Computational Mechanisms for Mental Models in Human-Robot Interaction

Matthias Scheutz

Department of Computer Science
Tufts University, Medford, MA 02155, USA
matthias.scheutz@tufts.edu
WWW home page: http://hrilab.tufts.edu/

Abstract. Mental models play an important and sometimes critical role in human-human interactions, in particular, in the context of human team tasks where humans need to interact with each other to achieve common goals. In this paper, we will describe some of the challenges involved in developing general computational mechanisms for mental models and their applications in the context human-robot interactions in mixed initiative tasks.

1 Introduction

The rapid advances in robot technology over the last two decades has enabled a shift in robot applications from very confined and constrained industrial settings to more open unconstrained human-like environments (from hospitals and elder-care settings, to offices and households). Increasingly, robots are also envisioned to become part of "mixed-initiative" teams, where humans and robot need to collaborate to achieve common goals. A case in point is the initial 2011 "National Robotics Initiative" funding program of the National Science Foundation in the US which explicitly targets "innovative robotics research and applications emphasizing the realization of [..] co-robots acting in direct support of and in a symbiotic relationship with human partners", where "co-robot" is a term specifically coined to denote robots that "work beside, or cooperatively with, people".[1]

Common to the wide range of co-robot applications (from search and rescue missions in disaster zones, to space exploration scenarios) is the role of the robot as a genuine helper, a true team member that acts in the interest of the team in a reliable and effective manner, just as a human team member would. Obviously, this is a very high bar for robots to meet, for many reasons. First of all, the robot has to be able to perform the task-based activities for its envisioned role. For example, a search and rescue robot might have to be able to perform triage on a wounded person or at least be able to instruct another human on how to do it [7]. This, itself, could be very challenging, e.g., if complex manipulation capabilities are required such as administering a syringe as part of the first-aid procedure or

[1] See http://www.nsf.gov/pubs/2011/nsf11553/nsf11553.htm

repairing broken pipes and valves in a power plant accident. Moreover, the robot has to be able to function reliably and autonomously for possible long periods of time without human intervention (e.g., in an underground sewer system, the rubble of collapsed buildings, or an extra-terrestrial space station setting with limited to no network connectivity).

While both of the above challenges will require sustained research efforts for years to come, there is at least one other challenge that has not received sufficient attention yet, even though it is potentially even more difficult to address than the two engineering challenges: that of *effective, natural human-robot interaction* (HRI) [11]. "Natural HRI" here means that humans will be able to communicate and interact with robots in the mixed initiative settings *as if* those robots were humans. For example, in a search and rescue scenario, a human commander might have to interact in natural language with a robot remotely located in a collapsed building about where to search for victims. And even though the dialogue exchanges will be focused on the task at hand, the underlying architectural and computational requirements on the robotic side for enabling even simple task-based natural dialogues are quite astounding [10].

Of the many open research problems in natural human-robot interaction, we will focus on a critical underlying capability that affects almost every aspect of the robotic control system: *the robot's ability to build and maintain mental models of other (human and robotic) team members*. We start by briefly reviewing the context of mental models in human teams and then describe the challenges involved in developing computational mechanisms for robotic control architectures for mixed initiative scenarios. Specifically, we will point to several different kinds of representations that are needed for building sufficiently accurate mental models of team members that, in turn, allow for making predictions and coordinating, at least up to some level, joint activities.

2 Motivation: Team Mental Models

It is well-known from extensive research in human teaming that for team members to coordinate their activities effectively and achieve overall high performance, they need to keep track of each other's mental states such as goals and subgoals, intentions, beliefs, and various others – a process commonly subsumed under the term "shared mental models" (e.g., [6]). "Shared mental models" are related to and could be subsumed under the concept of "mental model" as used in psychology, where it typically refers to the types of hypothesized knowledge structures humans build in order to make sense of their world, to make inference based on the available information and to make predictions about future states (e.g., [8]). However, while mental models research in psychology has more focused using mental models to explain various types of human reasoning, mental models in the context of teams have more to do with establishing and maintaining *common ground* (e.g., in Clark's sense [4]) and building *team mental models* [3] that aid in decision-making and the adjustment of one's behavior based on predictions made about the other team members' future activities and actions.

