine whether it has reached its goal, or if it has failed in the process, and what actions it can do in that situation. Access to this information provides the instructor with an opportunity to access the internal reasoning of the robot.

4. Question-answering mechanisms

Our first transparency mechanism allows the user to ask the robot questions about its perception, long term knowledge as well as the instantiation of that knowledge in its environment. To learn what the robot perceives in the environment, the user can ask "What do you see?". In response to this question in Figure 1 (a), the robot responds *A green loc_b location is below the blue block. A blue loc_c location is below the green block. A green loc_a location is below the red block. A blue block is on the green loc_b location. A green block is on the blue loc_c location. A red block is on the green loc_a location.* The user can also ask the robot about individual objects in terms of their properties and relations.

The robot can also describe its learned task knowledge. For example, let us assume the robot learned the definition of "matched" as "if the color of a location is the color of the block that is on the location then the location is matched." When the user asks the robot "What is 'matched'?", the robot describes it as *"If the color of an object is the color of a block and the block is on the object, then it is matched."*

5. Screen-based visual explanation mechanisms

The question-answering mechanisms can prove cumbersome when there are continuous scene changes that the user needs to keep track of. Here, we explore the idea of visual updates to a screen, where a user can keep track of the robot's perception and instantiation of knowledge. Figure 2 shows the robot's simulated perception. In Figure 1(a), one can observe that the user has continual access to

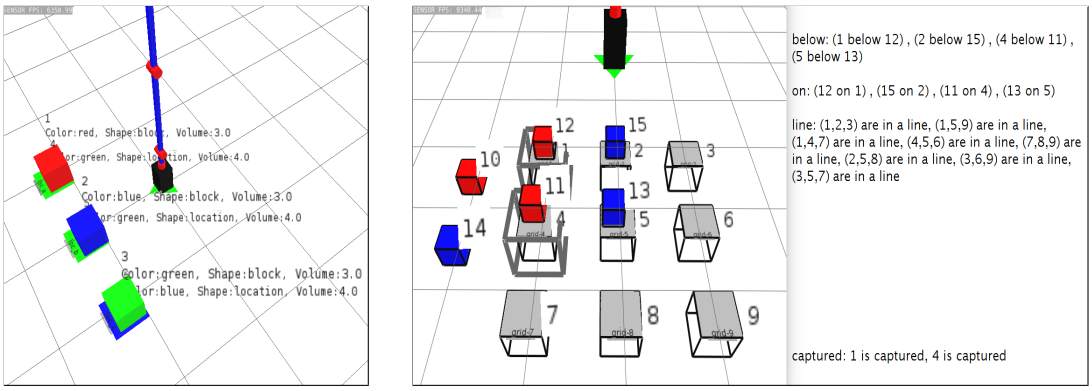

Figure 1: (a) A representation of the robot's perception of its environment annotated with the robot's labeling of objects, properties. (b) The *captured* locations have been highlighted in the environment and updated in the side panel.

the properties of objects. This would be useful for the user who wishes to learn the robot's model of the world in order to interact with it. In Figure 1(b), we show the implementation of a side panel, where the relations between objects in the simulated environment are displayed, and updated continually based on changes in the environment.

Finally we implemented mechanisms that allow the user to request to highlight objects that satisfy learned task predicates. This way, the user has access to the robot's instantiation of learned predicates. For example, Figure 1(b) shows the robot highlighting the objects in the environment that are *captured*. Here, if the object is below a red block, then it is *captured*. The robot updates the simulator by drawing a bigger 3d box around the listed objects and updates the side panel by listing the IDs of the objects that are *captured* at the bottom.

6. User Study

We conducted a user study on Mechanical Turk using LegionTools (Lasecki et al., 2014) in order to test the effects of the implemented transparency mechanisms on the user's mental model of the robot. We measure the user's accuracy and confidence in identifying the reason for 12 different mentor-robot interaction failures. We identified 13 features in the information required to identify the cause of failure in these 12 examples. We then classified them into the following broad three categories

- Commonly-defined features: These include properties such as 'color', binary relations or learned predicates like 'clear'. In interaction failures with these features, there is no need of any updates to the user's mental model.
- Uncommonly defined features: These are properties of the world for which the robot's representation can be fairly arbitrary, thus lending itself to misinterpretation or lack of information on the user's end. These include properties such as 'size' or relations such as 'diagonal'.
- Hidden features: Hidden features can be commonly or uncommonly defined features that are used to define learned predicates, but cannot be inferred from the predicate alone. Here, the

