Predicting Individual Human Performance in Human-Robot Teaming

Jack Kolb[1], Mayank Kishore[1], Kenneth Shaw[2], Harish Ravichandar[1], and Sonia Chernova[1]

Abstract— **Coordinating human-robot teams requires careful planning and allocation of tasks to the most appropriate agents. This challenge is exacerbated by the fact that, unlike their robot teammates, humans exhibit significant variation in their abilities. Existing work largely ignores this variation in favor of simpler aggregate models, failing to leverage specialized capabilities of different individuals. In this work, we introduce simple cognitive tests for measuring inherent variations in human capabilities related to human-robot teaming, specifically, the ability to maintain situational awareness and to mentally model latent network structures. We then demonstrate that user study participant performance on these cognitive tests is correlated with, and thus is a predictor for, their performance on human-robot teaming tasks. These findings have the potential to improve human-robot teaming algorithms (e.g., task allocation) by providing a mechanism to better leverage individual differences in human agents.**

I. INTRODUCTION

Human-robot teaming (HRT) enables groups of humans and autonomous robots to communicate, coordinate, and collaborate to perform a joint activity. In human-robot teams, humans and robots contribute complementary capabilities, allowing the team to accomplish a variety of complex tasks. As a result, HRT has been studied across a wide range of domains, including search and rescue [1], defense [2], and space exploration [3].

Unlike a team of homogeneous robots, agents within a human-robot team are heterogeneous, and thus not interchangeable. This heterogeneity introduces the need for careful planning and allocation of agents to the various tasks that need to be carried out [4]. This challenge is exacerbated by the fact that human agents often significantly vary from one another in terms of their capabilities. Indeed, prior work has shown that humans varied up to 87.5% in two traits associated with active robot path planning [5].

However, prior work on task allocation in human-robot teams has largely ignored this variation in favor of simpler aggregate models (e.g., [6]). In particular, it is often assumed that all human agents within a given category (e.g., soldier, firefighter, rescuer) have approximately equivalent attributes and can therefore be assigned arbitrarily. Treating all human operators as identical fails to account for individualized differences in capabilities across operators. As a result, such aggregate models fail to take advantage of the full potential of certain individuals, harming team performance.

This work was supported by the Army Research Lab under Grant W911NF-17-2-0181 (DCIST CRA)

[1]College of Computing at the Georgia Institute of Technology, North Avenue, Atlanta, GA 30332, USA

[2]The Robotics Institute at Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA

In this work, we model natural variations in human operator capabilities, and then study whether these variations translate into differences in individual performance on HRT tasks. Specifically, we introduce two cognitive tests that measure human cognitive capabilities that are pertinent to interacting with and coordinating multiple robots. The first test evaluates a user's ability to maintain situational awareness, and the second test evaluates their ability to mentally model latent network structures. Both tests are easy to implement, necessitate little instruction, and require only a few minutes of user time. Given user performance on these cognitive tests, we then evaluate how well users perform two common multiagent coordination tasks – remotely teleoperating multiple robots in an object retrieval mission, and forming an ad-hoc network using multiple mobile robots. Our central research question is to determine whether individual performance on certain cognitive tests serves as a predictor for HRT performance. If so, this information can serve to improve heterogeneous task assignment frameworks, such as [6].

Our results demonstrate that cognitive tests can in fact predict operator performance on certain HRT tasks. Specifically, the network test predicts performance on both the ad-hoc network task and the remote teleoperation task, however the situational awareness test does not predict performance on either task. This finding indicates that targeted cognitive tests can be developed to quickly and effectively probe individual human abilities prior to task assignment.

In summary, we contribute i) identification of two cognitive capabilities – maintaining situation awareness and modeling network structure – that affect performance in complex tasks involving HRT, ii) the design of two novel cognitive tests that are simple to deploy and effectively capture variations in human ability with respect to these two factors, and iii) analysis of correlation between human operator performance on the cognitive tests and HRT tasks.

II. RELATED WORK

Related work in robotics has identified cognitive traits that are expected to influence performance in HRT tasks (e.g., [7], [8], [9], [5]). However, little prior work has sought to verify these expectations. In contrast, our work explicitly analyzes the relationship between an individual's specific cognitive traits and performance in HRT tasks.

Many studies in the literature focus on how induced changes in a human's cognitive traits affect performance. For instance, traits such as workload have been studied to show how increasing a user's workload in a HRT task affects their performance [10] – finding that both too low and too high of a workload leads to decreased performance.

Similar studies have identified other attributes that affect user performance, such as situational awareness [8], [7], [9], prior experience with related tasks [8], [7], understanding of the robot autonomy [8], [9], and ability to context switch between tasks [8]. However, no prior studies have attempted to find correlations between these cognitive traits and a human's performance in HRT tasks.