Thus, team mental models are critical for making sense of team activities, for understanding the dynamic changes of team goals and team needs, the possibly dynamic changes of roles and functions of team members, and the overall state of the team.

While there is increasing evidence, especially from research in management and organizational behavior, that humans build and use team mental models, and that using them appropriately will lead to improved team performance, little is known about what these mental models look like, i.e., what information is represented about other team members, their intentions and goals, their knowledge and beliefs, and their activities. Moreover, it is not clear exactly how these representations influence the various cognitive processes involved in coordinating team activities, from task-based natural language dialogues, to distributed task and role assignments, to team decision-making. Yet, these aspects are critical for understanding the functional mechanisms of team mental models at a level that would allow for the development of similar mechanisms in robots. Note, however, an important difference in the aim of implementing team mental models in robots compared to providing computational models of human team mental models, say: since robotic and human cognitive architectures are (currently) very different, we do not claim nor intend to provide a computational account of how humans build and use team mental models; rather, we will attempt to lay out some of the important data representations and processes that have the potential to improve robot operation and might make robots better team mates.

3 How to formulate and represent mental models in robots

Mental models, to be usable in robotic architectures, effectively need to be broken down into two parts: the *data representations* that capture information about the task, the other team members, and the environment, and the *computational processes* that operate on these data structures to create, maintain, revise and discard them. The former will likely differ from task to task, while the latter are intended to be more general mechanisms that can be used across tasks. To be able to motivate the required data structures and processes better, we will use a team search task from our prior work [5].

The specific team search task requires at least two humans, a remotely located commander and at least one searcher located in the target environment that is to be searched. All team members must coordinate their activities through spoken natural language interactions via wireless audio links as their only interaction modality (given that they are spatially separated). The commander has a rough map of the target environment, while the searcher does not have any map. Neither commander nor search have been in the environment before (in our experiments, the indoor search environment consisted of several rooms and a surrounding hallway). The team has two main tasks to begin with (later a third task was added): (1) the searcher has to inform the commander of any encountered green boxes as the commander has to mark their locations on the

map; (2) the commander has to direct the searcher through the environment to a particular location where the searcher can pick up a container which then can be used to collect colored blocks from blue boxes located in the environment.

The following example is one of many from our CReST corpus [5] showing how commander and searcher collaborate via natural language dialogues to agree on actions in the interest of their task goals:

```
Commander: Okay.  Go through the open door and towards the steps
           that are right in front of you before the steps take
           a right.
Searcher:  Okay, right.  Right on the steps there's a green box
           number two.
Commander: Oh, number two right on the steps.
Searcher:  Yeah.
Commander: Okay, I got it.  Alright.  If you're looking at the
           steps you take a right there should be another open
           door.
Searcher:  So, don't actually go up the steps?
Commander: Don't actually go up the steps.
Searcher:  Okay.  Yep, I see the door.
```

Notice how the commander instructs the searcher where to go by describing the environment from the searcher's perspective based on a mental model of where the searcher is in the environment. These instructions frequently comprise descriptions of salient aspects of the environment as gleaned from the map such as "go through the open door and towards the steps that are right in front of you". Here, the director's description relies on a mental model of where the searcher is located and what the environment would look like from the searcher's perspective.

Also notice how the searcher interrupts the activity when she notices the green box on the stairs and communicates that to the director. This requires both parties to keep track of their overall activities as subactivities related to some of the task goals are initiated. After the commander confirms by repeating back the number and location of the box, the search confirms, and the subactivity finishes. This requires both parties to resume the previous activity of the commander directing the searcher through the environment. Here it is interesting to note that the searcher had a different goal in mind (namely to go up the stairs) from what the commander intended (to turn right), and the instruction to "take a right" by the commander revealed the goal discrepancy to the searcher, thus prompting the searcher to ask explicitly about the next action ("don't actually go up the steps?"), which in turn revealed her goal to the commander. Again, this interruption constitutes a sub-task of achieving goal alignment and consolidating the mental models of both parties about what the next actions and subgoals are. Once, agreement is reached, the searcher informs the commander of the fact that she can see the door, which directly answers the previous "hedged explain" dialogue move by the commander that "there should

be another open door". These types of dialogue moves specifically require the interactant to take verification action (i.e., to verify that there is another open door) and confirmation or disconfirmation action (i.e., that the open door was verified) [5].

