Towards Genuine Robot Teammates: Improving Human-Robot Team Performance Beyond Shared Mental Models with Proactivity

Gwendolyn Edgar and
Ayca Aygun and
Matthew McWilliams and
Matthias Scheutz[0000−0002−0064−2789]

Abstract Recent work in human-robot teaming has demonstrated that when robots build and maintain "shared mental models", the effectiveness of the whole human-robot team is overall better compared to a baseline with no shared mental models. In this work, we expand on this insight by introducing proactive behaviors in addition to shared mental models to investigate potential further improvements of team performance and task efficiency. We developed a set of proactive robot behaviors that we experimentally compared to baseline "reactive" behaviors, hypothesizing that, combined with shared mental models, robots with these more proactive behaviors will become even more effective teammates. The results from a human subject evaluation showed that proactive robot behaviors improves task efficiency and performance over mere reactive behaviors and objectively lowered human workload as measured by percentage change in the subject's pupil size.

1 Introduction

Human-robot teams have the potential to improve human lives in numerous areas like search-and-rescue missions, manufacturing, and education among others. Yet, they are still limited by the robots' lack of team awareness—current robots in mixed-

Gwendolyn Edgar
Tufts University, 177 College Avenue, Medford, MA 02155, e-mail: gwenfedgar@gmail.com

Ayca Aygun
Tufts University, 177 College Avenue, Medford, MA 02155 e-mail: ayca.aygun@tufts.edu

Matthew McWilliams
Tufts University, 177 College Avenue, Medford, MA 02155 e-mail: matthew.mcwilliams@tufts.edu

Matthias Scheutz
Tufts University, 177 College Avenue, Medford, MA 02155 e-mail: matthias.scheutz@tufts.edu

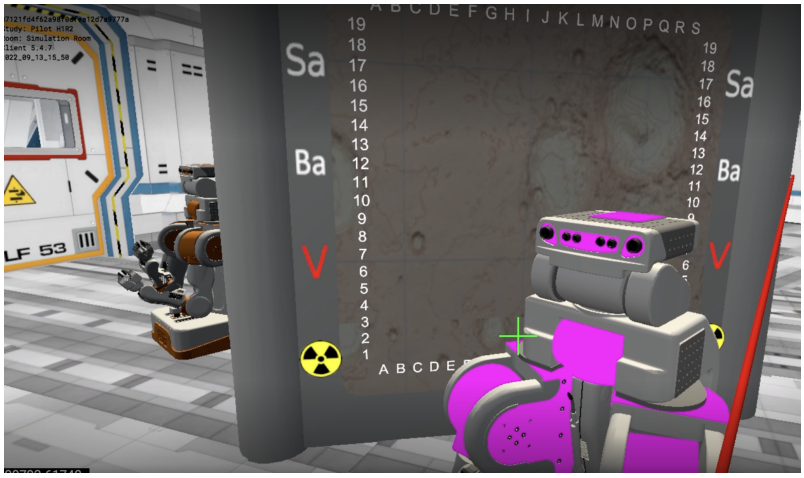

Fig. 1 The two simulated Willow Garage PR2 robots in the central area of the spaceship in the Unity3D simulation (different coloration and voice are used to distinguish them), see text for details.

initiative teams have at best a rudimentary understanding of the moment-by-moment team goals and the state of the task execution.

One way to improve their awareness of team and task states is endow robots with *Shared Mental Models* (SMMs) which are known to be essential for high-performing human teams. In the human teaming case, the term "shared mental model" refers to the synchronized individual mental models of all team members which track goals and subgoals, team and task states, team member capabilities and activities, amond others (e.g., [35]). Recent evidence from evaluations with robots using shared mental models indicate that in a team consisting of a human and multiple robots, SMMs can improve collaboration due to the shared understanding of the task environment, as measured in terms of enhanced task performance and team efficiency (e.g., Gervits et al. [19]).

However, shared mental models alone do not always improve the robots' ability to support human activities. Purely *reactive* robots, for example, might have a good understanding of the overall team and task state, but require explicit human instructions to take any action, while *proactive* robots could use that information to perform actions to further task performance (e.g., to proactively move to an area where the robot will be needed next, or to remind a human teammate of a task that still needs to be completed)—there is initial evidence from research investigating collaborations between human and robots that proactive robot behavior can improve team performance (e.g., [2]). Moreover, proactive robots might be able to lower human cognitive load compared to reactive robots, e.g., if people know that the robot will remind them of tasks they need to perform and when instead of having to remember those tasks themselves and track the time. This would have beneficial effects on team performance by allowing otherwise overloaded humans to perform at more effective load levels.

Hence, the goal of this chapter is to empirically examine the extent to which proactive robot behavior can improve the performance of mixed-initiative human-robot teams above and lower human cognitive load, beyond what shared mental models are able to do when they reduce redundancy and increase information distribution in robots. Specifically, we will introduce algorithms for three different autonomous teaming behaviors—*reactive*, *active* and *proactive*—that operate in conjunction with SMMs on two robots within a scalable teaming setting. We will then descibe the integration of the algorithms within a cognitive robotic architecture as well as the two autonomous robots it controls in a simulated Unity3D spaceship setting (see Fig. 1). We then report results from human subject evaluations that confirm our main hypotheses about the effectiveness of the proposed algorithms: that *proactive robots with SMMs* lead to better team performance and lower cognitive workload (as measured objectively by percentage change in pupil size) than *reactive robots with SMMs* on several team performance metrics.

2 Related Work

Teaming Behavior. One of the most important elements of a team is communication. The two main team communication methods [17] are reactive and proactive. Generally, a reactive teammate will only communicate when asked; a proactive teammate readily offers information and may also ask questions. A number of studies [15, 29, 41] have shown that effective teams exhibit proactive behavior. In human-robot teams, there is increasing evidence showing that humans prefer to work with proactive robots. Fong et al. [18] show how human-robot collaboration and dialogue improve multi-robot remote driving. Baraglia et al. [2] found that people collaborate best with a proactive robot on a joint preparation task. Cakmak et al. [3] show that in the context of table-top manipulation, proactive behavior yields the best collaboration results. On the other hand, Bhattacharjee et al. [6] show that in the context of robot-assisted feeding "more autonomy is not always better". Cakmak et al. [11] show that people overall prefer the robot taking initiative, but also do not like feeling bombarded with questions. There is a wide range of proactive behaviors, and it is not always clear what level of proactivity is best. Effective communication between humans and robots often depends on the quality of questions asked. Cakmak et al.[12] explore how robots can ask good questions. Rosenthal et al. [33] measure how robot questions affect the accuracy of human responses. In addition to an agent effectively communicating, a good proactive teammate can independently make choices about the best way to achieve a goal. Zhang et al. [47] show that in a simulated urban search and rescue task, humans prefer an agent teammate which adapts to the human's goals without them being explicitly communicated. Grosinger et al. [21] build a framework to allow agents to better choose their own goals and how to achieve them. We build upon this research on how to make the best robotic teammates, and design three different agent behaviors, each depicting different levels of proactivity when communicating with the human.