Another group of studies analyze how design elements of an environment or system affect a user's cognitive traits, such as how many robots a user can manage simultaneously ('fan-out ability') [11], [12], and how long a user takes to respond to tasks in the environment [13]. However, these studies are focused on improving task and interface design in an effort to improve the performance of an *average* human operator. As such, they do little to aid with modeling and assigning humans to HRT roles based on their individual differences.

Existing work in the literature has solely used cognitive traits as *explanatory* measures for a human's performance in HRT tasks. While such analyses are useful in designing HRT tasks to maximize performance, these works do not attempt to leverage individual human differences towards improving a human-robot team's performance. In contrast, our work explores a novel direction by viewing cognitive traits as *predictive* performance measures for HRT tasks.

III. STUDY OVERVIEW

Our objective is to develop a set of cognitive pretests that evaluate individual cognitive traits relevant to HRT, and analyze if those traits predict an individual's performance on HRT scenarios. To this end, we focus on two cognitive traits – *situational awareness* and *network inference*. While many other traits impact human performance, we avoid selecting traits that should affect a human's performance broadly across all HRT scenarios, as such traits will not help predict relative performance on specific HRT tasks. We consider two generic HRT task scenarios – *creating an ad-hoc robot network* and *controlling multiple robots for item retrieval*. These scenarios are representative of tasks that are commonly observed in domains such as search-and-rescue and defense.

In this work, we consider the following hypotheses:

- H1 Performance in the **situational awareness pretest** will correlate to performance in the **controlling multiple robots for item retrieval scenario**.
- H2 Performance in the **network inference pretest** will correlate to performance in the **creating an ad-hoc robot network scenario**.
- H3 Performance in the **situational awareness pretest** will **not** correlate to performance in the **creating an ad-hoc robot network scenario**.
- H4 Performance in the **network inference pretest** will **not** correlate to performance in the **controlling multiple robots for item retrieval scenario**.

To test our hypotheses, we conduct a two-way factorial within-subjects study where participants are exposed to one pretest and one human-robot task scenario. The pretests and scenarios return numerical scores for each user, allowing us to determine the correlation between each pretest and each scenario. A considerable correlation between a pretest and a scenario will indicate that the pretest can predict a participant's scenario performance.

Eighty participants were recruited using Amazon Mechanical Turk and were fully informed of the study. Each of the four study conditions (pretest-task pairing) was completed by twenty participants.

IV. COGNITIVE PRETESTS

To measure the operator's ability to maintain situation awareness and model underlying networks, we develop a browser-based pretest for each trait. The pretests are designed to have the following characteristics in order to keep pretests generalized and applicable to a variety of human-robot tasks:

- Each pretest is abstract and does not directly mimic a specific HRT task.
- Each pretest seeks to estimate a single human trait.
- Each pretest has high variability in user scores.

A. Situational Awareness Pretest

Situational awareness is a human's mental model of an environment. Situational awareness is often evaluated on Endsley's three-level model (Perception, Comprehension, and Projection) [14].

We expect that the effect of situational awareness on performance will vary depending the particulars of the scenario. Scenarios that require a human to be actively aware of multiple robots simultaneously should utilize the human's situational awareness ability more than scenarios where such active awareness is unnecessary.

Measuring situational awareness has been widely studied, and a number of metrics have been developed to quantify a user's situational awareness [15], [16], [17]. Overwhelmingly used is the Situation Awareness Global Assessment Technique (SAGAT), a test format where the user is periodically interrupted from a task and asked questions about the task's environment [17]. This format can test any levels of Endsley's situational awareness model, and can be applied to a wide range of task environments.

For this pretest, we develop an abstract task in which the SAGAT format is used to quantity a user's situational awareness. In our environment, shown in Fig. 1, the user watches 'packages' (represented by small shapes) be distributed through an abstracted warehouse network (represented by large shapes). Warehouses can only process packages of their own color and shape, and they forward incorrect packages to downstream warehouses. Warehouses with no downstream warehouses must store incorrect packages they receive. As warehouses have a limited capacity, once a warehouse has reached its limit, it can no longer accept more packages, removing itself from the network. Over time, these package buildups gradually break down the distribution network.