All of the above dialogue interactions serve the dual purpose of establishing what actions to take next (i.e., common subgoals in the interest of the overall task goals) and to establish *common ground* between searcher and commander [4]. Common ground here comprises several aspects such as where the searcher is located, what the searcher's perspective is, what objects the searcher can see, what goals the searcher/commander has, etc. Common ground is negotiated through dialogue interactions which eventually are finished with acknowledgments of both parties and various dialogue-based linguistic mechanisms are employed to communicate understanding or lack of confidence.

Aside from important dialogue-based mechanisms for negotiating goals and activities which are interesting in their own right, the above scenario points to several important aspects that mental models need to capture: (1) *facts*, about the task, events, objects, and the environment, including aspects about the location of team members; and (2) *beliefs*, about goals, activities, and beliefs about team members. The difference between facts and beliefs here is that facts are taken to be true from an agent's perspective while the content of beliefs might not be true. For example, the searcher believed that the goal was for her to go up the steps, while the commander intended for the searcher to turn right. For her to be able to detect the belief inconsistency, she had to represent this goal and keep it in her mental model of the commander. When the commander then gave an instruction that suggested another incompatible goal, the inconsistency was discovered. Hence, it is critical to represent the other agents' perspectives, which will allow for better detection of inconsistencies and thus improved alignment among agents.

In addition to the rich data representations, it is important for an agent to allow for belief maintenance and belief revision processes that can both synchronize beliefs and update them when new evidence arrives. This does not only apply to the beliefs of other agents as represented in the agent's mental models, but also to the facts the agent holds true (for it is possible that these factual representations were obtained from insufficient or flawed evidence, misinterpreted communications, etc.). Moreover, an agent not only has to correct her own inconsistent facts and beliefs, but those corrections might, in turn, trigger corrections of other agents' facts and beliefs (e.g., if the agent learned that a previously communicated fact is not true).

4 How to build and update mental models

The above interactions clearly showed that representations used to build mental models need to be sufficiently expressive to be able to capture goals, beliefs, desires, and knowledge-based states (such as rules and facts), as well as various modal operators (e.g., about beliefs of other agents and their beliefs about other

agents, but also possibilities, obligations, permissions, etc. of actions, but also). Here, we will build on our previous pragmatic and mental modeling framework [1, 2] to be able to discuss some of updating processes needed for mental models. To simplify the discussion, we will use $[[..]]_c$ to denote the "pragmatic meaning" of an expression (e.g., a natural language utterance) in some context c, which in team tasks typically includes task and goal information, as well as belief and discourse aspects. And we will use α to denote the agent under consideration (i.e., the agent whose whose perspective we will take for the discussion of how to update the agent's mental model).

Overall, updates to an agent's mental model will be triggered by various events, mediated through the agent's perceptual system. For example, the agent might perceive a new task-relevant object (as in the case of the searcher above noticing a green box). Assuming that the agent will store this perception, we can formulate a general principle that if α perceives an object o at location l at time t, then α will believe (B) that it perceived o at l at t:

$$Perceives(\alpha, o, l, t) \Rightarrow B(\alpha, Perceives(\alpha, o, l, t))$$

This principle can then be applied to all agents on the team. As a special case, if α perceived another agent β, it will form the belief that it perceived β and it might also form the belief that β perceived α. Similar principles can be defined for actions to capture their enabling conditions and effects (i.e., pre- and post-conditions).

