Toward Genuine Robot Teammates: Improving Human-Robot Team Performance Using Robot Shared Mental Models

Felix Gervits
Tufts University
Medford, MA
felix.gervits@tufts.edu

Dean Thurston
Tufts University
Medford, MA
Dean.Thurston@tufts.edu

Ravenna Thielstrom
Tufts University
Medford, MA
Ravenna.Thielstrom@tufts.edu

Terry Fong
NASA Ames Research Center
Moffett Field, CA
terry.fong@nasa.gov

Quinn Pham
Tufts University
Medford, MA
quinn.pham@tufts.edu

Matthias Scheutz
Tufts University
Medford, MA
matthias.scheutz@tufts.edu

ABSTRACT

Effective coordination is a critical requirement for human teaming, and is increasingly needed in teams of humans and robots. Building on decades of work in the behavioral literature, we have implemented a computational framework for coordination based on Shared Mental Models (SMMs) in which robots use a distributed knowledge base to coordinate activity. We also built a novel system connecting the robotic architecture, DIARC, to the 3D simulation environment, Unity, to serve as an evaluation platform for the framework implementation, and also for more general explorations of teaming with autonomous robots. Using this platform, we ran a user study to evaluate the framework by comparing performance of teams in which the robots used SMMs with those that did not. We found that teams in which the robots used SMMs significantly outperformed those without SMMs. This represents the first empirical demonstration that SMMs can be successfully used by fully autonomous robots interacting in natural language to improve team performance, bringing robots a step closer to genuine teammates.

KEYWORDS

shared mental models; coordination; human-robot teaming

1 INTRODUCTION

Robots are uniquely suited to performing dull, dirty, or dangerous tasks that can complement and enhance the work that humans do. As a result, they are increasingly needed to serve as joint partners with humans in a variety of domains, ranging from hospital operating rooms to space stations. The role of robots on human teams goes beyond serving as mere tools for information retrieval, but extends to scenarios in which the robots function as equal partners, or peers [39]. In terrestrial teams, effective human-robot teaming is needed for military applications, such as scouting and reconnaissance [23], disaster relief applications such as urban search and rescue [26], and assistive applications such as medical care [42]. Robots in these kinds of teams have a host of interaction requirements, including interpreting natural language instructions, carrying out actions in support of the team, and adapting to novel situations. Robots for space exploration have some of the most challenging requirements, as they need to coordinate their activities with humans as part of highly distributed teams that operate at multiple spatial ranges, time scales, and interaction modalities [9, 10]. These robots must be able to work before, during, and after human activity, and must have autonomous capabilities to operate with limited to no human intervention [13]. These challenges highlight the need for a coordination framework that can manage the interaction demands of these complex domains.

2 SHARED MENTAL MODELS FOR HUMAN-ROBOT COORDINATION

A promising approach to coordination is based on the concept of Shared Mental Models (SMMs), which are distributed knowledge structures that human teams build and maintain for effective coordination [4, 20, 25]. Decades of work from Organizational Psychology has demonstrated that SMMs serve to improve team coordination and performance in a variety of task domains [8, 21, 24, 38]. In a multi-agent human-robot interaction (HRI) context, an SMM can be between the robots (*Robot SMM*) or between the human-robot team (*Human-Robot SMM*). A Robot SMM represents the set of shared knowledge that is synchronized across all robots, and used to inform planning and decision-making. This allows for a level of alignment far beyond a human-human SMM (e.g., sharing internal representations). On the other hand, the Human-Robot SMM is the broader structure representing the common ground of all agents on the team, including each agents' knowledge and belief states and the extent to which they are aligned. This paper focuses on Robot SMMs, to specifically address the *SMM hypothesis for artificial agents*, which is that *SMMs improve coordination and performance in human-agent teams* [34].

Several lines of computational work have incorporated elements of SMMs into their approaches, though none have been sufficient to truly test the SMM hypothesis. In [45] and [13], agents share a knowledge base and use this to coordinate and make team-oriented decisions. While [13] used a comprehensive SMM framework based on [34] (including mental state representation, functional roles, obligations and norms, etc.), [45] used only select parts of an SMM, including team process, team structure, domain knowledge, and dynamic information needs. Moreover, both of the evaluations used

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 2020, Auckland, New Zealand

https://doi.org/doi

simulated agents (though [13] used a human-in-the-loop), which does not sufficiently demonstrate improved human-robot team performance. A proper evaluation would involve a user study in which humans interact with autonomous robots in a collaborative task. Another approach used shared mental representations for collaborative planning, with the robot inferring a human's plan and using it to inform its actions [40]. This is closer to a true Human-Robot SMM, although again, the scope of the SMM is limited and the evaluation used a simulated human in a proof-of-concept scenario. Finally, [27] used a POMDP to encode an SMM, however the formal specification of the SMM was limited to high-level variable descriptions such as "mental model variables" and "activity status variables", and the proposed assembly manufacturing task was not actually performed.

Recently, we introduced the first comprehensive formal and computational framework for SMMs (both Robot- and Human-Robot SMMs) [34]. This framework enables artificial agents to store and share a variety of representations, including agent capabilities and propensities, agent and task states, norms and obligations, activities and equipment types, and functional roles of agents in teams. Agents maintain these representations, update them based on inference and perception, and use them to adapt behavior. This framework is uniquely suited for testing the SMM hypothesis because it supports the same kinds of comprehensive representations that human teammates use, and informs how to incorporate these representations into a control architecture. While prior work has implemented only fragments of an SMM [40, 45], or failed to carry out a proper evaluation [13, 27] an implementation and evaluation of this framework can provide a true test of the SMM hypothesis.

3 METHOD

Building on our prior work [13, 17, 34], we first implemented and extended the SMM framework in a space robotics task domain (see Sec. 3.1 and 3.2). Next, we designed an evaluation platform in which robots running a cognitive robotic architecture could interact with humans in a virtual environment (see Sec. 3.3). Using this platform, we ran a user study to evaluate the SMM framework by comparing performance of teams in which the robots used SMMs with those that did not (see Sec. 4).

