Exploring Coordination in Human-Robot Teams in Space

Felix Gervits[*], Charlotte Warne[*], Harrison Downs[*],
Kathleen Eberhard[†], and Matthias Scheutz[*]
[*]*Tufts University, Department of Computer Science, Medford, MA 02155*
[†]*University of Notre Dame, Department of Psychology, Notre Dame, IN 46556*

Human-robot teams in space environments are difficult to evaluate, in large part because performance of these teams is influenced by a variety of factors, including team size, structure, and composition. We introduce and describe a novel experimental framework that is sensitive to these factors, and that serves as a testbed to facilitate the study of human-robot teaming in space. We also report on the results of a preliminary study in this framework that involves a human interacting with a simulated Mars rover. Our findings show that people exhibited great variation in strategy and performance, and point to the role that decision-making and task-switching may have played in this result. This study is the first in a larger effort to develop a rich multimodal corpus and to investigate various dimensions of teaming in this domain.

I. Introduction

Robots are being increasingly used in multi-agent teams with astronauts and ground crew to assist in the exploration of space. Success of these mixed-agent teams largely depends on how well they can communicate plans and goals, as well as coordinate actions to solve problems.[1] Though current approaches have shown some success in improving collaborative task performance,[2] we are still far from achieving the levels of coordination seen in human-human teams.[3] On top of the many challenges of coordination in human-robot teams, there are additional interaction challenges for teams operating in space environments.[4] One such challenge is that teams are often spatially distributed, making it difficult to achieve situational awareness and integrate perceptual context from multiple sources. The extent of this spatial distribution ranges from teammates in separate rooms on the same ship to astronauts in Mars orbit and ground control on Earth coordinating with astronauts and robots on the Martian surface. Certain team structures may also involve a variable degree of proximity, wherein teammates alternate between proximate and remote interaction in the same task.[5] Another challenge is varying time scales. Though co-located teammates can typically communicate without any latency, ground control may experience long delays or even signal loss when communicating across large distances (e.g., Earth to Mars). Since ground control and astronaut crews are a distributed team which must make decisions together, it is important that monitoring and control of the robots be shared between all teammates operating at different time scales. Another challenge is team heterogeneity, which involves variable team size, structure, and distribution. For example, some missions may involve only co-located humans, but others may involve mixed teams of spatially-distributed agents. Overall, additional work is clearly needed to understand the role that each of these factors independently (and collectively) plays in affecting team functioning. As it stands, these issues create a bottleneck in the future application of space robots, making it difficult to envision how critical tasks such as site preparation, habitat construction, and intra/extra-vehicular activity can be performed effectively.

While very little work has been done on addressing these particular coordination challenges in space domains (but see Ref. 6), there have been several efforts at studying the above issues in the broader human-robot teaming literature. One subset of the literature attempts to glean insight about fundamental issues of coordination and teaming by investigating human-human task performance.[7–9] These are useful first steps for understanding the kinds of interaction patterns that make effective partners, and the findings may certainly inform future robot policies. However, having robots on a team introduces many additional

variables that are not accounted for in these studies.[10] Thus, without including robots in the actual task it is difficult to apply the results directly to human-robot teams. Efforts involving robot teammates exist, but they are largely limited to domains in which spoken language route-instructions are given to remotely-located robots.[11–13] While insights from these studies can certainly be applied to search-and-rescue scenarios, the domains themselves are not particularly relevant for current space robotics applications. There have been a few studies that explored collaborative task execution between co-located humans and robots.[14,15] One study in particular investigated the kinds of heterogeneous, mixed-initiative teams of interest to space robotics.[5] However, the task was somewhat complex, involving specific robots with specialized roles, and so would not be able to scale to variable team sizes and structures without significant changes to the design. What is needed is a framework that is scalable and customizable in order to allow for a variety of experimental manipulations while retaining the same basic design and metrics.

To fill this gap, we have developed a novel experimental framework to study human-robot teaming in realistic space environments. The framework is designed to be flexible and scalable with respect to varying team sizes and structures, and also realistic with respect to the use of robots in current and future space operations.[1] The main benefit of such a customizable design is that it allows researchers to modify individual study parameters (e.g., team size) without needing to redesign the entire task. This enables the straightforward implementation of new experimental conditions, and the ability to compare results between these conditions to see how the changes affected performance and other metrics. The long-term goal of the project is to collect data in various experimental conditions and to develop a corpus of multi-agent teaming in a space domain. This will serve as a testbed to examine various empirical issues in the field of human-robot interaction, such as team cohesion under time pressure, the impact of a robot on team performance, and language as a coordination tool. Analysis of the corpus will also advance robotic technologies, informing computational mechanisms needed by robots to function as more effective teammates.[16] The present paper will serve as an introduction to the framework and task domain. We will also discuss the results of a pilot study in this domain to demonstrate the framework and how people approached the task. Data collection and corpus curation will be an ongoing process in the years to come, and it is expected that the resulting corpus will serve as a useful research tool in many related fields.