Shared Mental Models. A shared mental model [14] is a distributed knowledge structure that represents important aspects of the team and task states [36] and enables seamless communication by containing information for successful collaboration, like monitoring tasks, teammates and beliefs. Cohen et al. [13] define a team as "a set of agents having a shared objective and a shared mental state". SMMs have been shown to significantly improve human-human collaboration ([16, 28, 42, 31, 14]). Inspired by the improvement in collaboration between humans, researchers have also shown an improvement in collaboration between humans and agents, like software agents [26, 46], and robots [20, 30, 32]. Specifically, Gervits et al. [19] show that shared mental models can improve collaboration in distributed human-robot interaction improving task performance and team efficiency. In this work, the agents are only reactive. This line of work focuses on the robots having a shared mental model for more effective dissemination of information. However, the robots did not utilize their information to proactively engage humans in task-based or time sensitive interactions. In creating our human-robot teams, we give our agents shared mental models. Since Gervits et al. [19] showed that shared mental models do indeed improve human-robot teams, we build upon their work and explore if proactivity on top of shared mental models also improves human-robot teams.

Cognitive Workload. Robust estimation of cognitive workload has been considered an essential part of developing collaborative human-robot teaming since human performance can be significantly impacted by cognitive workload (e.g., [24]). Numerous objective measurements have been proposed to predict cognitive workload during task performance based on different types of physiological signals including pupillometry, electroencephalography, arterial blood pressure, heart rate and blood pressure variability, respiration rate, or skin conductance (e.g., [5, 8, 23, 43, 27, 10]). Among those, pupillometry is the most suitable physiological signal given that eye gaze is simple to record, easy to process in comparison to other physiological signal types, and, most importantly, because it has been shown to be the best indicator of cognitive workload compared to all other signal types [1]. We follow [1] and use *percentage change in pupil diameter* (PCPS) for cognitive workload assessment as it provides more reliable outcomes compared to the raw pupil diameter and eliminates subject-based variations [1, 48].

3 Technical Approach

We extend the cognitive robotic architecture DIARC [37] by implementing event memory, three algorithms for different robot teaming behaviors including proactive behaviors, methods for initiation of dialogue, initiation of shared goals, further goal management behavior, autonomous reactivation, and an extension of the shared mental models from [19]. The robots use a "supervisory control" policy [4] in which their actions and behaviors are carried out independently, but the human can intervene at any point.

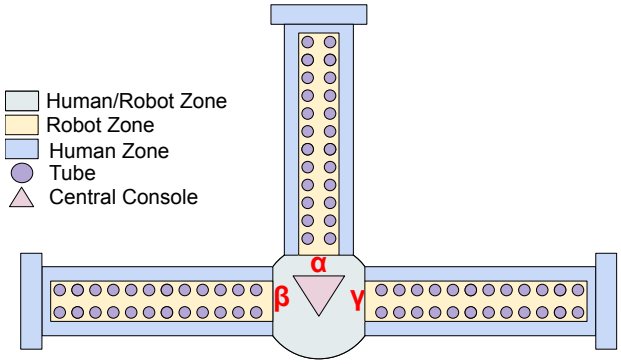

Fig. 2 Schematic of the spaceship.

3.1 Team and task setting

We utilize a spaceship environment similar to Gervits et al. [19] in the Unity3D virtual reality (VR) environment with **one** human and **two** robot teammates. The simulation was developed to simulate a use case for human-robot teams in space via a collaborative intravehicular activity (IVA) with a mission and a maintenance task. The spaceship has three wings and a central area that contains the mapping console (see Fig. 2). There are a number of ways for a robot to achieve effective communication (e.g., spoken language, gaze, gestures). Given evidence of effective communication through spoken language [7, 40, 45] in human robot teams, we chose to use natural language as a communication mechanism.

A human subject is told that their "primary task" is to record geological information given verbally by an off-ship rover on a map in the ship's central area. This is really a **distractor** task meant to increase the cognitive load of the participant. The rover reports data at preset times that are different across the trials. The human does not know the timing but must be at the central mapping console to mark the materials and landmarks given by the rover.

The subjects' **main** task is to keep the spaceship operational. In each of the three wings, there is a central corridor with 24 power tubes that, over time, will begin to break down. Once a tube starts to malfunction, it needs to be repaired within 90 seconds or otherwise it will be permanently broken and not produce power. Human subjects can decide if and when to repair a malfunctioning tube by turning it off (which makes it decay more slowly) and tell robots to repair it (Fig. 6). Once repaired, human subjects must turn them back on to produce power. Critically, the "tube repair subtask" was specifically designed to be a joint activity that requires both human and robot to take actions, and neither human nor robot alone can repair broken tubes.

All robots share information with each robots through the SMM, patrolling the three wings for breaking tubes and acting on the information depending on their particular **teaming behavior**: *reactive*, *active*, and *proactive* which *differ only with*

respect to their communication with the human. Based on [21, 47] showing that the best teammates try to independently identify the team's goal and take steps to achieve it, we designed all three robot configurations to be aware of the overall goal ("keep the spaceship operational") and independently choose which area to patrol next, using information from their shared mental models. Two of the teaming behaviors (the active and proactive ones) share unsolicited information with the human.

1. **Reactive:** Reactive robots silently patrol, only giving information when prompted.
2. **Active:** Active robots readily offer information to the human teammates, being mindful of becoming a distraction.
3. **Proactive:** Proactive robots offer information to the human teammates and attempt to initiate shared tasks when the situation calls for it, reminding the human to perform necessary actions that humans have already been committed to, but not finished.

Reactive robots are based on the agent that uses SMMs from Gervits et al. [19] which autonomously patrols the spaceship, but does not communicate with the human unless prompted. The agents use their SMM to share their task states (i.e., the knowledge about operational and broken tubes as well as their internal goals) which allows them to patrol the spaceship more effectively [19], yet leaving it up to the human teammate to query them for information and prompting them to initiate tasks.

In addition to spaceship and robot state, robots use their SMM to keep track of what was communicated by whom and when. Using this information, active or proactive robots can decide when an event (or a reminder) should be communicated and do not needlessly pester the human teammate with information that is not new or newly relevant. This is an important design choice as literature suggests that human teammates dislike too much communication (e.g., [11]).

We thus designed the **active** behavior based on team compositions in a workplace environment [15, 29, 41]. An active teammate will not only perform tasks when told, but readily communicates with their team on collaborative goals while being mindful of their teammates workload [18]. To this end, the active robot does not just silently patrol the spaceship, but also provides its human teammate time-relevant information about their shared task.

The **proactive** behavior builds upon the active behavior. A proactive teammate goes above and beyond what they are asked to do, often bringing innovation and leadership to a team to get better results [3]. Therefore, in addition to readily offering relevant information (with care not to overwhelm the human), proactive robots also initiates shared tasks. We specifically designed the question to be a closed form "yes/no" question for ease of response, as also Cakmak et al. [12] showed that "yes/no" is a positive property of a question.