More interesting are belief updates that have to do with perceptions that are communications, i.e., when α receives an update from β through an utterance Utt in a given context c. Given that all team members are collaborating on a common set of task goals and thus have no incentive to purposefully mislead or deceive their team members, we can assume that all communicated propositions $\phi \in [[Utt]]_c$ are true (at least from β's perspective). It is then necessary for α to check whether any of the communicated propositions are inconsistent with α's existing beliefs. For this purpose, α needs to employ an inference algorithm $\Rightarrow_\alpha^b$ that will, up to some bound b, check for inconsistencies.[2] Any inconsistent conflicting beliefs are then removed from the agent's sets of beliefs Bel_α (e.g., in the above example, the searcher would remove the goal to go the stairs as a result of the commander's correction telling the searcher to go right instead).

Moreover, α believes all propositions it can infer from (again within bound b) from propositions $\phi \in [[Utt]]_c$:

$$([[u]]_c \Rightarrow_\alpha^b \phi) \wedge Heard(\alpha, u) \Rightarrow B(\alpha, \phi)$$

By the same token, α also believes everything it says:

[2] Note that the bound b is intended to capture both the agent's reasoning limitations as well as other time-based limitations based on the current context.

$$([[u]]_c \Rightarrow^b_\alpha \phi) \wedge Said(\alpha, u) \Rightarrow B(\alpha, \phi)$$

In addition to updating its own beliefs, α needs to model any other agent γ also hearing β's utterance, i.e., α has to derive its mental model $\{\psi | B(\gamma, \psi) \in Bel_\alpha\}$ for all other agents $\gamma \neq \alpha$ and update it by using them same rules it applies to its own beliefs. The same is true if α notices that another agent has certain perceptions or performs certain actions. In general, α needs to update all its models whenever there is a change in its own beliefs (either through addition of a new fact or revision of a known one).

In team tasks, the situation often arises that another agent "intends to know" (IK) a proposition, i.e., either makes an explicit request to be informed or provides some other information from which such a request can be derived. The set of all propositions other agents want to know, Φ_{IK}, can be defined as:

$$\psi \in \Phi_{IK} \Leftrightarrow \exists \beta, \phi : \psi \in Bel_\alpha \wedge IK(\beta, \phi \in Bel_{alpha}) \wedge (\psi \Rightarrow^b_\alpha \phi \vee \psi \Rightarrow^b_\alpha \neg\phi)$$

This set of proposition is then part of what agent α needs to communicate to other agents, in addition to the set of all propositions that will correct false beliefs Φ_{rev} that agents holds, defined as:

$$\psi \in \Phi_{rev} \Leftrightarrow \exists \beta, \phi : B(\beta, \phi) \wedge \phi \in Bel_\alpha \wedge (\psi \Rightarrow^b_\alpha \neg\phi)$$

The above principles have been integrated in a cognitive robotic architecture for human-robot interaction and tested in simple human-robot interactions [1, 2]. However, a thorough experimental evaluation with naive subjects has not yet been performed.

5 Discussion

The previous sections briefly sketched the kinds of principles necessary for building mental models in computational architectures, in particular, what representations to build and maintain, how to update them based on events and internal changes (e.g., results of inferences), and when to communicate them to provide answers to requests or correct false beliefs of other agents. While the principles were stated in fairly general ways, there are important aspects of their computational implementation that were only hinted at. For example, computing deductive closures of any finite set of beliefs is usually not feasible for reasonable inference rules, hence a "bound" was imposed on the inference procedure. However, even such a bound might still make various computations intractable. E.g., if there is a large number of agents with many different belief and knowledge items, then even keeping track of all of them and detecting inconsistencies among them without any inference might not be feasible in a reasonable amount of time. On the other hand, it is doubtful that humans could do that either. Hence, restricting the above principles to small, human-like teams might be sufficient for addressing the computational overhead.