II. Task Domain

The task domain is an in-flight maintenance task to simulate robotic assistance on a spacecraft. It will take place in an indoor environment consisting of three rooms connected by various narrow passageways representing the interior of a spacecraft (see figure 1). The rooms are empty except for a computer in the starting room, which is used to complete the task. The scenario will work as follows: participants are told that they are astronauts aboard a spacecraft orbiting Mars, and will need to complete several objectives. The primary objective is to oversee a planetary rover carry out a geological survey mission on the Martian surface. The participant is told that this crucial mission is being done to prepare a site for an incoming colonization crew. They are also told that during this task, their ship will pass through space debris, which may cause damage to critical components on board the ship. When this happens, they must seek out the damaged components and repair any resulting air leaks. In short, the participant must balance two tasks: 1) update the map of the critical area on the planet via information from the remote rover, and 2) find and repair air leaks on the ship. It is not possible to suspend the survey task to fix air leaks because the rover proceeds along a scripted path and has limited time to navigate the critical area on the planet. It also is not possible to ignore the air leaks because both the duration and number decrease air pressure in the spacecraft, which must remain above a critical minimum level. If the cabin pressure drops too low, then the ship will need to make an emergency landing, thus ending the mission. Participants are instructed to persist in the task as long as possible so that the incoming colonization mission is not jeopardized. Task difficulty is increased over time by increasing the rate of air leaks. This is done in order to provide a consistent (and known) level of task-induced workload across participants and to make the task difficult so as to require coordination in the team conditions.

All participants will be equipped with a bluetooth headset to be able to hear the rover and ship announcements. The headset also has a built-in microphone, which will be used for communication in the team conditions. During the task, participants will be exposed to three types of announcements (described in the sections below): *landmark, pressure level*, and *air leak*. To prevent overlap, the announcements are scheduled in a queue and will sometimes be delayed if another announcement is in progress. Participants will

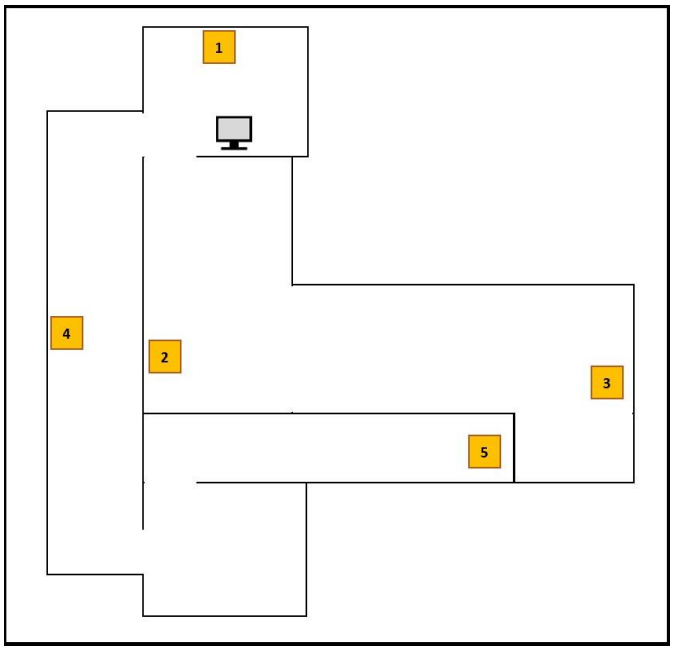


Figure 1. Schematic of the interior "spacecraft" environment. The computer is marked with a monitor icon, and the wall panels are numbered 1-5. Participants begin the task at the computer.

also be equipped with a head-strapped GoPro camera which records video for the duration of the trial. This video will be used to track participants' path management, task execution, and gaze location. In the team conditions it will be used to collect interaction data related to the teammates (e.g., proximity, gesturing, etc.).

II.A. Geological Survey Task

The primary task on which participants will be scored is the geological survey task. In the task, a simulated Martian rover explores some remote terrain, takes rock samples and radiation readings, and communicates map coordinates which correspond to the locations of the rocks and radiation zones. The "rover" is actually a series of scripts that sends transmissions (*landmark announcements*) at regular intervals to simulate movement on a map. An example landmark announcement is: "Transmitting data. Rover has encountered a rock at position G2". The rover script runs the same sequence for every participant, and is set up to send a new transmission every 30s. Rocks are transmitted at a rate of 2:1 relative to the radiation zones. The map coordinates increase slightly with each transmission to give the impression that the rover is moving within the remote environment. Note that while the rover can send messages to the participant, it does not respond to commands, as communication is only one-way in this condition.

When the rover sends a landmark announcement, a picture of the landmark is included in the transmission and sent to the on-board computer. This picture remains on the screen until the next landmark announcement. The participants' task is to mark the locations of these landmarks on a virtual map, and additionally classify the rocks into one of three categories depending on the visual data that is sent by the rover (see figure 2). To complete this categorization task, the participant is provided with a reference sheet next to the computer which shows a picture of each type of rock along with the corresponding label (volcanic, sedimentary, sandstone; see figure A2). This additional subtask was implemented to prevent memorization by requiring the participant to return to the computer after each announcement to view the image. There is only one type of radiation zone, and this is marked in the same way on the map. Thus, in principle, people can retain the coordinates for the radiation zones in memory without immediately returning to the computer. Before starting the task, participants will be allowed to practice marking landmarks on the map as long as they need until they are comfortable with the task.