For example, in the active behavior, $Robot1$ patrols area α and finds a new breaking tube. It will immediately convey the information. If, 15 seconds later $Robot2$ patrols area α and observes the breaking tube, which will have degraded a bit, it will also see that the human partner was told about this tube 15 seconds ago and did not take action. Here, the robot assumes the human knows the tube is broken

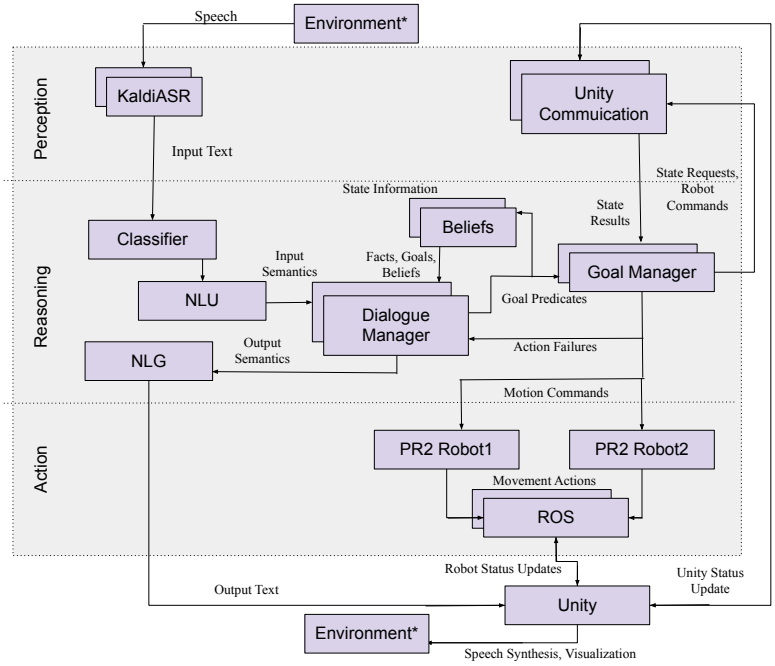

Fig. 3 DIARC Architecture as described in Section 3. Double tabbed sections are shared between robots as the SMM.

and adding more information would just be a distraction. However, if *Robot*2 went in 35 seconds later or the tube was close to fully breaking and becoming unusable, it would tell the human teammate again. In the proactive behavior, *Robot*1 and *Robot*2 would also ask the human if the human wanted them to repair the tube if the command was not already given. If the human responded "yes", the robot would initiate the action, otherwise it would resume its patrol behavior.

Evaluation Architecture. We use the DIARC architecture [37] for shared mental models (SMMs), natural language understanding (NLU), natural language generation (NLG), task-based inferences, goal planning, and action selection. DIARC provides a modular architecture for implementing multi-robot systems with shared architectural components (see Fig. 3). We use the Robot Operating System (ROS) for path planning and navigation, and algorithms in the Unity3D game engine for visualization and object avoidance. We use virtual robot models of the PR2 robot by Willow Garage, such that our algorithms and architectures can be equally run on physical robots. We enhanced the NLG system to initiate a dialogue, implemented further goal managing behavior and autonomy activation, and added configuration and implementation functions.

Shared Mental Models. All robots independent of proactivity levels use an SMM to be able to track the most up-to-date information possible, based on [19] who showed that teams with SMMs perform better than teams without on quanti-

tative metrics. SMMs ensure that robots do operate on or communicate outdated information, e.g., allowing robots to more efficiently distribute themselves in the spaceship—since both robots are aware of the other's current and past positions, they can better determine where they are most needed next. In the active and proactive cases, SMMs ensure that the human is not overwhelmed with information, as the robots know what information has already been communicated to the human. And finally, in the proactive case, SMMs form the basis that allows robots to coordinate their communication with the human, e.g., that at no point two robots will attempt to initiate the same task, thus reducing unnecessary communication and potentially reducing human workload and distraction. SMMs can thus make agents better teammates across two dimensions: (i) they can take more effective physical actions and (ii) they can become more effective communicators. In our implementation, when a robot queries its beliefs about the state of the world it also takes into account the other robot's beliefs (Fig. 3 shows that both robots share the same belief component and the goal manager). The robots use their SMM to add any new information, including any observations they make on the state of the spaceship (e.g., "Tube Alpha Left Two is damaged at 42%"), any information communicated to the user (e.g., information about the most damaged tube in area Alpha being at 42%), any questions asked (e.g., if the human like to repair tube Alpha Left Two), any answers the human communicates (e.g., no), any commands the human gives (e.g., "repair tube Alpha Left Two"), and any responses the agents give. Each piece of information is stored with a timestamp, who created it, and (in case of a communication) who was talking to whom.

3.2 Robot autonomy and proactive behavior

The fundamental role of the robots in the spaceship is to patrol the wings and fix broken tubes for which they using $AutonomyAction(R, pl)$ (Algorithm 1) which is a continuous process, started once at the beginning of each trial, with the name of the robot R and its *proactiveness level* pl as inputs. First, a robot decides which area is in most need to be patrolled next. After patrolling the next area, the robot will store any discovered information in the SMM. The reactive robot does not communicate with the human and thus will choose the next area to patrol while the active and proactive robots will share the outcome of patrolling with the human, if the human has not been recently informed and store the communication. Whereas active robots will go to patrol the next area, proactive robots will ask their human teammate if the most damaged tube in the patrolled area should be repaired (if there is any). After the question resolves (Algorithm 3), proactive robots will move on to patrol the next area. We note that every time robots communicate with the human, the communication (and the possible human response) is stored, and so are communications initiated by the human.

$AutonomyAction(R, pl)$ uses several subroutines. Algorithm $WhereNext(R)$ takes as input the robot's name R and returns the wing to be patrolled next by

Algorithm 1 *AutonomyAction*(R, pl)

1: $nextArea = WhereNext(R)$
2: wait for $InTransit(nextArea, R)$ to complete
3: $(damagedTubes, mdt, p) = PatrolArea(nextArea, R)$
4: **if** pl is $active$ or $proactive$ **then**
5: **if** human has not been informed recently **then**
6: inform human of the number of damaged tubes and which tube is most broken $(len(damagedTubes), mdt, p)$
7: **if** pl is proactive **then**
8: ask whether to repair the most damaged tube (mdt)
9: **if** received YES answer **then**
10: $repairTube(mdt, pl, R)$
11: wait 2 seconds, then go to line 1.

R. The robots use the SMMs to pick the area that has not been patrolled the longest. Algorithm $PatrolArea(nextLocation, R)$ (Algorithm 2) takes as input an area, updates the state of the tubes in it and aggregates information that is used in $AutonomyAction(R, pl)$ to inform the human and/or initiate a repair action. Algorithm $InTransit(location, R)$ takes robot R to the desired location, either an area or a tube.