The participant must keep track of the capacity levels of fifteen warehouses in the network. To evaluate the participant's situational awareness, the warehouse simulation is run for 30 seconds, then is paused and hidden. The

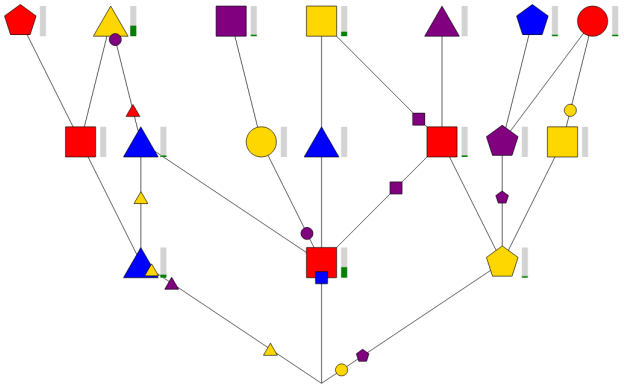

Fig. 1. Screenshot of the situational awareness pretest, showing the packages being distributed through the warehouse network. Packages start at the bottom and travel up through the network.

participant is asked to identify the capacity level of each warehouse as accurately as possible, or as 'uncertain'. The simulation then resumes and this process is repeated five times. The participant is scored by the accuracy of their labeling: $score = \#incorrect$, normalized to the maximum attainable score. Warehouses marked as 'uncertain' are not counted as incorrect.

This pretest focuses on capturing Level 1 and Level 2 of Endsley's situational awareness model. While the user is not directly asked comprehension questions about the warehouse network, as the network breaks down it becomes difficult to simply memorize the states of all fifteen warehouses. Users are therefore pushed to use their implicit comprehension of the network's congestion points and structure to accurately label the warehouse capacities.

B. Network Inference Pretest

Network inference is a human's ability to form mental models of a system's underlying network structure. Recent prior work in cognitive science has demonstrated that it is possible to model and predict a human's ability to learn abstract structures and relationships between a stochastic sequence of events [18]. Furthermore, a simple cognitive test was found to effectively quantify this ability. Many HRT tasks similarly require a user to mentally model complex structural information, such as communication networks, sensing capabilities, and relative influences of each agent on the swarm. We expect that a network inference pretest can be used to predict performance in HRT scenarios that heavily use underlying network structures.

In our preliminary work we implemented the network inference cognitive test from [18]. However, we found that test was challenging to apply in practice because it took up to 40 minutes of participant time. As a result we created a variant of the pretest, loosely inspired by the prior work.

Our pretest evaluates an individual's ability to determine the best starting node to efficiently propagate information across a hidden communication network. The pretest consists

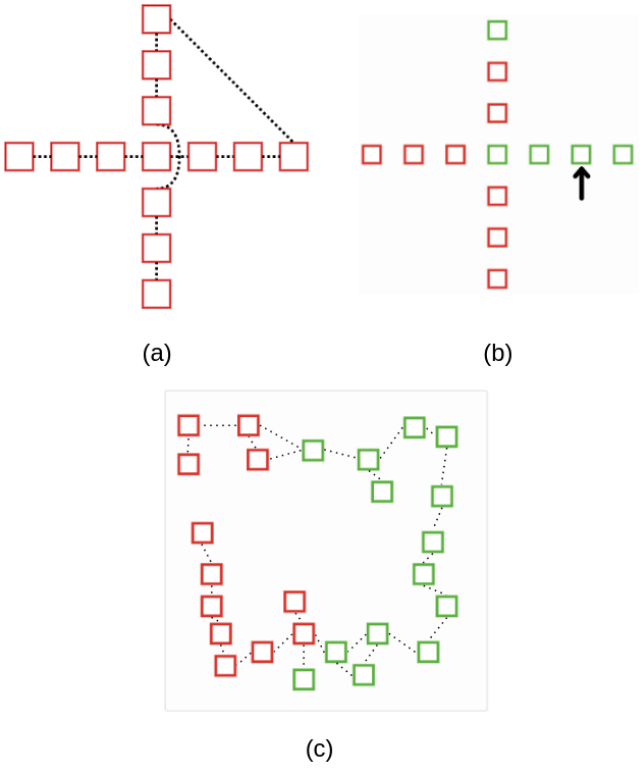

Fig. 2. Graphical overview of a network inference pretest stage. *(a)* shows the dotted edges that form the underlying node network, these edges are not shown to the user. *(b)* shows the network two time steps after the marked node was selected, with green color propagation. *(c)* shows a more complex network.

of two phases. In the first phase, the participant observes exemplar runs illustrating how information originating at various nodes propagates to the rest of the network. The propagation is shown by connected nodes changing color at each time step. In the second phase, the participant is asked to select the origin node such that information propagates to the rest of the network in the shortest amount of steps. This process is repeated for seven networks of differing complexity. Fig. 2 shows how the node networks are constructed and information is propagated. The underlying connectivity structure of the nodes (dotted lines in Fig. 2) is not shown to the user, thus they must learn the underlying structure of the graph, just as in a swarm interaction scenario the operator must maintain a mental model of robot connectivity, line of sight, or other latent features. The participant is scored by the total number of edges between their selected nodes and the closest correct node for each network, normalized to the maximum score attainable.