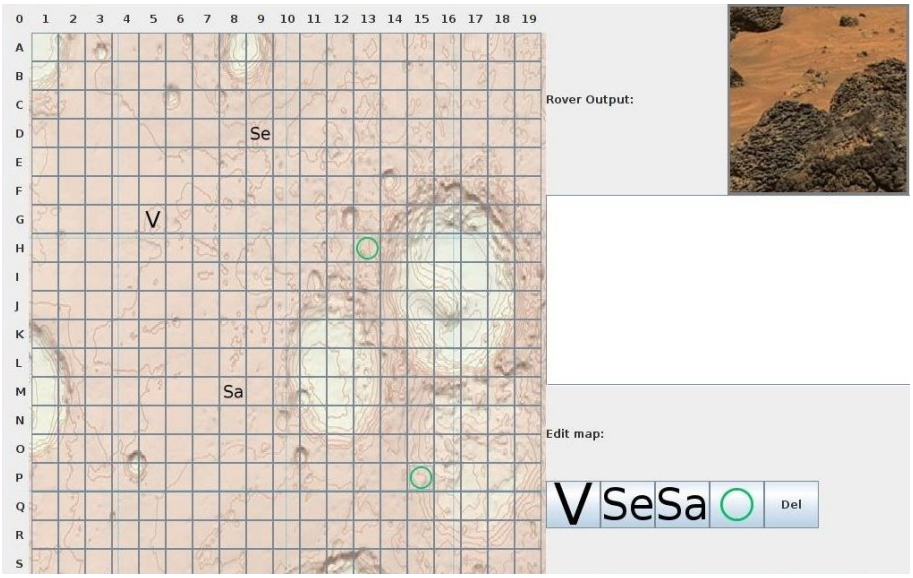


Figure 2. Study interface, including a map of the "Martian surface" for the geological survey task. Buttons V (volcanic), Se (sedimentary), and Sa (sandstone) represent the three types of rocks, and the green circle represents radiation zones. A picture of the rover's last sent image appears in the upper-right corner.

II.B. Air Leak Repair Task

While the participant is attending to the rover task, the ship will pass through orbiting debris, and will "take damage". Whenever the ship is struck by debris, a message will be transmitted to the participant's headset informing them that a wall panel has been damaged, resulting in an air leak that is causing a drop in the cabin pressure level. An example of such an *air leak announcement* is: "A new panel has been damaged. 1 panel is now damaged". The location of the damage will be one of five wall panels distributed throughout the ship environment, but the exact location of the damage will not be known to the participant. All participants will be familiarized with the ship layout and the location of each wall panel before the task begins.

The ship will also communicate periodic *pressure level announcements* as cabin pressure drops in increments of 10: "Air pressure at 60%". Cabin pressure will steadily decrease at a rate dependent on the number of damaged panels (each leak has a constant flow rate of 10% every 20s), and will increase at the same rate as the panels are repaired. If a panel is not repaired, then it is possible for multiple panels to be simultaneously damaged. This results in an accelerated flow rate of twice the normal rate for two panels, three times the normal rate for three panels, etc.

The air leaks are detected and repaired using an "ultrasonic scanner" (for the current study - a smartphone) that is moved across the surface of the panels in various compartments on the spacecraft. To simulate the repair process, each of the panels is represented by a unique QR code, and the smartphone runs a generic QR reader application to identify the particular panel. Once scanned, this information is transmitted to a webserver where it interacts with a central program to determine if the scanned panel is currently damaged. If so, a button will appear on the scanner that allows for the repair of the panel. Participants will then be able to repair the panel by pressing the button and waiting for 5 seconds (see figure 3). A final scan of the panel is then required to complete the repair process. This is done to prevent people from scanning a panel and then moving to search for others while completing the repair process. Participants will be allowed to practice scanning and repairing wall panels as long as they need until they understand the task.

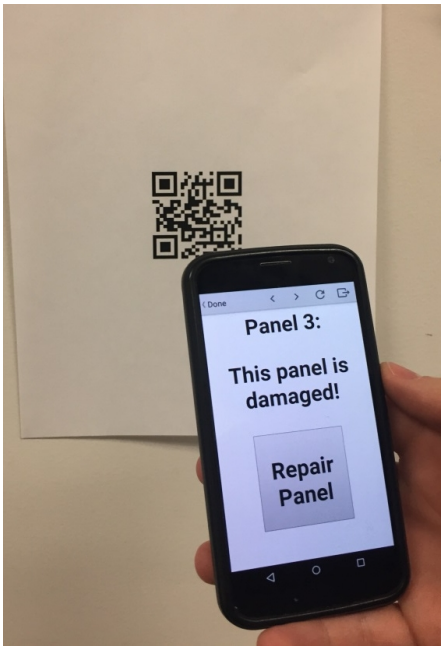


Figure 3. Sample wall panel and QR reader used to scan and repair wall panels.

The order of panel damage as well as the rate at which damage occurs is set by the program in advance and is consistent for every participant. This can be customized as needed, but for the current study the first panel (#3) becomes damaged at the 1:30 mark into the task. The rate then drops to a new panel every minute for the next two panels (#1 and #4), and then this drops to 45s for the following two panels (#5 and #4). The rate drops further to 30s after that, but no one in the current data set managed to last this long.

II.C. Performance Measures

All participants are instructed that their score will be entirely based on how well they perform in the geological survey task, i.e., the number of landmarks correctly marked on the map. The air leak repair task can be considered a kind of distractor task, but of course people will need to attend to it in order to maximize their score. We are also be recording a number of subjective measures including personality, using the Ten Item Personality Inventory (TIPI),[17] and demographics in a pre- experiment survey. After the experiment, participants fill out another survey to assess workload and situational awareness. Workload is measured using the NASA Task-Load Index (TLX) questionnaire,[18] and situational awareness is measured using the Situation Awareness Rating Technique (SART).[19] These metrics are important for understanding various team and performance dynamics, and we expect them to be customized depending on the experimental condition.

Table 1. Overview of pilot study participants.