3.1 Space Robotics Task Domain

The scenario we developed involves a collaborative maintenance intravehicular activity (IVA) task on a spacecraft, and is meant to loosely represent a use case for humanoid robots in space [5]. In the scenario, a human plays the role of an astronaut aboard a spacecraft orbiting Mars. The human must work together with a rover on the planetary surface to scout a territory for a future colonization mission. This involves performing a geological survey task in which the rover reports the location of landmarks on the surface (various rock types and radiation zones) and the human marks the coordinates on a map. While the rover task is the primary task, the human must also attend to a distractor task in which various tubes aboard the spacecraft become damaged and need to be repaired. Repairing damaged tubes requires coordination with two onboard autonomous robots. The human must first locate and shut off a damaged tube, and then instruct a robot to go to the tube and repair it. The human can then turn the tube back on after it has been repaired.

We created an environment using the Unity game engine to represent this task domain. The virtual spacecraft layout includes a central area (containing the rover task map) connected to three wings labeled Alpha, Beta, and Gamma. Each wing is identical and includes a hallway with 24 tubes that (people were told) contain fuel for a future Mars colony (12 on each side; see Fig. 1). The tubes are identified by the wing that they are in, the side that they are on, and the number, e.g., alpha left one. Only the robot can enter the hallway with the tubes as the fuel is "radioactive".

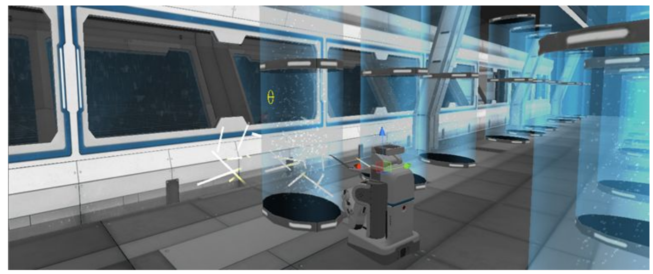

Figure 1: Wing layout showing a robot repairing a tube.

3.2 Framework Implementation

Knowledge in the SMM framework is represented using a set of logical primitives from several major categories, including *domain knowledge, agent capabilities, agent and task states, norms and obligations, activities*, and *functional role of agents in teams*. Below, we introduce the formal representations that were developed for each of the major categories as well as general rules that use these representations to update the SMM state.

Domain Knowledge is represented as a set of predicates including the agents, objects, locations, activities, and other domain-specific representations needed to model the task and environment. The set of agents, $A = \{H, R_1, R_2\}$, includes the human, *H*, and the two robots R_1 and R_2. Objects include the set of 72 tubes, *T*, where a tube $t \in T$ is identified by the tuple <*wing, side, number*> (where *wing* ∈ {alpha, beta, gamma}, *side* ∈ {left,right}, and *number* ∈ {1-12}), as well as its damage level *X*, *propertyOf(t,damaged(X))*, and its status (whether it is on or off), *propertyOf(t,status)*. The set of locations, *L*, represents areas in the spacecraft to which the robots can move; these include each of the tubes, as well as the entryway to each of the wings. If an agent, $a \in A$, is at one of these locations, $l \in L$, it is represented with the predicate *at(a,l)*.

Agent capabilities are represented using $Capable(a, X)$, which signifies that agent *a* is capable of carrying out action *X*. For example, R_1 is capable of monitoring the Alpha wing:
Capable(R_1,*monitorWing*(R_1,*Alpha*)). In general, the robots are capable of carrying out the actions listed in Table 1. *Perceivable*(a,ϕ,σ) signifies that agent *a* can perceive whether the proposition ϕ is true in situation σ. For example, robots can perceive if a tube is off if they are at that tube:
Perceivable(R_1,*propertyOf((alpha,left,one),off)*,*at*(R_1,*(alpha,left,one)*)).

Table 1: Robot actions, including preconditions and effects.

Action	Precondition(s)	Effect(s)
goTo(a,l)	$\neg at(a,l)$ $\neg repairing(a,t)$	*at(a,l)* $\neg movingTo(a,l)$
monitorWing(a,l)	*at(a,l)* $\neg repairing(a,t)$	*monitoring(a,l)*
repair(a,t)	*at(a,t)* *propertyOf(t,damaged(X))* *propertyOf(t,off)*	$\neg propertyOf(t,damaged(X))$ $\neg repairing(a,t))$

Agent[1] and Task States are represented using predicates for knowledge (*Knows(a,ϕ)*) and belief (*Believes(a,ϕ)*). In general, knowledge is used to represent the agent's own state, and belief is used to represent aspects of the task environment and the other agent. The predicate *Knows-Of(a,X)* is used to represent knowledge of the existence of agents, actions, etc. The robots in our domain know about one another and the starting status of the tubes on the spacecraft: *Knows-Of(R_1,R_2)*, *Knows-Of(R_1,(alpha,left,one),on)*, etc. *Common-Belief(ϕ)* is used to represent beliefs shared by all agents. This is used primarily in the SMM condition where robots share a knowledge base (see Sec. 4.5). Goals are represented using *Goal(a,γ)*, where γ is a goal state that includes the effects of actions in Table 1.

Obligations and norms are used to represent obligatory and permissible actions of the agents. We use *Superior(a_1,a_2)* to represent the command hierarchy of the team. In our domain, *Superior(H,R_1)* and *Superior(H,R_2)* indicate that the human is a superior to the robots. In addition, the predicates *Proposes(a_1,a_2,X)*, *Accepts(a_1,a_2,X)*, and *Rejects(a_1,a_2,X)* allow us to represent the outcome of a command from one agent to another. Following this, we introduce a rule that requires subordinates to accept the command of a superior if they are available for that goal (i.e., the pre-conditions are met): *Proposes(a_1,a_2,γ)* $\wedge$*Superior(a_1,a_2)* $\wedge$ *Available-For(a_2,γ)* $\Longrightarrow$ *Accepts(a_2,a_1,γ)*. This ensures that the robots will always accept the human's command if they can, including in situations when such a command will override their autonomy policy.

Activities are the set of actions that can be performed by the agents. The human can move to various areas in the spacecraft, instruct or request information from the robots, place landmarks on the rover map, and turn tubes on and off. The robots are capable of performing a number of actions including moving to a wing or tube, monitoring a wing (checking which tubes are damaged), and repairing a tube. Robot actions are represented by their preconditions and effects, and are defined in Table 1. We define a number of rules to enable the robots to track actions in progress, including: *Goal(a,monitorWing(a,l))* $\Longrightarrow$ *monitoring(a,l)*, *Goal(a,goTo(a,l))* $\Longrightarrow$ *movingTo(a,l)*, and *Goal(a,repair(a,t))* $\Longrightarrow$ *repairing(a,t)*. The robots use these inferred representations in their autonomy policy, described in Sec. 4.4. Finally, functional roles on the team are defined based on the corresponding goals, requirements, capabilities, actions, and obligations of each agent (see [34]). Team structure is defined in terms of the roles, command hierarchy, and equipment.