Algorithm 2 *PatrolArea*$(area, R)$

1: R patrols $area$ and identifies $damagedTubes$
2: R commits state of area to SMM
3: let mdt be the most damaged tube at percentage p
4: return $(damagedTubes, mdt, p)$

RepairTube$(tube, pl, R)$ (Algorithm 3) is triggered when the robot is ordered to repair a tube by the human or initiates the action itself. It will first go to the tube in question to check its current state. If the tube is damaged, the robot will check that the tube is off as tubes are normally on and only the human can turn them off. If the tube was not turned off after ten seconds (which is an appropriate amount of time for the human to travel to the tube from other parts of the spaceship), the reactive and active robots will run *AutonomyAction*(A, pl) while the proactive robot will check whether the human was planning to turn the tube off and may wait again. In all cases, robots will repair damaged tubes that are turned off and resume *AutonomyAction*(A, pl), leaving when a tube becomes permanently broken.

4 Human Subject Evaluation

In order to evaluate the effects of the different robot behaviors on mixed-initiative human-robot teams, we ran a within-subjects evaluation study where each human participant was paired with two robots in the simulated environment in three differ-

Algorithm 3 *RepairTube*$(tube, pl, R)$

```
1: R waits for InTransit(tube, R) to complete
2: if tube is not damaged or tube is fully broken then
3:     run Algorithm 1 // Repair tube fails
4: waitTime = 0
5: while tube not off and waitTime < 10 seconds) do
6:     query which tubes are off
7:     waitTime+ = elapsedTime
8: if tube is not off then
9:     if pl is reactive or active then
10:        run Algorithm 1 // Repair tube fails
11:    else
12:        // pl is proactive
13:        R asks if the human will turn the tube off
14:        if received YES answer then
15:            go to to line 2
16:        else
17:            run Algorithm 1 // Repair tube fails
18: repair tube, run Algorithm 1
```

ent randomly sequenced trials working with either two reactive, active, or proactive robots (both robots performed the main task, patrolling the spaceship and noting the broken tubes, but the way they communicated with the human changed based on the condition). Participants were not told about the different conditions, only that the robot behaviors *may* change between trials. Participants had two tasks: to collaborate with the robots on board and to record all coordinates of rock formations communicated by the remote rover on the map in the central area. As already mentioned, the goal of this latter "distractor task" was to add cognitive load throughout the experiment. The human participant wore an untethered Vive headset and interacted with their environment using the Vive controllers. An example video showing the different conditions can be found here: https://streamable.com/84x7bi.

4.1 Procedure

We recruited 28 participants from our University through posted flyers. 57% of our participants were female and 90% of our participants were aged 18-25. Of those 28, we removed 5 from our analysis (3 due to technical issues and the other 2 due to failure to participate in the distractor task in *any* of the trials). We gained approval for the study by our University's Internal Review Board. Each participant gave informed consent and went through a guided tutorial, and then through three trials. Subjects were told the robot behaviors *may* change between trials. Each participant was compensated $20 per hour. Specifically, when the participants arrived, they were given the consent form. After reading it alone, a researcher went through the form with them. If they wished to proceed, they signed the form and were given an introduction to the task space. They were set up with the Vive headset and the eye

tracking was calibrated. There was a tutorial period, where the participant learned to operate the controllers, received some background information and was allowed to practice for as long as desired. In the tutorial, every participant completed both the distractor and the primary repair task. The experimenter ensured that each participant was aware of the tasks and how to perform each basic function. While the participant was in VR, we took behavioral notes, paying particular attention to how interactions with the robots differed between different behavior conditions, how stressed the participant seemed, and any failed attempts to complete the tasks. Additionally, we kept a screen capture, microphone recording, eye tracking, and event logs. After the participant was comfortable in the tutorial, we ensured they were feeling okay (e.g., did not have motion sickness from the VR environment) and were ready to move on. The appropriate robot behavior was loaded and the trial begun. Each trial lasted 7 minutes and was followed by a brief in-VR survey on their perception of the robot that typically lasted 3 minutes. After finishing the survey, we checked in to make sure the participant was still feeling okay. Many participants took a break at some point between trials due to eye fatigue or mild VR sickness. Eye tracking calibration was set up again after each break. After the final trial, we removed the VR headset and checked in again. After a couple moments to readjust, we took participants back to the initial room for a post survey which collected subjective data about the task, workload, and robot behaviors. Finally, we compensated the participant for their time.

Table 1 Results of our human behavioral study.

	Reactive		Active		Proactive			
	M	SD	M	SD	M	SD	F-value	P-value
% Tubes Repaired	21.43	15.01	25.31	19.04	36.02	20.77	3.87	.0257*
Final Spaceship Power	27.04	15.26	33.13	17.21	41.35	20.89	3.68	.0304*
Task Efficiency	183.11	13.81	177.68	19.28	161.83	28.96	6.02	.0039*
Objective Workload	6.67	10.10	1.77	9.40	1.59	11.20	23.6	< .0001*
NASA-TLX: Workload	4.36	0.65	4.57	0.76	4.22	0.73	1.46	.2386
TWLQ: Team Workload	6.93	1.41	7.14	1.24	6.76	1.56	0.42	.6569
SART: Situational Awareness	4.89	0.93	4.84	0.96	4.85	0.90	0.02	.9784

4.2 Measures and Hypothesis

Overall we expected that robot proactivity with SMMs would improve performance of human-robot teams and lower human cognitive load beyond what SMMs are able to do alone, and we further hypothesized that the most significant differences would be between the reactive and proactive conditions, with potentially no significant difference between the active and proactive conditions.

For objective performance measures, we used *percent of tubes repaired* and *final spaceship power*. We measured the *task efficiency* of the participants by computing the average time to repair a tube (in case a tube was never repaired, we defaulted to 200 seconds which is approximately the maximum repair time). We also used

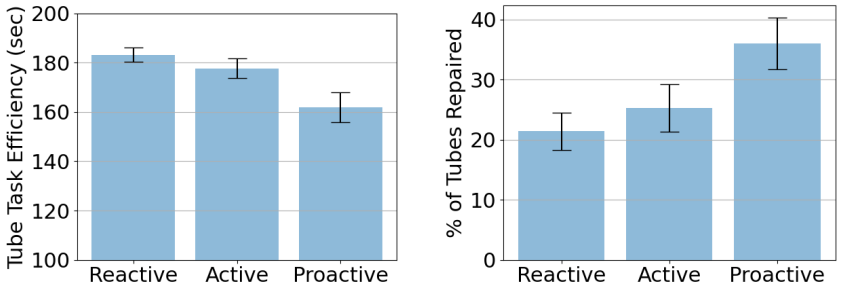

Fig. 4 (Right) We show the mean tube task efficiency per agent behavior (lower times are better). (Left) We show the mean percentage of tubes repaired per agent behavior (higher percentages are better). The error bars represent standard error. We show that the more proactive an agent is (vs reactive), the more efficient the team is and the better it performs.

moment-to-moment percent changes in pupil size (PCPS) as an *objective measure of cognitive load* throughout the experiment. Finally, we also used subjective measures of individual workload (NASA-TLX) [22], team workload (TWLQ) [38] and Situational Awareness (SA) [44] which participants provided at the end of the experiment.