ID	Age (yrs)	Gender	Duration (min:s)	Score	Landmarks	Distance (m)	Leaks repaired
1	41	F	4:20	6	8	74	2
2	21	M	5:10	4	10	67	4
3	22	F	5:20	5	10	88	3
4	18	F	4:12	5	8	17	1
5	20	F	3:16	3	6	89	0

III. Preliminary Results

We collected preliminary data from five participants running through the baseline condition of our task - four female, one male. The age range was 18-41, with an average age of 24 years ($SD = 9.4$). Table 1 shows the participant data, including demographic and performance measures. Score was reported as the total number of correct landmarks placed, although the total number of landmarks announced by the rover varied between participants depending on the duration of the trial. Those participants that lasted longer in the task naturally had more landmarks available to place. Despite the variability, we were mainly interested in correct landmarks placed, as these reflect the completeness of the map. 'Distance' represents the total distance traveled by the participant during the trial. This was approximated from the video files based on the path that the participant took, and known distances between various points in the environment. Due to the limited sample size in this exploratory study no statistical analyses were performed. We were mainly interested in using these pilot data to evaluate strategy and decision-making trends in the baseline condition of our task. We were also interested in investigating how people managed both subtasks, and the frequency with which they switched between them. These results will be used to further refine the experimental setup as well as to inform policies for a robot partner in future team conditions.

III.A. Task-Switching

Due to the fast-paced nature of the task, and the fact that the baseline condition involved only a single individual, frequent task-switching between the geological survey and air leak repair tasks was crucial for success. Since there were two main tasks, we define a task switch as any change in behavior that is associated with switching between the geological survey task and the air leak repair task. An example of a task-switching behavior includes leaving the computer to go search a panel. Similarly, moving back to the computer to add a landmark also constitutes a task-switch. Typically a prompt from the rover or ship was responsible for this task-switch - either an announcement of a new landmark, a new air leak, or an update on the pressure level.

In order to investigate this phenomenon more systematically, we extracted the prompted task-switching behaviors from the task tier in our transcripts (see figure 4). Whenever a participant switched tasks within 5s of an announcement, that event was marked as a task-switch.

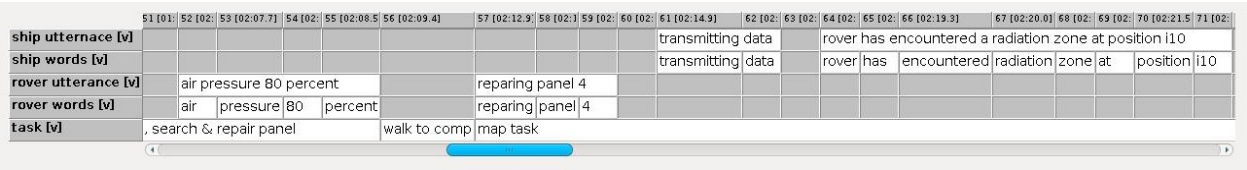


Figure 4. Sample transcript from the corpus showing the task (event) tier. Audio and video are also synchronized with the transcript, but are not shown here.

Our results indicate a wide range in task-switching frequency between the study participants (see table 2). Most of the task-switching occured immediately after a landmark or air leak announcement. This suggests that these were particularly salient messages that elicited an immediate change in behavior. Pressure level also sometimes led to an increase in panel searching, especially as the pressure level dropped to 50% or less. Interestingly, certain participants (1 and 3 in particular) engaged in prompted task-switching with greater frequency than others. This is reflected in their overall path-planning and strategy, which we explore in the Discussion section.

An important factor that reflects task switching behavior is the relative priority of each of the subtasks for each participant. To approximate this, we calculated the elapsed average time before a person started to move back to the computer following a landmark announcement (see figure 5). If they were already at the computer, that event was not included in the average, and if they started moving immediately then that event was marked as 1s. The lower times generally indicate that these people prioritized the survey task, whereas the higher times indicate a priority for the air leak repair task.

Another related indicator of subtask priority is how much time elapsed before a panel was successfully repaired following an air leak announcement (see figure 6). This gives an indication of how efficient people were at repairing the panels. Recall that if a panel is not repaired then air pressure drops fairly quickly, making proactive repair a high priority. We found that most people repaired the first panel within a minute, although one participant never repaired it. Generally, more panels repaired suggests that people were prioritizing the repair task, although participant 5 is an outlier to this trend as they unfortunately did not find even the first damaged panel. The impact of subtask priority and task-switching behavior on overall strategy is explored further in the Discussion section.

Table 2. Task-switching frequency by participant for each announcement prompt.

	Participant				
Announcement Type	**1**	**2**	**3**	**4**	**5**
Landmark	4	1	3	1	2
Pressure level	3	0	1	0	1
Air leak	4	2	3	1	1
Total task switches	11	3	7	2	4

III.B. Corpus Curation

Corpus curation is a work in progress, and we intend to make the data available for research purposes once it is ready. For the current dataset, participant audio and video data were transcribed using the EXMARaLDA[20] software, and coded for various features of theoretical interest (see figure 4 for example transcription). These features include task-relevant events such as the multiple announcement types from the ship and the rover, and task/path management of the participant (e.g., moving back to computer). Aligning these events with the announcements allowed us to evaluate prompted task-switching and subtask prioritization. Future conditions with dialogue interaction will have more robust annotations, including dialogue moves,[21] disfluencies,[22] syntax[23] and prosody.[24] Because of this fine-grained annotation, the corpus will serve as a valuable resource to researchers in human-robot interaction and space robotics. For example, in future team conditions, it will be possible to track what the human said to their teammate at the specific

