Bi-Directional Mental Model Reconciliation for Human-Robot Interaction with Large Language Models

Nina Moorman, Michelle Zhao, Matthew B. Luebbers, Sanne Van Waveren, Reid Simmons, Henny Admoni, Sonia Chernova, and Matthew Gombolay

Abstract

In human-robot interactions, human and robot agents maintain internal mental models of their environment, their shared task, and each other. The accuracy of these representations depends on each agent's ability to perform theory of mind, i.e. to understand the knowledge, preferences, and intentions of their teammate. When mental models diverge to the extent that it affects task execution, reconciliation becomes necessary to prevent the degradation of interaction. We propose a framework for bi-directional mental model reconciliation, leveraging large language models to facilitate alignment through semi-structured natural language dialogue. Our framework relaxes the assumption of prior model reconciliation work that either the human or robot agent begins with a correct model for the other agent to align to. Through our framework, both humans and robots are able to identify and communicate missing task-relevant context during interaction, iteratively progressing toward a shared mental model.

Introduction

Mental models are abstract representations of reality, used for reasoning about cause and effect, and for making decisions in an individual's environment (Wilson and Rutherford 1989). Though the term originates from human psychology, it can also be applied to robotic agents to describe their formalized world and task models, programmed to support autonomous decision-making (Tabrez, Luebbers, and Hayes 2020). Prior work in human factors has shown that the degree of mental model synchronization between collaborators on a task is correlated with team performance (Mathieu et al. 2000). To achieve this synchronization, humans rely on their theory of mind capacity to infer the mental models of their teammates through observation, communicating when disagreements are identified (Andrews et al. 2023). To achieve fluent human-robot teaming, we must develop systems with a similar capacity for identifying and reconciling mental model discrepancies during interaction.

Prior human-robot model reconciliation methods have typically been uni-directional: either a robot's model is aligned with an expert human's model (e.g., in learning from demonstration (Argall et al. 2009)), or a human's model is aligned with an expert robot's model (e.g., in autonomous decision support or behavior elicitation/coaching (Tabrez, Agrawal, and Hayes 2019; Sreedharan, Chakraborti, and Kambhampati 2021)). However, in real-world human-robot interactions, the diversity of environments and users means neither the human nor the robot is likely to start with a complete mental model for the task.

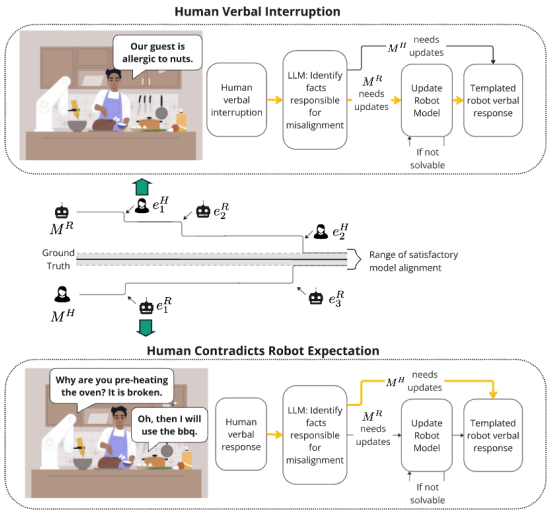

Figure 1: In our pipeline, the robot and human can prompt mental model reconciliation via natural language.

We propose a framework for bi-directional mental model reconciliation between human and robotic agents. The framework facilitates iterative updates of both human and robot models through semi-structured natural language dialogue, initiated either by verbal interruptions from the human or upon the observation of human actions that contradict the robot's expectation. This iterative process allows both humans and robots to share knowledge and preferences during the interaction, and gradually form a shared, mutually satisfactory mental model for the task.

The proposed contribution of our work is the following:

1. A theoretical framework for bi-directional human-robot mental model reconciliation.
2. An instantiation of that framework which represents the robot's model via Planning Domain Definition Language (PDDL), represents shared mental model context as structured facts (Knepper et al. 2017), and leverages a

large language model (LLM) to process natural language dialogue between the human and robot agents.

3. A human-subjects experiment evaluating the performance of the proposed method for facilitating iterative model updates via natural language communication.

Methodology

Problem Formulation In our setup, a human-robot team shares a collective task, specified within a ground-truth task context, c^{GT}. In practice, c^{GT} comprises knowledge involving the task, environment, and each agent's capabilities and preferences, such that the task can be completed to each agent's satisfaction. Neither agent is assumed to fully know c^{GT}; instead, each begins with their own understanding of the context, c_0^R and c_0^H.

The robot and human mental models, M^R and M^H, combine each agent's current context with a decision-making capacity. Throughout the interaction, M^R yields both a policy for the robot to follow π^R, and a prediction of the human's policy $\pi^{R(H)}$. Likewise, M^H yields a human policy π^H and predicted robot policy $\pi^{H(R)}$.

The solution to the bidirectional model reconciliation problem is a set of explanations $E^R \cup E^H = \{e_1^R, ...e_n^R\} \cup \{e_1^H, ...e_m^H\}$, that minimizes $d(\pi^{H(R)}, \pi^R) + d(\pi^{R(H)}, \pi^H)$, with each explanation aimed at communicating missing contextual information to the other agent, thus updating that agent's mental model. The reconciliation is deemed complete when $d(\pi^{H(R)}, \pi^R) < \epsilon$, and $d(\pi^{R(H)}, \pi^H) < \epsilon$.

Research Questions In this work, we investigate the following research questions.