H_1**: Task Performance.** We hypothesized that participants would perform better on the tube repair subtask as the robots' proactivity levels increased, with reactive robots exhibiting the worst task performance and the proactive robots the best. Thus, we predicted that the percent of tubes repaired and the final spaceship power would improve as robots become more proactive.

H_2**: Task Efficiency.** We hypothesized that participants would have an increased task efficiency, i.e., the speed at which they perform tasks and address issues, in teams with proactive agents. In the active and proactive cases, robots inform human teammates of issues needing attention, which should allow humans to address those issues sooner. Additionally, we hypothesized that in the proactive case, the robot offering to start the task should improve the efficiency. Therefore, we predicted that the task completion efficiency would improve with increased robot proactivity.

H_3**: Workload.** We collected pupillometry signals from 15 subjects (for whom usable data was available for all three experimental trials, for 8 data was only partial) during the whole experiment at a sampling frequency of 120 Hz and applied the three-step pre-processing described in [1] to calculate PCPS (instead of using the pupil diameter itself to reduce the subject-based variations, see [48]) for the prediction of cognitive workload. We hypothesized that the task load imposed on humans by the multi-tasking paradigm would be reflected in high cognitive workload and that proactive robot behavior would be able reduce it.

Since it is more challenging to evaluate subjective assessments due to ordinal scaling used in subjective surveys [25], we also hypothesized that subjective workload measures might show different results due to the overall high task load, e.g., because participants might subjectively experience workload differently, or because they

might not be able to accurately compare their workload across the three different conditions and differentiate them.

H_4**: Situational Awareness.** Since robots guide the maintenance task, we worried that humans might be less aware of the state of the ship as a whole the more proactive a robot gets, thereby potentially decreasing situational awareness as proactivity increases.

5 Results

We performed several one-way ANOVAs together with Tukey HSD tests on the above performance measures to determine whether the data supported our hypotheses. In all one-way ANOVAs, the independent variable was the robot behavioral condition. See Table 1.

Supported H_1**: Task Performance.** We found that the percent of tubes repaired was higher as the robots' proactivity increased ($F = 3.87, p = .0257, \eta^2 = .10$). Using Tukey's HSD test, we found that there was no statistically significant difference between the reactive and active behaviors ($p = .73$) or the active and proactive behaviors ($p = .12$). However, the difference between reactive and proactive behaviors was significant ($p = .024$). We plot the % of tubes repaired metric in Fig. 4. We found that the final spaceship power was higher as the robots' proactivity increased ($F = 3.68, p = .0304, \eta^2 = .10$). Using Tukey's HSD test, we found no significant difference between the reactive and active behaviors ($p = .48$) or the active and proactive behaviors ($p = .27$), but again a significant difference between the reactive and proactive behaviors ($p = .023$). Our results support H_1, i.e., teams with proactive agents achieve better task performance than ones with reactive agents.

Supported H_2**: Task Efficiency.** The task efficiency significantly improved when the agents were proactive ($F = 6.02, p = .0039, \eta^2 = .15$). Tukey's HSD test showed no significant difference between the reactive and active behaviors ($p = .65$), but there was a statistical significance between the active and proactive behaviors ($p = .04$) and the reactive and proactive behaviors ($p = .0039$). We plot the task efficiency metric in Fig. 4. Our results thus support H_2, i.e., teams of humans and proactive agents achieve the best task efficiency compared to teams with reactive or active agents.

Supported H_3**: Workload.** Objective workload was significantly improved when agents were more proactive ($F = 23.6, p < .0001, \eta^2 = .05$). A Tukey's HSD test showed significant differences between reactive and active ($p < 0.0001$) and reactive and proactive ($p < .0001$), but no significant difference between active and proactive behaviors ($p = .974$). Fig. 5 illustrates an example violin plot captured from one subject and indicates the change in the PCPS for three robot behavior conditions which demonstrates the robustness of the PCPS in assessing different workload levels. We did not find any significant effect on the participants' subjective workload assessments, i.e., the NASA-TLX showed little difference across conditions ($F = 1.46, p = .23, \eta^2 = .04$), and the team workload metric (TWLQ) also did not

show a significant effect across different conditions ($F = .42, p = .65, \eta^2 = .012$), supporting our suspicion that subjective workload measures might diverge from the objective ones, supporting H_3.

Not supported H_4: Situational Awareness. There is no statistical significance on the situational awareness metric across conditions ($F = .02, p = .978, \eta^2 = .0006$). It appears that the situational awareness of the participants is not affected by the different conditions. Thus, the results do not provide evidence for H_4 that proactive robot behaviors decrease human situational awareness.

Additional analyses. We performed additional analyses to ensure that the additional communications by proactive robot did not affect the above results because proactive robots to spend more time communicating, which increases their response time (since all robots will first finish their utterance before answering a question or responding to a human command), compared to reactive robots which can respond faster. Additional ANCOVAs analyses with "response time" as a covariate (defined as the interval from the time of a human speaking compared to the time of a robot responding) on the task performance measures showed that the performance results are still significant, even after accounting for the robot response times (Table 2). Specifically, for the % of tubes repaired ($p = .022$), for the final spaceship ($p = .025$) and for the task efficiency ($p = .003$). Thus, it was indeed the robot behaviors that were responsible for the performance effects as opposed to the response times.

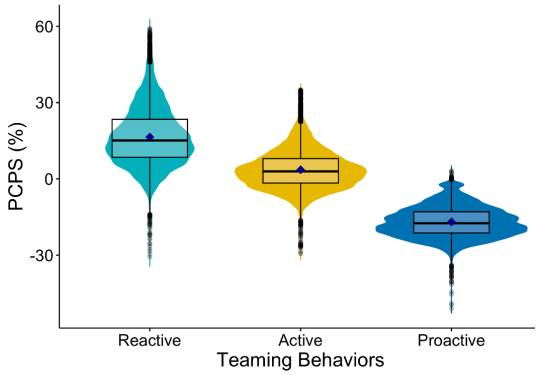

Fig. 5 An example violin plot from one subject showing the differences in workload as measured by PCPS for the three robot behavior conditions.

Participant Feedback. We can only briefly report on some of the written and oral feedback we received. Based on the post survey, 67% of the participants thought of the robots as "teammates", as opposed to "subordinates" or "supervisors", in at least one of their trials, but tended to only think the robots communicated well in the proactive case (50% vs 30% for the other trials and 18% for never communicating well). A participant after the proactive trial, which was not their first trial said *"It was much easier. The robots thought for themselves so I didn't have to instruct as much."*. Participants' quotes suggest that they appreciated more proactive teammates. They

Table 2 Ancova on Conditions of Behaviors vs Response Time.