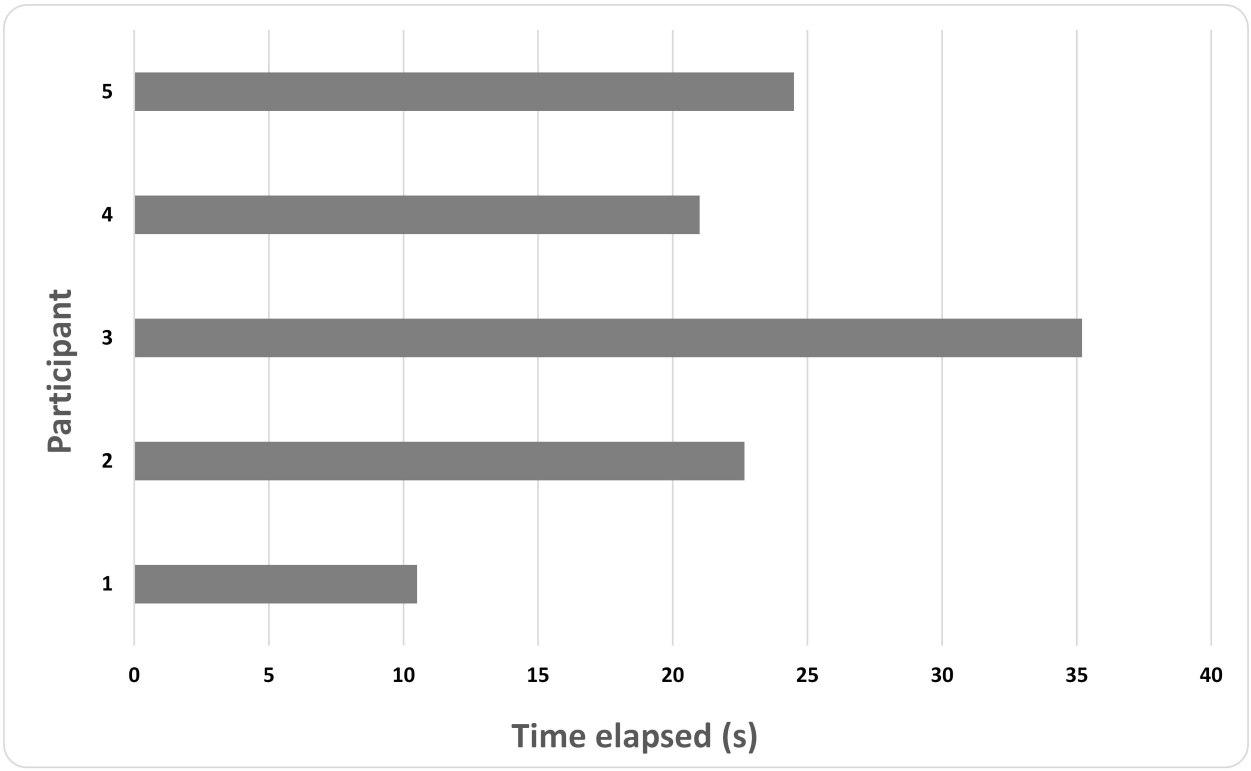


Figure 5. Average time elapsed between landmark announcement and participant starting to move back to computer.

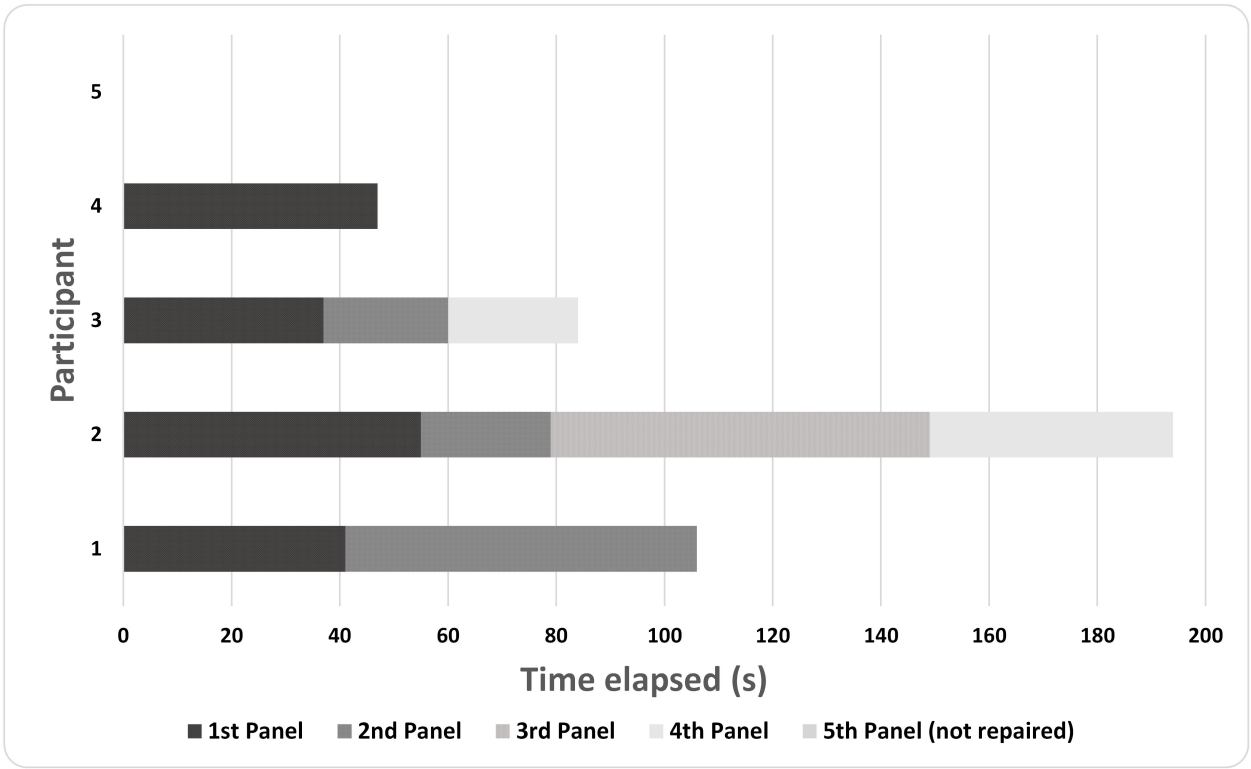


Figure 6. Time elapsed from the air leak announcement to the moment in which each panel was repaired.