[1]Representations of human knowledge and goal states are not included here, but are an important topic for future work (see Sec. 5.3).

3.3 Evaluation Platform

In order to evaluate the SMM framework in our task domain we developed an evaluation platform consisting of several components described below. We use the DIARC robot architecture [33] for natural language understanding (NLU), inference, and action selection. The Robot Operating System (ROS) [31] is used for path planning and navigation, and the Unity 3D game engine for visualization. We created an interactive 3D environment in Unity to simulate the interior of a spacecraft based on our task domain (see Sec. 3.1). Virtual robot models of the PR2 robot by Willow Garage are used to ensure that our evaluation results are extensible to physical robots in the real-world. Architecture components are implemented in the Agent Development Environment (ADE), which is a middleware for the DIARC architecture. For this task, we developed a configuration of DIARC which connects the set of components shown in Fig. 2[2]. For an overview of the architecture, see [35], but here we highlight the key components that were modified for our task.

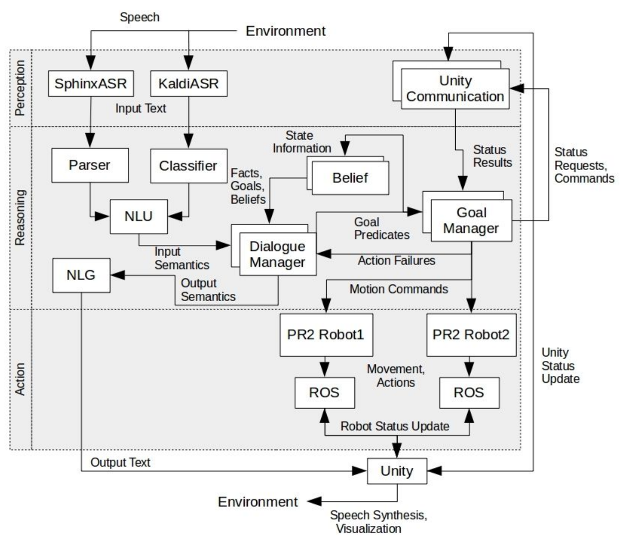

Figure 2: DIARC architecture diagram for the task domain. Components with multiple "tabs" were shared by both robots in the SMM condition, whereas in the Non-SMM condition, each robot had its own instance of these.

In order to add flexibility to the NLU, we implemented a dual-NLU pipeline into ADE. In one branch of the pipeline, an incoming utterance is transcribed to text by an automatic speech recognition (ASR) component using a custom in-domain language model in the *Kaldi* toolkit [30]. The text is then sent to the *Classifier* component, which uses the NPC-Editor [22] software to predict the semantic meaning of the text in logical predicate form, having been trained on a data set of in-domain text-to-semantics links (see [14] for a similar approach). In the other parallel branch of the NLU pipeline, the utterance is transcribed by the *Sphinx4* ASR component [43], which matches words using a dictionary and grammar constructed from the task domain. The resulting text is parsed by the *Parser*

[2]While the diagram only shows two robots, the architecture can scale to any number.

component, which uses a symbolic rule-based parser to produce a semantic translation in predicate form. An additional NLU component merges the two branches of the pipeline by choosing between the two possible semantics based on parser confidence. The advantage of this approach is that simple, structured utterances (e.g., “Move to area Alpha") are processed quickly by the parser, whereas atypical utterances (e.g., “Um can you like come to Alpha") are processed by the classifier.

Another key component in the architecture is the *Goal Manager*. The Goal Manager handles the high-level goals of the robots and manages the execution of actions which are used to achieve those goals. The actions in Table 1 were defined for the task, and the Goal Manager interfaces with several components to carry them out, including the *PR2 Robot* components for low-level movement, the *Unity Communication* component for providing/requesting information from the simulation environment, and the *Belief* component for updating the knowledge base with the effects of actions.

4 HUMAN-SUBJECTS EVALUATION

To evaluate the benefit of a Robot SMM on team performance, we conducted a between-subjects user study in which teams of humans and robots performed a collaborative task. In one condition, humans were partnered with two robots that had an SMM, and in another condition the two robots did not have an SMM (see Sec. 4.5 for more details about the conditions). We sought to test the SMM hypothesis by evaluating whether Robot SMMs improve performance.

To maximize immersion, we ran the study in a Mechdyne virtual reality (VR) CAVE[3] (see Fig. 3 for CAVE setup). Participants wore eye tracking glasses, which tracked gaze location and also head position for the CAVE. A game controller was used to navigate and interact with the environment [4].

4.1 Task Overview

The task domain from Sec. 3.1 was implemented in our Unity environment. The rover task involves the participant marking the location of three types of rocks (sedimentary, sandstone, and basalt) and a radiation zone on a map based on coordinates verbally announced by a planetary rover. The “rover" is just a script that reports landmarks on a set interval (one every 60 s). The abbreviation for the corresponding rock (or radioactive symbol for radiation zones) appears above the map along with the coordinates, and this information stays on top of the map until the next landmark announcement.

4.2 Participants

Overall, 36 participants were recruited from a University campus through posted fliers, online advertisements, and snowball sampling. The study was approved by the Tufts University Institutional Review Board. All participants gave informed consent and were randomly assigned to one of the two conditions (SMM or Non-SMM). Participants received a base rate of $10 for their time, plus an additional $5 as a performance incentive if they achieved an accuracy of 75% or greater on the rover task *and* a task duration of 480 s (8 minutes) or greater.

[3] https://www.mechdyne.com/hardware.aspx?name=CAVE
[4] Video of the task recorded from the mobile eye tracker can be found at the following link: https://vimeo.com/360632866.

4.3 Procedure

After reading and signing the consent form, participants took a preliminary survey consisting of demographic questions. Next, participants were set up with the equipment, including SMI mobile eye-tracking glasses, a Shure push-to-talk wireless microphone, and a wireless Xbox controller (see Fig. 3). The experimenter then read the task instructions, which included the backstory and how to interact with the environment and the robots. Participants were then given a tutorial in which they were allowed to practice any part of the task as long as they needed until they were ready to start. In general, the experimenter made sure that participants could at least navigate and perform all aspects of the primary and distractor tasks, including talking to the robots.