Condition	Source	F	P-unc	Np2
Task Efficiency	Behavior	6.2803	.0032*	.1619
	Response Time	0.9010	.3460	.0136
% Tubes Repaired	Behavior	4.0424	.0221*	.1106
	Response Time	0.6875	.4100	.0104
Final Spaceship Power	Behavior	3.8826	.0255*	.1067
	Response Time	0.9720	.3278	.0147

also reported finding robots more helpful as proactivity increased, participants on the post survey, with half of the participants liking the proactive robot behavior best. One participant noted *"The robots were thinking for themselves and being proactive towards addressing the damage."* The participants' preference for proactive behavior was sometimes exhibited verbally during their proactive trial with one participant mentioning *"These robots are helping me out this time and I really appreciate that."*. The post survey also asked participants whether their views on robots changed after participating in our study. One participant wrote *"There's definitely room for humans and robots to work together to advance human goals."*.

6 Discussion and Future Work

Overall, there experimental evaluation demonstrated for the first time that proactive robot behaviors paired with shared mental models on robots can significantly improve the performance of mixed-initiative human-robot teams and lower human cognitive workload compared to robots with just shared mental models.

Proactive robots with SMMs increase task performance. We found improvement in our main task performance metrics: percent of tubes repaired and final spaceship power. A person could always tell if a tube was broken, but only through a robot could they tell exactly how much the tube was broken. This means that in the active and proactive cases, the human teammate could prioritize their tasks. However, based on our data, only the proactive case actually made a statistically significant difference to our task performance metrics. Our results show that volunteering information to the human does not suffice to improve teamwork. Robots need to actively take responsibility in shared tasks to be better teammates and achieve more effective teams.

All robots patrol the spaceship and record information about damaged tubes but the reactive agent does not offer any information voluntarily. Rather, human teammates have to ask the robot in order to get any information about broken tubes (or patrol the spaceship) and then they have to select a damaged tube in need of repair and ask the robot to go repair it. An active (or proactive) robot, in addition to

patrolling, will inform the human about any damaged tubes, and which is the most damaged one, before leaving a spaceship wing. This behavior relieves the human from the burden of having to figure out which tubes are damaged on their own. Additionally, the proactive agent, will ask the human if the tube should be repaired. The combination of the human not needing to determine which tube to be repair and being relieved from some decision-making, allows the proactive human-robot team to repair more tubes and thus perform better than the reactive human-robot team. Interestingly, we found that there is no significant difference in performance of the reactive versus active agents, or the performance of the active versus proactive agents. This leads us to believe that just informing the human or just initiating tasks alone does not suffice to make an impact on task performance but it is the conjunction of both behaviors that is more effective.

Future work. Future work could examine if a robot which initiate tasks without provide relevant information prior to initiation would fare significantly better than the reactive robot, and whether this robot's behavior would affect situational awareness. It would also be interesting to investigate the limits of proactivity, i.e., when robots that talk too much block humans from completing other tasks that require communication because humans can only process one conversation at a time [9].

Proactive robots with SMMs improve task efficiency. Proactive robots improve the task efficiency over reactive robots. Additionally, teams with proactive robots perform significantly better than teams with active robots. Proactive robots both inform the human of damaged tubes, but also initiate the task of fixing them. The salient difference between active and proactive robots is this task initiation. Our results show that the simple act of initiating a task, without changing any of the steps, had a positive impact on team efficiency. For example, the active robot will inform the human of how many tubes are breaking in a specific area, and how broken the most damaged tube is. Human teammates then needs to determine if the robot should repair it. Then, they have to remember which tube was broken and decide which robot to ask to repair the tube. If it is the robot that reported the broken tube, while it might have waited a bit for a human command to repair it, by the time it gets the command it might already be on its way to patrol a different wing. In contrast, the proactive robot will ask the human specifically if it should repair the particular tube and wait for a response. The human only needs to respond "yes" or "no". In case of a positive response, the robot will go to the tube and wait for it to be turned off by the human. We believe that making initiating a tube repair task easier in addition to the robot not leaving the area, significantly contributed to team efficiency.

Future work. While our results indicate that more proactive agents lower objective workload, further research could attempt to better classify what level of proactivity is best for different settings. For example, in our case, the distractor task required the humans to add a symbol on a map in a specific place on the spaceship. If, instead, the human had a more involved and time-sensitive task, like mixing sensitive chemical compounds, then a less distracting level of proactivity might be more appropriate.

Proactive robots with SMMs reduce objective workload. As hypothesized, objective workload was significantly reduced with proactive robots as measured in terms of PCPS which has been recently shown to be one of the most effective ob-

jective workload measures (e.g., [1]). Importantly, this was the case even for only 20 random samples taken from over 40,000 pupil size measurements recorded in each trial (higher sample numbers only increase the significance of the results), thus helping teammates to maintain a high level of performance even in stressful environments [39]. Interesting, as we considered possible, subjective workload measures did not show a significant difference between conditions, likely because the subjective experience of an overall highly stressful and difficult task did not allow subjects to distinguish their different cognitive states in retrospect. This is important going forward for teaming studies as subjective workload assessments, as they are often used in HRI studies, are not adept at distinguishing moment-by-moment cognitive effort. Yet, it is this moment-by-moment cognitive load that is an important determinant of human performance human errors and can be the source of errors and performance decrease when not kept within optimal levels.

Future work. We focused on a high cognitive load environment with a distractor task. Future work could investigate the extent to proactive behavior is able to reduce cognitive load as a function of task load, thus helping to quantify when proactive behaviors cease to confer cognitive load benefits to humans.

Proactive robots with SMMs do not seem to negatively affect situational awareness. We were initially worried that robots taking initiative would lead humans to pay less attention to the state of the spaceship and thus the overall task state, lowering their situational awareness. However, levels of human situational awareness were similar across trials, suggesting that humans overall remained informed about the task state even with proactive robots.

Future work. Future studies could develop paradigms for online probing of SA during task performance to determine the extent to which proactivity might have an impact on it.s

Shared Mental Models and Proactivity. We know based on past published work (e.g., [19]) that SMMs improve performance in mixed-initiative human-robot teams because SMMs allow the robots to be aware of each other's actions, dialogue, and knowledge and thus prevent them from overloading human teammates with communication of redundant or outdated information, which can lead to information fatigue, loss of trust [34], and cause the human to ignore their teammates. Moreover, SMMs allow robots to coordinate their actions better (e.g., distributing themselves more effectively in our spaceship setting as opposed to robots without SMMs who may end up moving to the same area, and then informing the human of the same information, which leads to lost time and two agents effectively doing the work of one.

We also know that based on past published work (e.g., [2]) that proactive robot behavior improves performance, and in this study we showed for the first time that proactive robot behavior paired with SMMs can further improve the performance of mixed-initiative teams over robots just SMMs.