point in the task when the ship announced that a new panel was damaged and the air pressure was at 60%. This precise knowledge of situational awareness, task-relevant events, and team communication will make the corpus a useful testbed for researchers interested in task-oriented dialogue (grounding, perspective-taking, referential language), teaming (coordination, strategy, effectiveness), performance (workload, situational awareness, fatigue), and the interesection of these areas. One of the main goals of this project is to use this rich multimodal dataset to develop and validate a metric for *team cohesion*. This phenomenon has proven difficult to operationalize in the past,[25] but we believe that our corpus will be particularly well suited for developing such a composite measure.

IV. Discussion and Ongoing Work

IV.A. Task-Management Strategies

The baseline condition of this task was designed to be extremely difficult for a single person to perform. This is evidenced in the short trials, low scores, and high ratings for categories like "Effort" and "Temporal load" in the workload survey (see figure 7). It is reassuring at least that, while difficult, we did find a performance range, and that people approached the task in different ways to varying degrees of effectiveness. As a result, we were able to identify several basic task-management strategies in the data. One strategy involved a particular emphasis on the geological survey task. This strategy was characterized by a short overall distance traveled, low task-switching, and a short study duration. Participant 4 best exemplified the use of this strategy. The participant spent most of her time in front of the computer adding landmarks, and largely ignored the air leaks. This person lasted only 4 minutes and 12 seconds (the second-shortest time), but they also got a decent score of 5 (the median). While the study does place importance on the survey task, it's important to also repair the air leaks in order to prolong the task.

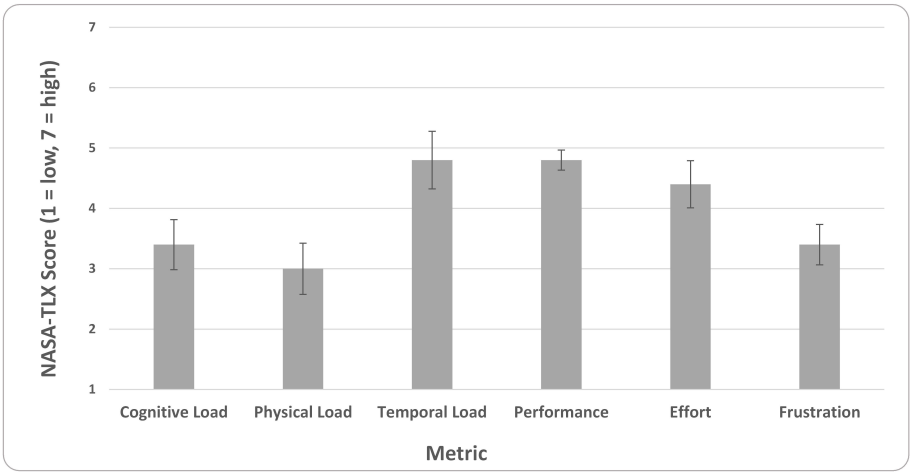


Figure 7. Average scores for the NASA-TLX survey. Error bars represent S.E.M.

Another observed strategy was, conversely, one in which the emphasis was on repairing the air leaks. This was typically characterized by a large distance traveled, low task-switching, and a variable study duration which depended on early success of finding the damaged panels. Participants 2 and 5 employed this strategy, to varying degrees of effectiveness. While participant 5 traveled the greatest distance, most of this time was spent finding the first damaged panel (which was never repaired). She was unfortunate in that the damaged panel was the last one searched, so this naturally set her back. As a result, this person finished in only 3 minutes and 16 seconds with the lowest score. Participant 2 on the other hand was fortunate in that he successfully found the damaged panel on the first try for three out of the five panels. Though this prolonged the task significantly, he missed a number of the landmarks as a result, and only scored a 4. Figure 6 is informative here, as it demonstrates the relative priority that people assigned to repairing air leaks. Here again we see that participant 2 spent much longer than anyone else searching and repairing leaks, and that participant 5 was unsuccessful in repairing any leaks.

The final strategy we observed was balancing both subtasks, which was characterized by a large distance traveled and a high degree of task-switching. Participants 1 and 3 used this strategy, and they achieved the

highest scores. While both participants balanced the two subtasks, they each slightly emphasized one over the other, which may explain differences in their performance. Participant 1 was very efficient, and managed to repair several panels while accurately placing 6 landmarks. She prioritized the survey task, and was able to put her panel-searching on hold when a landmark was announced in order to rush back to the computer. On average, when a landmark was announced by the rover she started moving to the computer within 10.5s (see figure 5). On the other hand, participant 3 took on average 35.2s to start moving to the computer after a landmark was announced. Though she lasted a minute longer in the task than participant 1, she also scored a point less due to her slight preference for the air leak repair task. Overall, the high performance of the two participants that employed this strategy suggests that balancing both subtasks is required for effective performance in this condition; however, there seems to be a benefit to slightly prioritizing the survey task.