Next, participants performed the task. A typical trial saw participants using the game controller to navigate between the central map and the various wings, and giving verbal instructions to repair damaged tubes. The minimum task duration was 327 s and the maximum was 1015 s. At the end of the task, participants completed a survey in which they answered questions about their workload, situational awareness (SA), and attitudes about the robots (see Sec. 4.6). Finally, payment of $10 was given, plus an additional $5 performance incentive if they scored above the set threshold.

Figure 3: Task setup in the VR CAVE showing the human navigating in the central area. The rover map is visible in the center, as is one of the robots, and the entryway to two wings, Alpha and Beta.

4.4 Robot Behavior

4.4.1 Robot Autonomy. Robots in the task performed the actions defined in Table 1 autonomously. The robots used a “supervisory control" policy [1] in which all actions are carried out independently, but the human can intervene if needed. If the human does intervene, then the robots will always carry out the human’s instruction.

The autonomy policy that the robots use is shown in Algorithm 1. First, the robot checks if it has received a new command (line 2).

If not, it monitors [5] the current wing for damaged tubes (line 3). If one or more damaged tubes are detected then the robot sorts them by damage and moves to the most damaged tube (lines 4-7). When at the tube, the robot checks if the tube is off, in which case it repairs the tube (lines 8-9). If the tube is still on, then it notifies the human about this (line 11) and moves to the next most damaged tube. If there are no damaged tubes in a wing or all damaged tubes have been visited, the robot moves to the next wing that does not contain a robot (lines 15-16). The robots are programmed to wait 5 s after finishing an autonomous action before moving on to the next action, and the human can interrupt them at any point to issue an overriding command.

Algorithm 1 Robot Autonomy Policy

```
 1: procedure AUTONOMY(Human H, Robot self, Location l)
 2:     while ¬Proposes(H,self,X) do                ▷ No new command
 3:         monitorWing(self,l) ▷ monitor wing for damaged tubes
 4:         if damaged tubes are found then
 5:             tubes = ordered list of damaged tubes in wing
 6:             for all t ∈ tubes do
 7:                 goTo(self,t)          ▷ Move to most damaged tube
 8:                 if propertyOf(t,off) then        ▷ If off, repair it
 9:                     repair(self,t)
10:                 else
11:                     Notify H that tube is on
12:                 end if
13:             end for
14:         else
15:             l = next wing not containing a robot
16:             goTo(self,l)
17:         end if
18:     end while
19: end procedure
```

4.4.2 Human-Robot Communication. Robots cannot directly communicate among themselves (explicitly or implicitly), but they are capable of engaging in dialogue with the human in mixed-initiative interaction. The human communicates with both robots on the same channel, though robots only parse messages addressed to them. Since there are two robots, the human must preface their utterance with “Robot One" or “Robot Two" to address them. Robots can take commands to perform any of the actions in Table 1. They will provide feedback after accepting an instruction as a form of grounding, e.g., “Robot two here, okay I am moving to area Alpha". This is useful to inform people if the robot understood the command or if an ASR/parsing error occurred.

Robots can also answer a variety of status inquiries including “Where are you?", “What are you doing?", “Which tubes are damaged?", and others. Importantly, the robots can be told facts about themselves and the other robot that they will assert to their knowledge base. For example, Robot One can be told that “Robot Two is moving to Alpha" or “Tube alpha left four has been repaired". This allows the function of the SMM to be preserved in the Non-SMM condition, albeit with the additional requirement of updating the robots each time an action occurs. Robots will accept the knowledge they are given unless it conflicts with something they know to be true. This only occurs in the SMM condition when given knowledge about the other robot, and in the Non-SMM condition when given knowledge about itself. Finally, robots can initiate dialogue to inform the human about their current action. They do this at the onset of each action (e.g., “Robot Two here, I am moving to area Alpha") and also at the end of an action (e.g., “Robot Two here, I have moved to area Alpha"). Another time the robots initiate dialogue is when they arrive at a tube to repair it but the tube is still on; in this case they notify the human of this issue.

[5]The autonomous version of the *monitorWing* action is similar to the standard action, except that it does not hang indefinitely for a new command.

4.5 Conditions

We employed a between-subjects design in which participants were randomly assigned to one of two conditions - SMM or Non-SMM. The conditions were nearly identical except for a few key distinctions which are discussed below.

Robots in both conditions used the same autonomy policy described in Algorithm 1. The main difference between conditions was the architectural configuration. In the SMM condition, all of the components with “tabs" (see Fig. 2) were shared between the two robots. The Non-SMM condition used the same set of components except that the tabbed ones were duplicated (one instance for each robot). The most critical difference lies in the Belief component being duplicated in the Non-SMM condition. Since this component serves as the world knowledge base, having separate instances ensures that the robots do not have access to a source of shared knowledge. To offset this limitation, participants in both conditions were told that they could give the robots information about what the other robot is doing, or about states of the world (see Sec. 4.4.2 for details). This ensured that the Non-SMM robots could behave exactly as the SMM robots, but with the added overhead of the human providing extra information. Finally, participants in both conditions were read nearly identical task instructions, with people in the SMM condition being read an additional line informing them that the robots can share information and use that in the task.

4.6 Measures and Hypotheses

The *SMM hypothesis* is that *SMMs improve coordination and performance in human-agent teams.* We used a number of objective and subjective metrics to test this hypothesis. The main performance metric was *Score*, which was *task duration* (in seconds) multiplied by *task accuracy*. Task accuracy (in the rover task) was calculated as the total number of correctly-placed landmarks divided by the total announced landmarks plus any erroneous placements. We used a composite score measure because neither of these metrics alone is sufficient. It is possible to get a high task duration by ignoring the rover task entirely, and the converse is also true. We also used objective metrics of task efficiency, including the *percentage of tubes repaired* (total tubes repaired / total tubes damaged) and the *mean tube repair time*, or how long (in seconds) it took to repair tubes after they became damaged. Finally, we included subjective measures of *workload* (NASA-TLX) [18], *team workload* (TWLQ) [36], and *SA* (SART) [41]. These scales were administered in surveys at the conclusion of the study.