Future work. It would be interesting to perform additional studies to determine how well teams with any of our three robots behaviors would perform without SMMs as an additional yard stick for the utility of SMMs. Future studies could also test even more proactive behaviors, e.g., for the robot to take on more of the decision-

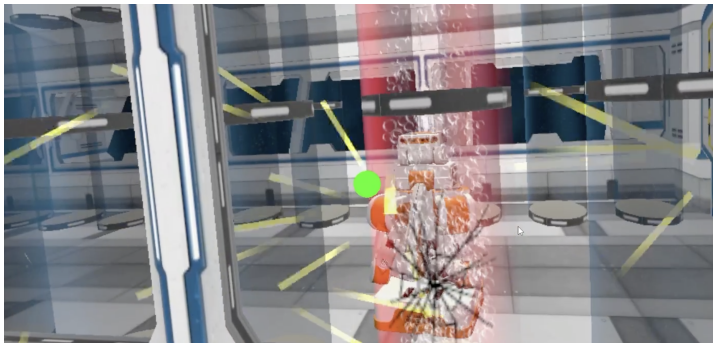

Fig. 6 Human-robot team repairing a tube (human perspective), also shown in https://streamable.com/84x7bi.

making process, only asking the human to turn on/off certain tubes, allowing them to better focus on their main (or in our case distractor) task, to determine the limits of proactive behavior (before robots essentially end up doing all the work).

7 Conclusion

We demonstrated for the first time that mixed-initiative human-robot teams where fully autonomous natural language enabled robots use shared mental models and proactive behaviors to support their human teammate perform significantly better than robots with just shared mental models without proactivity. Specifically, we developed several methods for proactive robot behavior in teaming contexts and integrated them into a cognitive robotic architecture which was used to control two fully autonomous PR2 robots. The utility of the developed algorithms was evaluated in human subject virtual reality study in the Unity3D environment in a complex high cognitive load team task. The results showed that more proactive behaviors lead to better objective team performance and to objective reduction in human cognitive workload without negatively affecting human situational awareness. This is an important result contributing evidence to the growing literature on human-robot teaming in HRI that more autonomous, more task and human-aware robots will make overall better teammates and significantly improve the performance of mixed-initiative human-robot teams.

8 Acknowledgment

Funding for this work was in part provided by AFOSR grant # FA9550-18-1-0465.

References

1. Aygun, A., Nguyen, T., Haga, Z., Aeron, S., Scheutz, M.: Investigating methods for cognitive workload estimation for assistive robots. Sensors **22**(18), 6834 (2022)
2. Baraglia, J., Cakmak, M., Nagai, Y., Rao, R., Asada, M.: Initiative in robot assistance during collaborative task execution. In: 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 67–74 (2016). DOI 10.1109/HRI.2016.7451735
3. Baraglia, J., Cakmak, M., Nagai, Y., Rao, R.P., Asada, M.: Efficient human-robot collaboration: When should a robot take initiative? The International Journal of Robotics Research **36**(5-7), 563–579 (2017). DOI 10.1177/0278364916688253. URL https://doi.org/10.1177/0278364916688253
4. Beer, J.M., Fisk, A.D., Rogers, W.A.: Toward a framework for levels of robot autonomy in human-robot interaction. Journal of human-robot interaction **3**(2), 74 (2014)
5. Berka, C., Levendowski, D.J., Lumicao, M.N., Yau, A., Davis, G., Zivkovic, V.T., Olmstead, R.E., Tremoulet, P.D., Craven, P.L.: Eeg correlates of task engagement and mental workload in vigilance, learning, and memory tasks. Aviation, space, and environmental medicine **78**(5), B231–B244 (2007)
6. Bhattacharjee, T., Gordon, E.K., Scalise, R., Cabrera, M.E., Caspi, A., Cakmak, M., Srinivasa, S.S.: Is more autonomy always better? exploring preferences of users with mobility impairments in robot-assisted feeding. In: Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, HRI '20, p. 181–190. Association for Computing Machinery, New York, NY, USA (2020). DOI 10.1145/3319502.3374818. URL https://doi.org/10.1145/3319502.3374818
7. Bisk, Y., Yuret, D., Marcu, D.: Natural language communication with robots. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 751–761 (2016)
8. Bitkina, O.V., Park, J., Kim, H.K.: The ability of eye-tracking metrics to classify and predict the perceived driving workload. International Journal of Industrial Ergonomics **86**, 103193 (2021)
9. Brodbeck, C., Hong, L.E., Simon, J.Z.: Rapid transformation from auditory to linguistic representations of continuous speech. Current Biology **28**(24), 3976–3983.e5 (2018). DOI https://doi.org/10.1016/j.cub.2018.10.042. URL https://www.sciencedirect.com/science/article/pii/S096098221831409X
10. Buerkle, A., Matharu, H., Al-Yacoub, A., Lohse, N., Bamber, T., Ferreira, P.: An adaptive human sensor framework for human–robot collaboration. The International Journal of Advanced Manufacturing Technology pp. 1–16 (2022)
11. Cakmak, M., Chao, C., Thomaz, A.L.: Designing interactions for robot active learners. IEEE Transactions on Autonomous Mental Development **2**(2), 108–118 (2010). DOI 10.1109/TAMD.2010.2051030
12. Cakmak, M., Thomaz, A.L.: Designing robot learners that ask good questions. In: 2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 17–24 (2012). DOI 10.1145/2157689.2157693
13. Cohen, P.R., Levesque, H.J., Smith, I.: Sia on team formation. In: J. Hintikka, R. Tuomela (eds.) Contemporary Action Theory. Kluwer Academic Publishers (1997)
14. Converse, S., Cannon-Bowers, J., Salas, E.: Shared mental models in expert team decision making. Individual and group decision making: Current issues **221**, 221–46 (1993)
15. Dickinson, T.L., McIntyre, R.M.: A conceptual framework for teamwork measurement. In: Team performance assessment and measurement, pp. 31–56. Psychology Press (1997)
16. Espinosa, J.A., Kraut, R.E., Slaughter, S., Lerch, F.J., Herbsleb, J.D., Mockus, A.: Shared mental models, familiarity, and coordination: A multi-method study of distributed software teams. In: F. Miralles, J. Valor (eds.) Proceedings of the International Conference on Information Systems, ICIS 2002, Barcelona, Spain, December 15-18, 2002, p. 39. Association for Information Systems (2002). URL http://aisel.aisnet.org/icis2002/39