Another interesting result is that, contrary to the well established finding that task-switching incurs a cognitive cost,[26] our subjective workload results did not indicate that the highest scoring participants were under any more workload than the others. It is possible that the time pressure and workload already inherent in the task masked any effect of task switching, or that the cognitive cost of switching may not have manifested in any of our measures. This finding is worth exploring in future studies with larger, controlled sample sizes. If there is a cognitive cost of task-switching, then it is unclear whether this burden will be alleviated in the team conditions. This is because people will still need to switch between their current task and the task of keeping track of their partner's activities and (in the case of the robot) managing their actions. Though this may increase workload more than doing the task alone, we would still expect to see a performance improvement from having multiple teammates. It will be interesting to examine the results in future conditions and see if having a partner (human or robot) may serve to offset the cognitive cost associated with task-switching, as well as to examine the extent to which a partner can improve performance (if at all).

IV.B. Ongoing Work

The current work describes the general task setup for the baseline condition, which involves a two-agent team consisting of one human and one (simulated) robot. Moving forward, we will introduce several experimental conditions, which involve changes to the team size and structure. In one condition, we will add a teleoperated robot on board the spacecraft to assist the participant. The robot will be a full-size PR2 (see figure A1) and will be able to interpret natural language commands (via Wizard-of-Oz) to help navigate the ship and repair air leaks. Having a robot partner may alleviate some of the task-switching burdens, but it will introduce new constraints involving the management of responsibilities, as well as the need to coordinate through dialogue. We expect a degree of variability in the way that people interact with the robot, with some teams coordinating more effectively than others. To compare how people interact with a physical robot vs. a human we will have another condition in which the onboard robot is replaced by a human. Now, the human-human team must still manage the task of coordination, but they can communicate in a less restricted manner, and also divide the task responsibilities more evenly. This condition is important because one of our central aims is to understand how effective teams coordinate their actions. If people perform better in this condition compared to the robot condition then the data will allow us to systematically investigate the differences (e.g., communication, movement speed, decision-making, etc.). We will be using these conditions to examine a host of multimodal interaction techniques (dialogue, gesture, etc.) that people use, in order to inform the design of an autonomous system capable of interacting in a natural way.

While we will focus on these three conditions for the time being, the task is scalable to additional team sizes and structures. In the future, we can add additional agents to the task, potentially with fixed roles, and that operate at different time scales. An example of this would be a ground control team on Earth that can help pinpoint the location of a damaged panel, but that has a delay in the time it takes to transmit this information. We can also modify the rate of air leaks or landmark transmissions, as well as the order of which panels are damaged, or even the number of panels in the ship. This will be necessary in order to scale the difficulty to the corresponding team size and structure, as additional teamamtes will likely make the task easier. Fortunately, the framework is designed to be customizable, allowing for such modifications to the study parameters.

Finally, we also plan to redesign certain elements of the task in order to increase the realism of the scenario. We are currently working on a custom scanning device which will replace the smartphone. The QR codes on the wall panels will also need to be replaced with more realistic-looking paneling that retains the same functionality. Concurrent with the present work, we are implementing a similar task setup in a fully

immersive virtual reality (VR) environment. Not only will this streamline data collection, but it will also allow us to explore larger, more varied, and more immersive environments (e.g., a Mars base). We intend to run both the physical and virtual experiments in parallel, using the results to develop a rich multimodal dataset.

V. Conclusion

The novel experimental framework described here is uniquely suited as a platform to study human-robot team interaction in space domains. The customizable design allows for the evaluation of factors such as team performance, communication under changing workload, and coordination between multiple spatially distributed agents operating at different time scales. The task domain was specifically designed to scale with variable numbers of human and/or robot teammates (including a co- located human, robot, or ground control agent on Earth), so it will serve as a testbed for theories about team coordination as well as an evaluation platform for robot technologies. We have presented preliminary data of the baseline condition, which provide insight into decision-making and task management used by participants in the task. These results show the importance of task-switching in order to manage the multiple subtasks, and point to the importance of coordination in future team conditions. As part of the larger project, the next step is the design and implementation of additional experimental conditions involving teams of co-located humans and robots in both physical and virtual environments. This will facilitate the study of various dimensions of multi-agent coordination and task-oriented dialogue, and will lead to the development of a multimodal corpus of human-robot teaming in space domains.

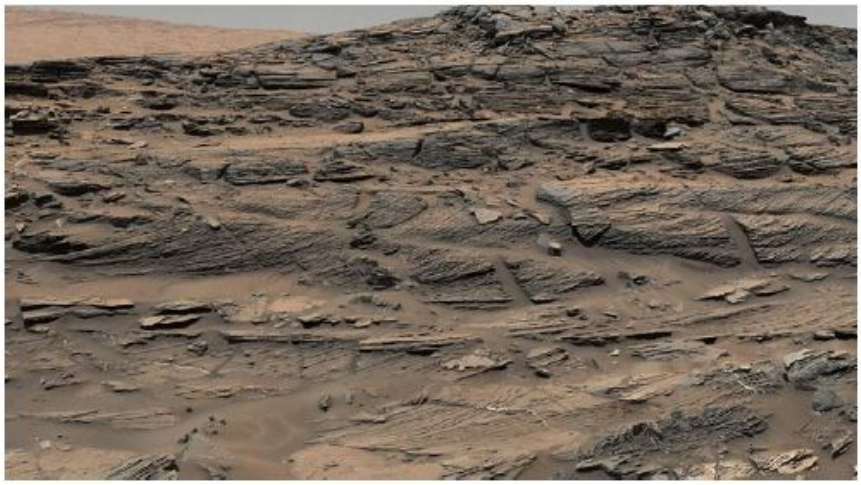
Appendix

A: Extra Figures

Figure A1. PR2 robot in the planned team condition.