We had three main hypotheses, each relating to the broader *SMM hypothesis for artificial agents*:

H_1: Task performance, including score, task duration, and task accuracy would increase in the SMM condition (following similar effects found in the human literature, e.g., [24]). Moreover, we predicted that these effects would not be due to people in the SMM condition being better at talking to the robots. To test this, we looked at the *percentage of correctly-formed instructions*, which is the total number of instructions that were interpreted correctly by the robot divided by the total number produced. We expected no difference in this percentage across both conditions.

H_2: Task efficiency would improve in the SMM condition, as indicated by an increase in the percentage of tubes repaired, and a decrease in the mean tube repair time. Here we predict that teams in the SMM condition would be more vigilant about repairing tubes as soon as they become damaged due to the reduced overhead of managing the robots. This would lead to more tubes being repaired over the trial, and also a reduced tube repair time.

H_3: Workload, team workload, and SA would not change across conditions. Though the SMM hypothesis might predict an improvement in these measures, it is likely that any effort saved by the SMM would be re-applied to the primary task. In support of this, SMMs have been shown to increase task productivity under high workload [37] (without reducing workload), so we do not expect to see a decrease in workload or team workload in the SMM condition. In terms of SA, while robots in the SMM condition have more accurate task knowledge, this knowledge is mainly used in their autonomy policy to guide behavior. The human has limited access to this knowledge, except when receiving responses to queries. In general, we predict that the imposing time pressure will cause people in both conditions to have high demands on their attentional resources, ultimately limiting their understanding of the situation.

4.7 Results

We excluded from analysis a total of 10 participants who either failed to follow instructions or that experienced technical issues during their trial. Due to the involved nature of the task and the complex technical setup, this number was higher than anticipated. Failure to follow instructions was determined by a task duration within one minute of the minimum possible time of 326 s (so < 386 s) and/or scoring 0% on the rover task[6]. Eight participants' data were excluded based on this criterion, and two were excluded due to technical issues during the task. The remaining 26 participants (13 per condition) were used for the analysis. The mean age of the final sample was 24.9 (*SD* = 8.6) and 19 of the participants were male. We conducted between-subjects ANOVAs (N=26) to compare the effect of our independent variable, Condition (SMM vs Non-SMM), on our objective and subjective measures (see Sec. 4.6). For our analysis of the subjective survey measures, we averaged the various subscales and report a single composite score for each.

4.7.1 H_1: Task Performance. In terms of our objective task performance measures, we observed a significant increase for *task accuracy* [$F(1,24) = 5.796$, $p < .05$, $\eta_p^2 = .195$] and *task duration*

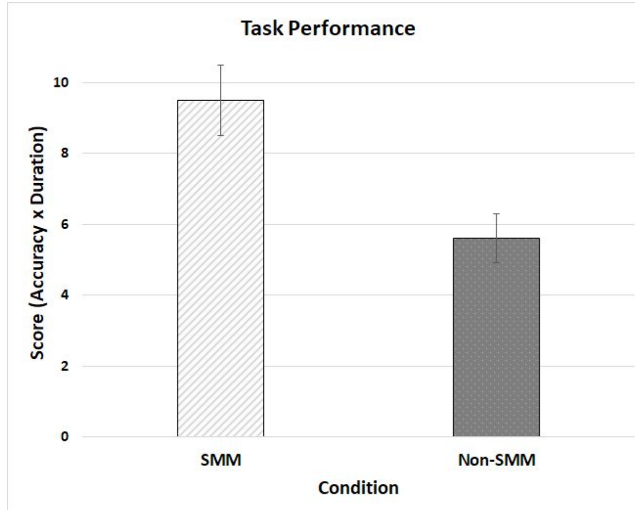

Figure 4: Results for task performance. Error bars represent standard error.

[$F(1,24) = 10.893$, $p < .005$, $\eta_p^2 = .312$] in the SMM condition. Not surprisingly, *score* (accuracy x duration) was also significantly increased [$F(1,24) = 10.432$, $p < .005$, $\eta_p^2 = .303$] (see Fig. 4). Based on these scores, only 4 of 13 people received the performance bonus in the Non-SMM condition, compared to 10 of 13 in the SMM condition. Finally, we found no difference in the *percentage of correctly-formed instructions* that participants gave to the robots between both conditions [$F(1,24)$=.001, $p > .05$, $\eta_p^2 < .001$][7]. Overall, these results support H_1 - teams in which the robots had SMMs displayed improved performance in the task over teams in which the robots did not have SMMs. Moreover, the SMM benefit was not due to people in that condition being better at talking to the robots. See Table 2 for an overview of the results.

4.7.2 H_2: Task Efficiency. There was a significant effect for the *percentage of tubes repaired* [$F(1,24) = 13.175$, $p < .005$, $\eta_p^2 = .354$], with the SMM group averaging .54 (*SD*=.12) and the Non-SMM group averaging .35 (*SD*=.15). There was no significant difference in *average repair time* [$F(1,24) = 2.267$, $p > .05$, $\eta_p^2 = .086$], but there was a numeric reduction of 22 s in the SMM condition compared to the Non-SMM condition. These results partially support H_2 in that there was a higher percentage of tubes repaired in the SMM condition, but the mean tube repair time was not different between conditions (see Table 2).

4.7.3 H_3: Workload and Situational Awareness. We did not find a statistically significant effect for any of our subjective survey measures, including *NASA-TLX* [$F(1,24) = .3419$, $p > .05$, $\eta_p^2 = .014$], *SART* [$F(1,24) = .1255$, $p > .05$, $\eta_p^2 = .005$], and *TWLQ* [$F(1,24) = 1.5645$, $p > .05$, $\eta_p^2 = .061$]. These results support H_3 in that there was no difference in workload, team workload, and SA between conditions (see Table 2).

[6] The minimum possible task duration is achieved by not turning off or repairing any tubes, and the minimum accuracy is achieved by not marking anything on the rover map (or marking everything incorrectly).

[7] While 50% seems low, we distinguish *correctly-formed* from *misinterpreted* and *unactionable* instructions, which denote parsing errors and failed preconditions, respectively. Unactionable instructions were interpreted correctly, but simply could not be executed.

Table 2: Table of results for task measures.