1. **RQ1)** As a function of the number of iterations, how does bidirectional model reconciliation impact the accuracy of the robot's and the human's mental model, as compared to ground truth?
2. **RQ2)** As a function of the number of iterations, how does bidirectional model reconciliation impact the alignment between the robot's and the human's mental model?
3. **RQ3)** As a function of the number of iterations, how does bidirectional model reconciliation impact user attitudes towards and perceptions of the robot?

Approach Our proposed approach is depicted in Figure 1. To evaluate our framework, we implement the robot mental model M^R using a common planning language (PDDL (Aeronautiques et al. 1998)); solving the planning problem affords π^R and $\pi^{R(H)}$. The human mental model M^H represents the human's internal decision-making. To facilitate the alignment of task-relevant context, we represent c_t^R and c_t^H as sets of facts (fact-based models) that reflect knowledge believed by an agent, similar to Knepper et al. (2017).

Given their initial fact-based model contexts, the human and robot formulate their respective plans and begin executing them concurrently. Model reconciliation is initiated in two ways: (1) when the human interrupts with a verbal utterance and (2) when the robot notices a deviation from expected human behavior ($\pi^{R(H)} \neq \pi^H$). In this second case, the robot provides a templated verbal interruption that communicates the anticipated and actual human behavior, asking the human to clarify the discrepancy.

Upon receiving either the interruption or the clarification from the human, the pipeline employs an LLM to input the human's utterance, and output whether the robot or human contexts are missing information, and what fact(s) could be added to either to rectify the discrepancy. If the robot's context has been updated, another LLM takes the new c_t^R, and returns an updated robot mental model M^R. Once updated, the robot provides a templated verbal explanation of the update. On the other hand, if the human's context has been updated, the robot provides the human with a templated verbal explanation of the new fact(s). Finally, the human is asked to restate what the robot has indicated, ensuring mutual understanding of the respective model updates.

Proposed Evaluation We propose a human subject experiment to evaluate the accuracy of and alignment between the robot and human mental model, and to investigate the resulting user perceptions of and attitudes toward the robot. After obtaining participants' consent and demographics, the human and robot are each given an initial mental model. In this work, we conduct mental model reconciliation in the cases where both mental models contain correct but incomplete information. To accomplish the collaborative task, the human and robot must identify when their mental models lack information, prompt the other agent, and exchange the missing information.

We define the ground truth mental model as the union of the facts initially given to the robot and the human. To evaluate mental model accuracy we report the edit distance[1] between the ground truth mental model and the final human mental model. To evaluate the alignment between the robot and human mental models, we report the edit distance between the two fact-based models, and visualize the changes in edit distance over time.

Our evaluation domain involves organizing and hosting a dinner party, with tasks such as picking a dish, cooking, setting the table, and loading the dishwasher. We propose to evaluate our mental model reconciliation system in scenarios where either, both, or neither models have missing information. At the end of each task, a post-task questionnaire is administered that measures the human's perceived task success, and the human's mental model using the Situation Awareness Global Assessment Technique (SAGAT) (Endsley 1988) over the content of the fact-based model. At the end of the study, we administer a questionnaire that measures the human's perceptions of and attitudes toward the robot, including perceived workload (Hart 1986), acceptance (Belanche, Casaló, and Flavián 2012), and trust (Jian, Bisantz, and Drury 2000).

[1] We define edit distance here as the number of facts that would need to be edited such that the two mental models are the same.

References

Aeronautiques, C.; Howe, A.; Knoblock, C.; McDermott, I. D.; Ram, A.; Veloso, M.; Weld, D.; Sri, D. W.; Barrett, A.; Christianson, D.; et al. 1998. Pddl— the planning domain definition language. *Technical Report, Tech. Rep.*

Andrews, R. W.; Lilly, J. M.; Srivastava, D.; and Feigh, K. M. 2023. The role of shared mental models in human-AI teams: a theoretical review. *Theoretical Issues in Ergonomics Science*, 24(2): 129–175.

Argall, B. D.; Chernova, S.; Veloso, M.; and Browning, B. 2009. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5): 469–483.

Belanche, D.; Casaló, L. V.; and Flavián, C. 2012. Integrating trust and personal values into the Technology Acceptance Model: The case of e-government services adoption. *Cuadernos de Economía y Dirección de la Empresa*, 15(4): 192–204.

Endsley, M. R. 1988. Situation awareness global assessment technique (SAGAT). In *Proceedings of the IEEE 1988 national aerospace and electronics conference*, 789–795. IEEE.

Hart, S. G. 1986. NASA task load index (TLX).

Jian, J.-Y.; Bisantz, A. M.; and Drury, C. G. 2000. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics*, 4(1): 53–71.

Knepper, R. A.; Mavrogiannis, C. I.; Proft, J.; and Liang, C. 2017. Implicit communication in a joint action. In *Proceedings of the 2017 acm/ieee international conference on human-robot interaction*, 283–292.

Mathieu, J. E.; Heffner, T. S.; Goodwin, G. F.; Salas, E.; and Cannon-Bowers, J. A. 2000. The influence of shared mental models on team process and performance. *Journal of applied psychology*, 85(2): 273.

Sreedharan, S.; Chakraborti, T.; and Kambhampati, S. 2021. Foundations of explanations as model reconciliation. *Artificial Intelligence*, 301: 103558.

Tabrez, A.; Agrawal, S.; and Hayes, B. 2019. Explanation-based reward coaching to improve human performance via reinforcement learning. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 249–257. IEEE.

Tabrez, A.; Luebbers, M. B.; and Hayes, B. 2020. A survey of mental modeling techniques in human–robot teaming. *Current Robotics Reports*, 1: 259–267.

Wilson, J. R.; and Rutherford, A. 1989. Mental models: Theory and application in human factors. *Human factors*, 31(6): 617–634.