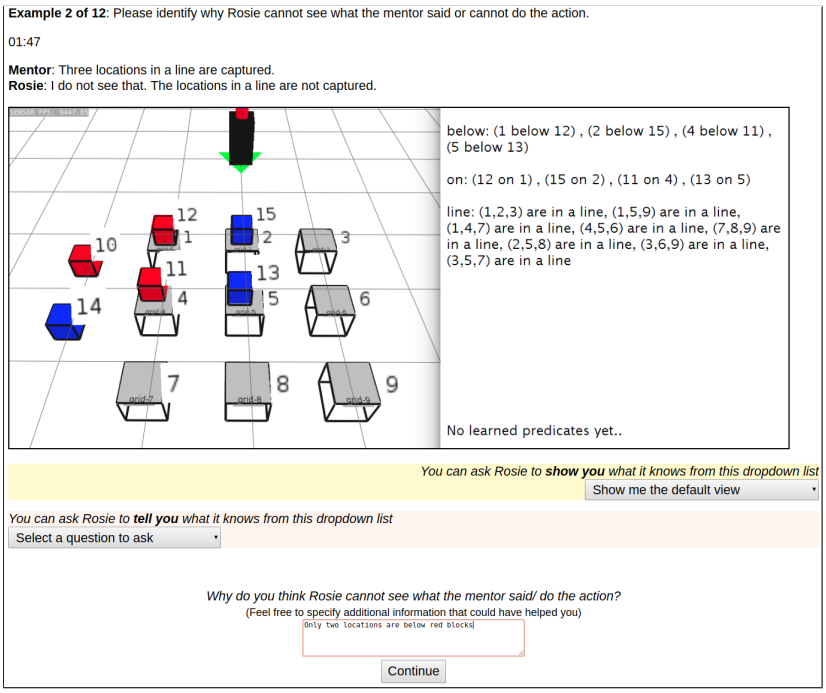

Figure 2: Mechanical Turk user interface presented to the user

user needs to explicitly update their mental model with the robot's definition. For example, in figure 1 (b), the definition of *captured* uses both the features of color as well as binary-below relation, but it cannot be inferred.

We hypothesize that *the average accuracy and confidence of the users will be significantly higher in examples of failures with only commonly-defined features in comparison to the failures with hidden or uncommonly defined features*.

In order to study the effect of transparency mechanisms on the user's mental model, we varied their availability on the user interface. Each user encountered examples with the following conditions: no transparency mechanisms available, only TM1(Q-A mechanisms) available, only TM2 (visual explanation mechanisms) available and both TM1 and TM2 available. We replicated the robot capabilities in the web interface by pre-populating the visual explanations and question-answer combinations for each interaction failure such that it could be requested by the user, depending on the condition.

We had a total of 64 participants, 59 (33 male, 26 female) of whom completed the post-survey questionnaire and provided demographic information. The age range distribution of our participation was: 18-24: 6, 25-34: 26, 35-44: 16, 45-54: 10, 55-64: 1. Figure 2 shows the interface of a specific example where both transparency mechanisms were available, where the interaction presented is as follows:

Mentor: Three locations in a line are captured.
Rosie: I do not see that. Three locations in a line are not captured.

Here, the user needs to identify why Rosie failed to understand the mentor's instruction. The user is expected to use the available transparency mechanisms to learn about the robot's understanding of the environment. If none are available, they are expected to guess why this failure occurred. Our goal was to *learn whether the transparency mechanisms contribute to higher accuracy and confidence*. We ask the user to provide their potential reasoning for what led to the interaction failure - through a description in their own words, as well as selecting an option from a multiple-choice drop-down list of available reasons. There is only one correct option per example. The user is also asked to provide their confidence in their selected answer on a Likert scale of range 1-5.

7. Study Results

We used t-tests, ANOVA and post-hoc Tukey's HSD tests to test our hypotheses about the normally distributed accuracy data, and non-parametric Wilcoxon Rank Sum tests to test our hypotheses about the non-normally distributed confidence data. We calculated the accuracy by dividing the number of correct responses by the total number of responses per example. We confirm our first hypothesis that in the None condition, the average accuracy of the users in failures with commonly-defined features (M = 77.72%, SD = 17.2%) is significantly higher than the accuracy in failures with hidden or uncommonly-defined features (M = 28.9%, SD = 14.2%), t(10)=5.35, $p < 0.001$. The confidence of the users in failures with commonly-defined features (Median = 5) is also *significantly higher* than the confidence of the users in failures with hidden or uncommonly-defined features (Median = 4), W(378, 363)=85320, $p < 0.001$.