	Non-SMM		SMM			
Objective Measures	M	SD	M	SD	F-Value	P-Value
Task Performance (Score)	5.60	2.51	9.51	3.56	10.43	.004*
Task Duration (s)	516.20	92.58	689.16	164.72	10.89	.003*
Rover Task Accuracy	.63	.21	.81	.17	5.80	.024*
% Tubes Repaired	.35	.15	.54	.12	13.17	.001*
Mean Tube Repair Time (s)	159.64	35.98	138.06	37.09	2.27	.145
% Correctly-Formed Instructions	.50	.13	.50	.16	<.001	.976
Subjective Measures	M	SD	M	SD	F-Value	P-Value
Workload (NASA-TLX)	4.24	.50	4.12	.61	.34	.564
Team Workload (TWLQ)	4.24	.75	4.60	.72	1.56	.223
Situational Awareness (SART)	5.73	1.09	5.59	.98	.13	.726

5 DISCUSSION

The results of our user study support the SMM hypothesis in that teams in which the robots used SMMs outperformed teams without SMMs. While the SMM partially improved task efficiency, it had no impact on workload (at the individual or team level), or SA. Overall, these results support the findings of [13] in which SMM-like policies were found to improve performance and efficiency but not reduce workload. However, this is the first time that a comprehensive SMM framework has been implemented and evaluated in autonomous robots with natural language capabilities. Below, we interpret the findings and discuss future work.

5.1 Interpretation of Results

5.1.1 H_1 Supported: Robot SMMs Improve Task Performance. The finding that Robot SMMs improve task performance in human-robot teams serves as the first empirical support for the SMM hypothesis to date. We found improvements in all of our performance measures, including *task duration*, *rover task accuracy*, and the composite *score* measure (see Table 2). These improvements were largely due to robots in the SMM condition having more accurate knowledge of the task and team state, and accessing that knowledge through shared architectural components to coordinate more effectively. As a result, they required less monitoring, allowing the human to attend to the primary rover task.

Despite the fact that robots in both conditions used the same autonomy policy (described in Algorithm 1), the resulting behavior was different due to the robots having different knowledge. For example, when determining the next wing to move to, the robots use their belief about the location of the other robot, e.g., *Believes(self, at(R_2,Beta))*. However, in the SMM condition, this belief was more accurate since both robots knew one another's location at all times through sharing a knowledge base; this is represented as *CommonBelief(at(R_2,Beta))*. In the Non-SMM condition, knowledge of the other robot's location was usually outdated unless the human gave frequent verbal updates. As a result, robots in the Non-SMM condition were more likely to end up in the same wing, which could lead to inefficiency since no one is monitoring or repairing tubes in other wings. Note that the SMM robots could still be in the same wing if the human instructed them to do so, however, their autonomous actions were still more efficient since they would immediately know when tubes were repaired by the other robot.

Another factor explaining the performance difference between conditions was greater alignment of the Human-Robot SMM, which was partly mediated by improved accuracy of responses to the humans' queries. Since a specific robot is addressed for each query, it will respond with the knowledge that it has. However, in the SMM condition, robots have access to their own knowledge as well as that of the other robot. So if the human asks "Robot Two which tubes are damaged?", Robot Two will respond with all the tubes it knows to be damaged as well as all the tubes that Robot One knows to be damaged. Providing more accurate information in this way may have helped the humans to plan which tubes to prioritize for repair. Overall, H_1 is supported, suggesting that Robot SMMs serve to improve task performance.

5.1.2 H_2 Partially Supported: Robot SMMs Improve Some Measures of Task Efficiency. As discussed in the previous section, increased inefficiency in the Non-SMM condition was an important factor in the observed performance difference. Another source of inefficiency occurred when a robot moved to a wing that was just handled by the other robot. This is wasteful because robots take about 30 s to move from one wing to another, so arriving at a wing with no damaged tubes will prompt the robot to move again, thus wasting more time. Note that this could still happen in the SMM condition, but it was usually less problematic because the robots spread out more effectively. This problem could be alleviated if the human micromanaged the robots at every step or gave them information about the other robot, but the complexity of the task tended to prevent micromanagement, and people in both conditions generally did not provide information to the robots. As a result, teams in the Non-SMM condition repaired a fewer proportion of damaged tubes (35%) compared to the SMM condition (54%). Despite this, *mean tube repair time* was not significantly different between conditions, although we observed a small qualitative difference. This suggests that teams in both conditions repaired tubes at about the same rate, but that the Robot SMM enabled more tubes to be repaired, likely due to more efficient actions. As a result, H_2 is partially supported, though further analysis and/or studies are needed to track the exact nature of the efficiency gain.

5.1.3 H_3 Supported: Robot SMMs Do Not Improve Workload or Situational Awareness in the Tube Repair Task.

We found no improvements in workload, team workload, or SA in the SMM condition, supporting H_3 (see Table 2). In the SMM condition, people repaired more tubes, indicating that they were *more* productive and did not benefit from increased downtime (or workload reduction). Regarding SA, since robot behavior was the same in both conditions, the only way SA could be reduced was if people frequently asked the robots about task status, which they generally did not. Another factor was that the surveys were taken at the end of the task (when workload was highest), so this may have also contributed to increased perceptions of workload and reduced SA in both conditions. In future studies, perhaps it may be useful to supplement the existing surveys with a freeze probe method like SAGAT [6] for measuring SA or a physiological measure of workload (e.g., [19]).

Importantly, these results do not suggest that Robot SMMs *cannot* be used to modulate workload or enhance SA, as evidence from human teams shows that these concepts are highly related [7, 29, 32]. For example, [2] found that SMMs are one factor among others (shared displays, communication, etc.) that influences SA. Other evidence suggests that SMMs may not explicitly *reduce* workload, but may enable a team to maintain its performance under stress and time pressure [12, 28, 37]. Our results support this position, as teams in the SMM condition were able to maintain a high level of performance throughout the increasing workload of the task.

5.2 Contributions

This study offers several contributions to the fields of HRI and multi-agent systems. First, it provides the first empirical support for the SMM hypothesis for artificial agents. Since the only difference between our experimental conditions was the shared architectural components, the results suggest that Robot SMMs have a practical benefit for teaming - a result that has not previously been reported. Importantly, since the robots in the task did not communicate with each other, this result is not simply a replication of prior findings that communication improves performance [3]. Similarly, this result does not demonstrate that robot "telepathy" [44] improves performance, since telepathy is a form of covert communication that requires a sender and a receiver, which is unlike how the robots accessed the SMM in our study. Instead, the system described here is most closely related to blackboard architectures or centralized planning, with key differences being that it does not involve iterative problem solving and we do not use a planner. To our knowledge, there has not been any prior empirical evidence that such systems serve to improve performance in a human-robot collaborative task.