A related question that can also lead to computational escalation is the level of nested beliefs an agent should keep track of (e.g., it is necessary to keep track of beliefs of beliefs of beliefs?). The answer here is largely pragmatic, for in some cases this type of information may not even be available or if it is, it might be too short-lived to be of interest (e.g., if a third agent observed the dialogue in the human search task between the search and the commander about whether the searcher should go up the stairs, it could have represented the goals of the searcher and the commander before, during, and after the dialogue, and updated them throughout as it became clear that the searcher had to revise its goal; but none of that processing might have impacted the activity of this third agent). Hence, there is a critical notion of "relevance" to one's own activity as part of the team that should be taken into account when building and updating mental models. What this notion is and how it can be cast in algorithmic terms is, however, an open question at this point.

Other interesting questions are related to the extent to which multiple agents will manage to keep their mental models synchronized depending on their communication abilities and the task demands. It might be possible to impose hierarchical structures that would allow agents to focus on a subset of the team members and only track their activities and goals. Such hierarchical structures could also help curb the computational overhead of mental model processing.

Finally, it is also important to consider ways to evaluate the efficacy and efficiency of mental models employed on robots. Here evaluation measures that have been used in human teams might be applicable (e.g., [9]) to the extent that the measures are objective and can be answered directly from observations. Subjective measures (such as questions asked of human team members after the tasks) could also be employed as long as they can be obtained from logged information about architecture-internal states (e.g., the set of beliefs about other agents beliefs could be examined with respect to the accuracy of those beliefs throughout the task).

6 Conclusions

In this paper, we discussed some of the computational mechanisms required for building and maintaining mental models of team members in mixed-initiative human-robot teams. We started by examining typically task-based natural language interactions in human teams and derived some representational and processing requirements from those interactions. We then introduced a formal framework for representing and updating mental models in a way that they can be integrated into a robotic architecture. Finally, we discussed some of the challenges involved in employing such mental models given the computational and real-time constraints of robots working with humans and pointed to some interesting future directions that will require extensive experimentation and modeling of human-robot teams in order to determine the extent to which mental models need to be built and maintained in order to improve team performance.

References

1. Briggs, G. and Scheutz, M. (2012). "Multi-modal Belief Updates in Multi-Robot Human-Robot Dialogue Interactions". *Proceedings of AISB 2012*.
2. Briggs, G. and Scheutz, M. (2011). "Facilitating mental Modeling in Collaborative Human-Robot Interaction through Adverbial Cues". *Proceedings of 12th Annual SIGDIAL Meeting on Discourse and Dialogue*, 239–247.
3. Cannon-Bowers, J. A., and Salas, E. (1990). "Cognitive psychology and team training: Shared mental models in complex systems." Paper presented at the *Annual meeting of the Society for Industrial and Organizational Psychology*, Miami, FL.
4. Clark, H. H. (1996). *Using language* Cambridge University Press, Cambridge.
5. Eberhard, K., Nicholson, H., Kübler, S., Gundersen, S., and Scheutz, M. (2010). "The Indiana 'Cooperative Remote Search Task' (CReST) Corpus". In *Proceedings of Language Resources, Technologies and Evaluation LREC 2010*.
6. Mathieu, John E., Heffner, Tonia S., Goodwin, Gerald F., Salas, Eduardo and Cannon-Bowers, Janis A. (2000). "The Influence of Shared Mental Models on Team Process and Performance". *Journal of Applied Psychology*, 85,2, 273–283.
7. Harriott, Caroline E., Zhang, Tao, and Adams, Julie A. (2011). "Evaluating the applicability of current models of workload to peer-based human-robot teams." In *Proceedings of the 6th international conference on Human-robot interaction*, 45–52.
8. Johnson-Laird, Philip N. (1983) *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press
9. Lim, Beng-Chong and Klein, J. Katherine (2006). "Team mental models and team performance: A field study of the effects of team mental model similarity and accuracy." *Journal of Organizational Behaviour* 27, 403–418.
10. Scheutz, M., Cantrell, R. and Schermerhorn, P. (2011) "Toward humanlike task-based dialogue processing for human robot interaction" *AI Magazine*, 32, 4, 77–84.
11. Scheutz, M., Schermerhorn, P., Kramer, J. and Anderson, D. (2007). "First Steps toward Natural Human-Like HRI" *Autonomous Robots*, 22, 4, 411–423.