Our second hypothesis is regarding the claim that transparency mechanisms would result in a significant difference in the average accuracy and confidence on interaction failures with hidden and uncommonly-defined features. A one-way Anova test shows that there is a significant difference in the means of the accuracy measured across the transparency mechanisms, F(3, 6)=7.392, $p < 0.01$. The post-hoc Tukey's HSD tests show that the mean accuracy in the Q-A condition (M=53.3%, SD=6.8%) and Both condition (M=55.5%, SD=12.7%) are significantly greater than the mean accuracy in the None condition (M = 28.9%, SD = 14.2%). We do not observe significant difference between the Visual(M=43.7%, SD=8.01%) and None condition. After using Bonferroni correction for multiple tests, with a modified alpha of 0.016, we observe a significant difference between the user's confidence in the Q-A condition (Median=5) and the None condition (Median=4), W(87,95)=5006, $p < 0.01$. We did not observe significant difference between the pairs of None and Both(Median=5), and None and visual(Median=4) condition.

8. Conclusion

The goal of this work was to learn how one could improve a non-expert user's mental model during interaction failures using transparency mechanisms. We learned that **commonly-defined features are more understandable**. We were able to verify this using the user's average accuracy and confidence which was higher compared to failures with uncommonly-defined or hidden features. We also learned that **transparency mechanisms help when terms are unclear**: We were able to observe a significant difference in accuracy when the users had access to Q-A and both mechanisms in comparison to the None condition, in case of failures with uncommonly-defined or hidden features.

References

Chai, J. Y., Fang, R., Liu, C., & She, L. (2016). Collaborative language grounding toward situated human-robot dialogue. *AI Magazine*, *37*.

Chao, C., Cakmak, M., & Thomaz, A. L. (2010). Transparent active learning for robots. *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on* (pp. 317–324). IEEE.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*, 1–39.

Gouravajhala, S. R., Yim, J., Desingh, K., Huang, Y., Jenkins, O. C., & Lasecki, W. S. (2018). Eureca: Enhanced understanding of real environments via crowd assistance. *Sixth AAAI Conference on Human Computation and Crowdsourcing*.

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

Kirk, J. R., & Laird, J. E. (2016). Learning general and efficient representations of novel games through interactive instruction. *Advances in Cognitive Systems*, *4*.

Laird, J., et al. (2017). Interactive task learning. *IEEE Intelligent Systems*, *32*, 6–21.

Laird, J. E. (2012). *The soar cognitive architecture*. MIT press.

Lasecki, W. S., Gordon, M., Koutra, D., Jung, M. F., Dow, S. P., & Bigham, J. P. (2014). Glance: Rapidly coding behavioral video with the crowd. *Proceedings of the 27th annual ACM symposium on User interface software and technology* (pp. 551–562). ACM.

Lazewatsky, D. A., & Smart, W. D. (2012). Context-sensitive in-the-world interfaces for mobile manipulation robots. *RO-MAN, 2012 IEEE* (pp. 989–994). IEEE.

Mininger, A., & Laird, J. (2016). Interactively learning strategies for handling references to unseen or unknown objects. *Adv. Cogn. Syst*, *5*.

Mohan, S. (2015). *From verbs to tasks: An integrated account of learning tasks from situated interactive instruction.*. Doctoral dissertation, Department of Computer Science and Engineering,University of Michigan, Ann Arbor.

Mohan, S., Mininger, A. H., Kirk, J. R., & Laird, J. E. (2012). Acquiring grounded representations of words with situated interactive instruction. *Advances in Cognitive Systems* (pp. 113–130). Citeseer.

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

Perlmutter, L., Kernfeld, E., & Cakmak, M. (2016). Situated language understanding with human-like and visualization-based transparency. *Robotics: Science and Systems*.

Ramaraj, P., & Laird, J. E. (2018). Establishing common ground for learning robots. *RSS 2018: Workshop on Models and Representations for Natural Human-Robot Communication*.

Rosenthal, S., Veloso, M., & Dey, A. K. (2012). Acquiring accurate human responses to robotsâĂŹ questions. *International journal of social robotics*, *4*, 117–129.

Tellex, S., Knepper, R. A., Li, A., Rus, D., & Roy, N. (2014). Asking for help using inverse semantics. Robotics: Science and Systems Foundation.

Thomaz, A. L., & Breazeal, C. (2008). Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence*, *172*, 716–737.

Wu, E., Gopalan, N., MacGlashan, J., Tellex, S., & Wong, L. L. (2016). Social feedback for robotic collaboration.