Another contribution is the novel task domain, which serves as a test bed for studies on human-robot teaming. The task involves interdependency of action, ramping workload, and is scalable with respect to the number and kinds of agents on the team. In general, the domain is extensible to cover a wide range of interaction requirements expected for HRI in space domains [9, 10]. Additional robot and/or human agents can be added either in the virtual environment, or remotely (simulating ground operators). While the present study focused on co-located teaming (with periods of remote interaction), such additions support the study of truly distributed interaction over various spatial ranges and time scales.

A set of design requirements for evaluating SMMs in human-robot teams has also come out of this work, along with a system that implements the requirements. These requirements are specifically aimed at teams in which the robots are expected to serve as human-like partners, and include the need for: 1) an SMM that is formally defined within the context of a robotic architecture and implemented in real or virtual robots (not tele-operated or simulated agents), 2) a team structure that includes a combination of real humans and robots in which the robots possess robust NLU and autonomous capabilities that leverage the SMM to adapt behavior, 3) a collaborative task requiring interdependency of action, and 4) an evaluation that involves the human-robot team performing the specified collaborative task (not a simulation or proof-of-concept), and which measures various objective and subjective aspects of team functioning. These are challenging requirements, but we feel that they provide for a true test of the robot-as-partner paradigm.

5.3 Future Work

Several directions for future work are currently being explored. First, we plan to use our evaluation platform to run additional conditions in the current domain, further exploring the relationship between an SMM and team coordination. These studies will also help to understand the relationship between task structure (e.g., communication constraints, time pressure, etc.) and SMM. Another important direction for future work is to include more aspects of the humans' mental states into the SMM. We are currently exploring epistemic planning to generate robot behavior based on models of all agents' knowledge states. With this approach, it becomes possible to track what each agent on the team believes, and to take actions to align these beliefs, thus supporting Human-Robot SMMs. Finally, we seek to extend the system's natural language capabilities in order to better handle phenomena such as disfluency and speech overlap, which are common in team discourse [11, 15, 16].

6 CONCLUSION

We have demonstrated for the first time that a comprehensive computational framework for SMMs in human-robot teams serves to improve performance in a collaborative task. This is the first demonstration of the *SMM hypothesis for artificial agents*, and shows that shared knowledge representations in robots can support coordination and improve team performance in complex domains. In testing this hypothesis we have developed a novel, scalable evaluation platform for studying human-robot teaming that allows for the modification of team organization and task parameters to further explore the SMM and other aspects of teaming. We hope that our design guidelines, platform, and results will spur further research toward the goal of making robots genuine teammates.

ACKNOWLEDGMENTS

This work was partly funded by a NASA Space Technology Research Fellowship under award 80NSSC17K0184, NASA grant C17-2D00-TU, and AFOSR grant FA9550-18-1-0465. We are especially grateful to Evan Krause, Bradley Oosterveld, and Zachary Haga for their technical assistance.

REFERENCES

[1] Jenay M Beer, Arthur D Fisk, and Wendy A Rogers. 2014. Toward a framework for levels of robot autonomy in human-robot interaction. *Journal of human-robot interaction* 3, 2 (2014), 74–99.

[2] Cheryl A Bolstad and Mica R Endsley. 1999. Shared mental models and shared displays: An empirical evaluation of team performance. In *proceedings of the human factors and ergonomics society annual meeting*, Vol. 43. 213–217.

[3] Abhizna Butchibabu, Christopher Sparano-Huiban, Liz Sonenberg, and Julie Shah. 2016. Implicit coordination strategies for effective team communication. *Human factors* 58, 4 (2016), 595–610.

[4] Sharolyn Converse, JA Cannon-Bowers, and E Salas. 1993. Shared mental models in expert team decision making. *Individual and group decision making: Current issues* 221 (1993).

[5] Myron A Diftler, JS Mehling, Muhammad E Abdallah, Nicolaus A Radford, Lyndon B Bridgwater, Adam M Sanders, Roger Scott Askew, D Marty Linn, John D Yamokoski, FA Permenter, et al. 2011. Robonaut 2-the first humanoid robot in space. In *Proceedings of the 2011 IEEE international conference on robotics and automation.* IEEE, 2178–2183.

[6] Mica R Endsley. 1988. Situation awareness global assessment technique (SAGAT). In *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference.* IEEE, 789–795.

[7] Mica R Endsley. 2000. Situation models: An avenue to the modeling of mental models. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 44. SAGE Publications Sage CA: Los Angeles, CA, 61–64.

[8] J Alberto Espinosa, Robert E Kraut, Sandra A Slaughter, Javier F Lerch, James D Herbsleb, and Audris Mockus. 2001. Shared Mental Models, Familiarity and Coordination: A Multi-Method Study of Distributed Software Teams. In *Proceedings of the International Conference on Information Systems.*

[9] Terrence Fong and Illah Nourbakhsh. 2005. Interaction challenges in human-robot space exploration. *Invited article for ACM Interactions special issue on human-robot interaction* (2005).

[10] Terrence Fong, Jennifer Rochlis Zumbado, Nancy Currie, Andrew Mishkin, and David L Akin. 2013. Space telerobotics: unique challenges to human–robot collaboration in space. *Reviews of Human Factors and Ergonomics* 9, 1 (2013), 6–56.

[11] Felix Gervits, Kathleen Eberhard, and Matthias Scheutz. 2016. Disfluent but effective? A quantitative study of disfluencies and conversational moves in team discourse. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers.* Osaka, Japan, 3359–3369.

[12] Felix Gervits, Kathleen Eberhard, and Matthias Scheutz. 2016. Team communication as a collaborative process. *Frontiers in Robotics and AI* 3 (2016), 62.

[13] Felix Gervits, Terry Fong, and Matthias Scheutz. 2018. Shared Mental Models to Support Distributed Human-Robot Teaming in Space. In *2018 AIAA SPACE and Astronautics Forum and Exposition.* 5340.

[14] Felix Gervits, Anton Leuski, Claire Bonial, Carla Gordon, and David Traum. 2019. A Classification-Based Approach to Automating Human-Robot Dialogue. (2019).

[15] Felix Gervits and Matthias Scheutz. 2018. Pardon the interruption: Managing turn-taking through overlap resolution in embodied artificial agents. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue.* 99–109.