17. Fan, X., Yen, J.: Modeling and simulating human teamwork behaviors using intelligent agents. Physics of Life Reviews **1**(3), 173–201 (2004)
18. Fong, T., Thorpe, C., Baur, C.: Multi-robot remote driving with collaborative control. IEEE Transactions on Industrial Electronics **50**(4), 699–704 (2003). DOI 10.1109/TIE.2003.814768
19. Gervits, F., Thurston, D., Thielstrom, R., Fong, T., Pham, Q., Scheutz, M.: Toward genuine robot teammates: Improving human-robot team performance using robot shared mental models. In: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '20, p. 429–437. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2020)
20. Goodrich, M.A., Yi, D.: Toward task-based mental models of human-robot teaming: A bayesian approach. In: International conference on virtual, augmented and mixed reality, pp. 267–276. Springer (2013)
21. Grosinger, J., Pecora, F., Saffiotti, A.: Making robots proactive through equilibrium maintenance. In: IJCAI, vol. 16, pp. 3375–3381 (2016)
22. Hart, S.G.: Nasa-task load index (nasa-tlx); 20 years later. In: Proceedings of the human factors and ergonomics society annual meeting, no. 9 in 50, pp. 904–908. Sage publications Sage CA: Los Angeles, CA (2006)
23. Hoover, A., Singh, A., Fishel-Brown, S., Muth, E.: Real-time detection of workload changes using heart rate variability. Biomedical Signal Processing and Control **7**(4), 333–341 (2012)
24. Howard, Z.L., Innes, R., Eidels, A., Loft, S.: Using past and present indicators of human workload to explain variance in human performance. Psychonomic Bulletin & Review **28**(6), 1923–1932 (2021)
25. Jahedi, S., Méndez, F.: On the advantages and disadvantages of subjective measures. Journal of Economic Behavior & Organization **98**, 97–114 (2014)
26. Jonker, C.M., van Riemsdijk, M.B., Vermeulen, B.: Shared mental models. In: M. De Vos, N. Fornara, J.V. Pitt, G. Vouros (eds.) Coordination, Organizations, Institutions, and Norms in Agent Systems VI, pp. 132–151. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)
27. Lin, C.J., Lukodono, R.P.: Classification of mental workload in human-robot collaboration using machine learning based on physiological feedback. Journal of Manufacturing Systems **65**, 673–685 (2022)
28. Mathieu, J.E., Heffner, T.S., Goodwin, G.F., Salas, E., Cannon-Bowers, J.A.: The influence of shared mental models on team process and performance. Journal of applied psychology **85**(2), 273 (2000)
29. McIntyre, R.M., Salas, E.: Measuring and managing for team performance: Emerging principles from complex environments. Team effectiveness and decision making in organizations **16**, 9–45 (1995)
30. Nikolaidis, S., Shah, J.: Human-robot teaming using shared mental models. ACM/IEEE HRI (2012)
31. Orasanu, J.: Shared mental models and crew decision making. Princeton, NJ (1990)
32. Ososky, S., Schuster, D., Jentsch, F., Fiore, S., Shumaker, R., Lebiere, C., Kurup, U., Oh, J., Stentz, A.: The importance of shared mental models and shared situation awareness for transforming robots from tools to teammates. In: Unmanned Systems Technology XIV, vol. 8387, pp. 397–408. SPIE (2012). DOI 10.1117/12.923283
33. Rosenthal, S., Dey, A.K., Veloso, M.: How robots' questions affect the accuracy of the human responses. In: RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication, pp. 1137–1142 (2009). DOI 10.1109/ROMAN.2009.5326291
34. Salem, M., Lakatos, G., Amirabdollahian, F., Dautenhahn, K.: Would you trust a (faulty) robot? effects of error, task type and personality on human-robot cooperation and trust. In: 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 1–8 (2015)
35. Scheutz, M., DeLoach, S., Adams, J.: A framework for developing and using shared mental models in human-agent teams. Journal of Cognitive Engineering and Decision Making **11**, 203–224 (2017)
36. Scheutz, M., DeLoach, S.A., Adams, J.A.: A framework for developing and using shared mental models in human-agent teams. Journal of Cognitive Engineering and

Decision Making **11**(3), 203–224 (2017). DOI 10.1177/1555343416682891. URL https://doi.org/10.1177/1555343416682891

37. Scheutz, M., Williams, T., Krause, E., Oosterveld, B., Sarathy, V., Frasca, T.: An overview of the distributed integrated cognition affect and reflection diarc architecture. Cognitive architectures pp. 165–193 (2019)
38. Sellers, J.M.: Team workload questionnaire (twlq): Development and assessment of a subjective measure of team workload. Ph.D. thesis, University of Canterbury. Psychology (2013)
39. Serfaty, D., Entin, E.E., Volpe, C.: Adaptation to stress in team decision-making and coordination. Proceedings of the Human Factors and Ergonomics Society Annual Meeting **37**(18), 1228–1232 (1993). DOI 10.1177/154193129303701806. URL https://doi.org/10.1177/154193129303701806
40. She, L., Cheng, Y., Chai, J.Y., Jia, Y., Yang, S., Xi, N.: Teaching robots new actions through natural language instructions. In: The 23rd IEEE International Symposium on Robot and Human Interactive Communication, pp. 868–873 (2014). DOI 10.1109/ROMAN.2014.6926362
41. Smith-Jentsch, K.A., Johnston, J.H., Payne, S.C.: Measuring team-related expertise in complex environments. In: J.A. Cannon-Bowers, E. Salas (eds.) Making decisions under stress: Implications for individual and team training, pp. 61–87. American Psychological Association (1998)
42. Stout, R.J., Cannon-Bowers, J.A., Salas, E., Milanovich, D.M.: Planning, shared mental models, and coordinated performance: An empirical link is established. Human Factors **41**(1), 61–71 (1999). DOI 10.1518/001872099779577273. URL https://doi.org/10.1518/001872099779577273
43. Stuiver, A., Brookhuis, K.A., de Waard, D., Mulder, B.: Short-term cardiovascular measures for driver support: Increasing sensitivity for detecting changes in mental workload. International Journal of Psychophysiology **92**(1), 35–41 (2014)
44. Taylor, R.: Situational awareness rating technique(sart): The development of a tool for aircrew systems design. In Proceedings of Situational Awareness in Aerospace Operations (AGARD) (1990)
45. Tellex, S., Gopalan, N., Kress-Gazit, H., Matuszek, C.: Robots that use language. Annual Review of Control, Robotics, and Autonomous Systems **3**, 25–55 (2020)
46. Yen, J., Fan, X., Sun, S., Hanratty, T., Dumer, J.: Agents with shared mental models for enhancing team decision makings. Decision Support Systems **41**(3), 634–653 (2006). DOI https://doi.org/10.1016/j.dss.2004.06.008. URL https://www.sciencedirect.com/science/article/pii/S0167923604001368. Intelligence and security informatics
47. Zhang, Y., Narayanan, V., Chakraborti, T., Kambhampati, S.: A human factors analysis of proactive support in human-robot teaming. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3586–3593 (2015). DOI 10.1109/IROS.2015.7353878
48. Zhao, M., Gao, H., Wang, W., Qu, J., Chen, L.: Study on the identification of irritability emotion based on the percentage change in pupil size. In: Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing, pp. 20–24 (2020)