Sedimentary

Volcanic



Sandstone

Figure A2. Reference sheet displaying the three rock types in the geological survey task.

Acknowledgments

This work was in part funded by a NASA Space Technology Research Fellowship under award 80NSSC17K0184.

References

[1]Fong, T., Rochlis Zumbado, J., Currie, N., Mishkin, A., and Akin, D. L., "Space telerobotics: unique challenges to human–robot collaboration in space," *Reviews of Human Factors and Ergonomics*, Vol. 9, No. 1, 2013, pp. 6–56.

[2]Fong, T., Nourbakhsh, I., Kunz, C., Flückiger, L., Schreiner, J., Ambrose, R., Burridge, R., Simmons, R., Hiatt, L. M., Schultz, A., et al., "The peer-to-peer human-robot interaction project," *In Proceedings of AIAA Space*, 2005.

[3]Cohen, P. R. and Levesque, H. J., "Teamwork," *Nous*, Vol. 25, No. 4, 1991, pp. 487–512.

[4]Fong, T. and Nourbakhsh, I., "Interaction challenges in human-robot space exploration," *Invited article for ACM Interactions special issue on human-robot interaction*, 2005.

[5]Fong, T., Scholtz, J., Shah, J. A., Fluckiger, L., Kunz, C., Lees, D., Schreiner, J., Siegel, M., Hiatt, L. M., Nourbakhsh, I., et al., "A preliminary study of peer-to-peer human-robot interaction," *In Proceedings of Systems, Man and Cybernetics (SMC)*, Vol. 4, IEEE, 2006, pp. 3198–3203.

[6]Fong, T., Kunz, C., Hiatt, L. M., and Bugajska, M., "The human-robot interaction operating system," *In Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, ACM, 2006, pp. 41–48.

[7]Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al., "The HCRC map task corpus," *Language and speech*, Vol. 34, No. 4, 1991, pp. 351–366.

[8]Eberhard, K. M., Nicholson, H., Kübler, S., Gundersen, S., and Scheutz, M., "The Indiana "Cooperative Remote Search Task"(CReST) Corpus." *In Proceedings of LREC*, 2010.

[9]Gervits, F., Eberhard, K., and Scheutz, M., "Team communication as a collaborative process," *Frontiers in Robotics and AI*, Vol. 3, 2016, pp. 62.

[10]Goodrich, M. A. and Schultz, A. C., "Human-robot interaction: a survey," *Foundations and trends in human-computer interaction*, Vol. 1, No. 3, 2007, pp. 203–275.

[11]Marge, M. R. and Rudnicky, A. I., "The teamtalk corpus: Route instructions in open spaces," *In Proceedings of the Workshop on Grounding Human-Robot Dialog for Spatial Tasks*, 2011.

[12]Marge, M., Bonial, C., Foots, A., Hayes, C., Henry, C., Pollard, K., Artstein, R., Voss, C., and Traum, D., "Exploring Variation of Natural Human Commands to a Robot in a Collaborative Navigation Task," *In Proceedings of the First Workshop on Language Grounding for Robotics*, 2017, pp. 58–66.

[13]Voss, C. R., Cassidy, T., and Summers-Stay, D., "Collaborative Exploration in Human-Robot Teams: Whats in Their Corpora of Dialog, Video, & LIDAR Messages?" *EACL 2014*, 2014, pp. 43.

[14]Green, A., Hüttenrauch, H., Topp, E. A., and Eklundh, K. S., "Developing a contextualized multimodal corpus for human-robot interaction," *In Proceedings of LREC*, 2006.

[15]Hoffman, G. and Breazeal, C., "Collaboration in human-robot teams," *In Proceedings of the AIAA 1st Intelligent Systems Technical Conference*, 2004.

[16]Gervits, F., Eberhard, K. M., and Scheutz, M., "Disfluent but effective? A quantitative study of disfluencies and conversational moves in team discourse." *In Proceedings of COLING*, 2016, pp. 3359–3369.

[17]Gosling, S. D., Rentfrow, P. J., and Swann, W. B., "A very brief measure of the Big-Five personality domains," *Journal of Research in personality*, Vol. 37, No. 6, 2003, pp. 504–528.

[18]Hart, S. G., "NASA-task load index (NASA-TLX); 20 years later," *In Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50, 2006, pp. 904–908.

[19]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.

[20]Schmidt, T., "The transcription system EXMARaLDA: An application of the annotation graph formalism as the Basis of a Database of Multilingual Spoken Discourse," *In Proceedings of the IRCS Workshop On Linguistic Databases*, 2001, pp. 219–227.

[21]Carletta, J., Isard, S., Doherty-Sneddon, G., Isard, A., Kowtko, J. C., and Anderson, A. H., "The reliability of a dialogue structure coding scheme," *Computational linguistics*, Vol. 23, No. 1, 1997, pp. 13–31.

[22]Lickley, R. J., "HCRC disfluency coding manual," *Technical Report HCRC/TR-100*, 1998.

[23]Santorini, B., "Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision)," *Technical Report MS-CIS-90-47*, 1990.

[24]Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J., "ToBI: A standard for labeling English prosody," *In Proceedings of the Second International Conference on Spoken Language Processing*, 1992.

[25]Carless, S. A. and De Paola, C., "The measurement of cohesion in work teams," *Small group research*, Vol. 31, No. 1, 2000, pp. 71–88.

[26]Monsell, S., "Task switching," *Trends in cognitive sciences*, Vol. 7, No. 3, 2003, pp. 134–140.