[16] Felix Gervits and Matthias Scheutz. 2018. Towards A Conversation-Analytic Taxonomy of Speech Overlap. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

[17] Felix Gervits, Charlotte Warne, Harrison Downs, Kathleen Eberhard, and Matthias Scheutz. 2017. Exploring Coordination in Human-Robot Teams in Space. In *AIAA SPACE and Astronautics Forum and Exposition.* 5309.

[18] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting.* 904–908.

[19] Jamison Heard, Rachel Heald, Caroline E. Harriott, and Julie A. Adams. 2019. A Diagnostic Human Workload Assessment Algorithm for Collaborative and Supervisory Human–Robot Teams. *ACM Trans. Hum.-Robot Interact.* 8, 2, Article 7 (June 2019), 30 pages. https://doi.org/10.1145/3314387

[20] Catholijn M Jonker, M Birna Van Riemsdijk, and Bas Vermeulen. 2011. Shared mental models. In *Coordination, organizations, institutions, and norms in agent systems VI.* Springer, 132–151.

[21] M Lee, Tristan Johnson, and Myung H Jin. 2012. Toward understanding the dynamic relationship between team and task shared mental models as determinants of team and individual performances. *International journal of information technology and business management* 8, 1 (2012), 1–14.

[22] Anton Leuski and David Traum. 2011. NPCEditor: Creating virtual human dialogue using information retrieval techniques. *Ai Magazine* 32, 2 (2011), 42–56.

[23] Stephanie M Lukin, Felix Gervits, Cory Hayes, Pooja Moolchandani, Anton Leuski, John Rogers, Carlos Sanchez Amaro, Matthew Marge, Clare Voss, and David Traum. 2018. ScoutBot: A Dialogue System for Collaborative Navigation. In *Proceedings of ACL 2018, System Demonstrations.* 93–98.

[24] J.E. Mathieu, T.S. Heffner, G.F. Goodwin, E. Salas, and J.A. Cannon-Bowers. 2000. The influence of shared mental models on team process and performance. *The Journal of applied psychology* 85, 2 (2000), 273–83. https://doi.org/10.1037/0021-9010.85.2.273

[25] Susan Mohammed, Lori Ferzandi, and Katherine Hamilton. 2010. Metaphor no more: A 15-year review of the team mental model construct. *Journal of management* 36, 4 (2010), 876–910.

[26] Robin R Murphy. 2004. Human-robot interaction in rescue robotics. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 34, 2 (2004), 138–153.

[27] Stefanos Nikolaidis and Julie Shah. 2012. Human-robot teaming using shared mental models. *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (2012).

[28] Judith Orasanu. 1990. *Shared mental models and crew decision making.* Technical Report Technical Report 46.

[29] Scott Ososky, David Schuster, Florian Jentsch, Stephen Fiore, Randall Shumaker, Christian Lebiere, Unmesh Kurup, Jean Oh, and Anthony Stentz. 2012. The importance of shared mental models and shared situation awareness for transforming robots from tools to teammates. In *Unmanned Systems Technology XIV*, Vol. 8387. International Society for Optics and Photonics, 838710.

[30] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding.* IEEE Signal Processing Society.

[31] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. 2009. ROS: an open-source Robot Operating System. In *ICRA workshop on open source software*, Vol. 3. Kobe, Japan, 5.

[32] Eduardo Salas, Carolyn Prince, David P Baker, and Lisa Shrestha. 1995. Situation awareness in team performance: Implications for measurement and training. *Human factors* 37, 1 (1995), 123–136.

[33] Matthias Scheutz. 2006. ADE: Steps toward a distributed development and runtime environment for complex robotic agent architectures. *Applied Artificial Intelligence* 20, 2-4 (2006), 275–304.

[34] Matthias Scheutz, Scott A DeLoach, and Julie A Adams. 2017. A framework for developing and using shared mental models in human-agent teams. *Journal of Cognitive Engineering and Decision Making* 11, 3 (2017), 203–224.

[35] Matthias Scheutz, Thomas Williams, Evan Krause, Bradley Oosterveld, Vasanth Sarathy, and Tyler Frasca. 2019. An overview of the distributed integrated cognition affect and reflection diarc architecture. In *Cognitive Architectures.* Springer, 165–193.

[36] James Michael Sellers. 2013. *Team Workload Questionnaire (TWLQ): Development and Assessment of a Subjective Measure of Team Workload.* Ph.D. Dissertation. University of Canterbury. Psychology.

[37] Daniel Serfaty, Elliot E Entin, and Catherine Volpe. 1993. Adaptation to stress in team decision-making and coordination. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 37. SAGE Publications Sage CA: Los Angeles, CA, 1228–1232.

[38] Renée J Stout, Janis A Cannon-Bowers, Eduardo Salas, and Dana M Milanovich. 1999. Planning, shared mental models, and coordinated performance: An empirical link is established. *Human Factors* 41, 1 (1999), 61–71.

[39] K. Sycara and G. Sukthankar. 2006. *Literature Review of Teamwork Models.* Technical Report CMU-RI-TR-06-50. Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.

[40] Kartik Talamadupula, Gordon Briggs, Tathagata Chakraborti, Matthias Scheutz, and Subbarao Kambhampati. 2014. Coordination in human-robot teams using mental modeling and plan recognition. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems.* IEEE, 2957–2962.

[41] RM Taylor. 1990. Situational Awareness Rating Technique(SART): The development of a tool for aircrew systems design. *In Proceedings of Situational Awareness in Aerospace Operations (AGARD)* (1990).

[42] Simon Thiel, Dagmar Häbe, and Micha Block. 2009. Co-operative robot teams in a hospital environment. In *2009 IEEE International Conference on Intelligent Computing and Intelligent Systems*, Vol. 2. IEEE, 843–847.

[43] Willie Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, and Joe Wölfel. 2004. Sphinx-4: A flexible open source framework for speech recognition. *Sun Microsystems* (12 2004).

[44] Tom Williams, Priscilla Briggs, Nathaniel Pelz, and Matthias Scheutz. 2014. Is robot telepathy acceptable? Investigating effects of nonverbal robot-robot communication on human-robot interaction. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication.* IEEE, 886–891.

[45] John Yen, Xiaocong Fan, Shuang Sun, Timothy Hanratty, and John Dumer. 2006. Agents with shared mental models for enhancing team decision makings. *Decision Support Systems* 41, 3 (2006), 634–653.