V. HUMAN-ROBOT TEAMING DOMAIN

To validate participant performance in a HRT scenario, we developed a simulation of a "search and retrieval" operation using the WeBots simulation environment [19]. In this domain, participants control unmanned aerial vehicles and unmanned ground vehicles to retrieve several supply caches hidden throughout a 3D environment. To independently validate different human skill sets, we split the operation

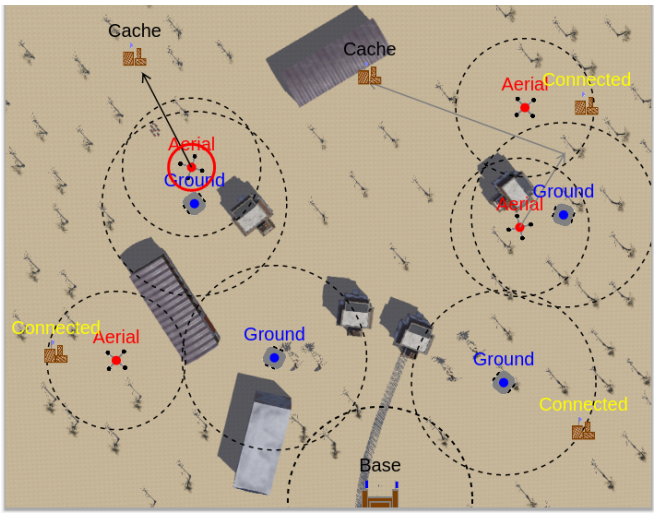

Fig. 3. The *creating an ad-hoc robot network* user interface. The user controls robots by clicking them and then clicking waypoints on the map for them to navigate to. The dashed circles around each robot indicate their network connection ranges. The user must extend the network from 'Base' to all five caches.

into two scenarios. These scenarios represent two different typical operator roles:

- **Creating an ad-hoc robot network**: Construct a communications relay network that extends to the caches.
- **Controlling multiple robots for item retrieval**: Leverage the relay network to retrieve the caches.

To reduce score variation due to learning effects, prior to starting their scenario participants are given a visual tutorial covering the contents of each scenario, and complete a set of navigation tasks in a simple demo world.

A. HRT Scenario: Creating an ad-hoc robot network

User Capability Being Assessed: Mental models of latent robot network; understanding how network coverage will change with robot placement.

Overview: This first scenario focuses on constructing an ad-hoc network, or communication relay, made up of a set of robots – a common task in multi-robot missions. The participant is given a complete overhead map of the environment, including the locations of the caches. The participant controls two robot species, aerial robots and ground robots. However, the robots have a limited communication range. To maintain control of a robot the operator must keep the robot connected to the base station, either directly or via a network of robots within each others communication ranges. The operator's objective is to arrange available robots in a spatial configuration that forms a relay network reaching all of the caches. Four aerial robots and four ground robots are used for this scenario – just enough robots to cover all five caches on the map – challenging the operator's ability to spatially arrange the robots.

The participant controls the robots by simply setting waypoints on a 2D overhead map. The robots autonomously travel along the waypoints to their destinations. By abstracting the robot control to a simple high level interface we aim to limit the performance variation between users due to their abilities to control robots. If an aerial or ground robot exceeds the boundaries of the relay network, it becomes disconnected and stops; the disconnected robot will remain out of contact until it is reconnected to the network. The participant can retrieve disconnected robots by moving other robots to reestablish the network. The participant receives no information about the robots other than their locations and orientations as shown on the 2D overhead map.

Solving this scenario requires the network to be built gradually from the robot base to the caches. While the aerial robot movements are not affected by obstacles in the environment, the ground robot movements are. As a trade off, the ground robots have a larger network relay range than the aerial robots due to their greater payload capacity. The user must therefore remain aware of the movements of the robots and the overall efficiency of the relay network. The scenario completes when the relay network extends to cover all five caches and the robot base, or 10 minutes have elapsed.

We expect that performance in this scenario can be predicted by network inference pretest scores, but not situational awareness pretest scores. Since creating and applying a mental model of the robot communication network is key to completing this scenario, it is reasonable to expect that network inference pretest scores would correlate to performance in this task. Alternatively, since this scenario is more focused on path planning than active awareness, we do not expect the situational awareness pretest scores to correlate to this scenario's outcomes.

Metrics: The participant is scored by the time it takes to complete the relay network. Participants who failed to complete the task in 10 minutes were discarded and marked as "x" in Fig. 6. A lower score is better.

B. HRT Scenario: Controlling multiple robots for item retrieval

User Capability Being Assessed: Situational awareness in the context of low-level robot control; context switching between multiple largely independent tasks.

Overview: The second scenario focuses on low level robot control with the objective of item retrieval. The operator is

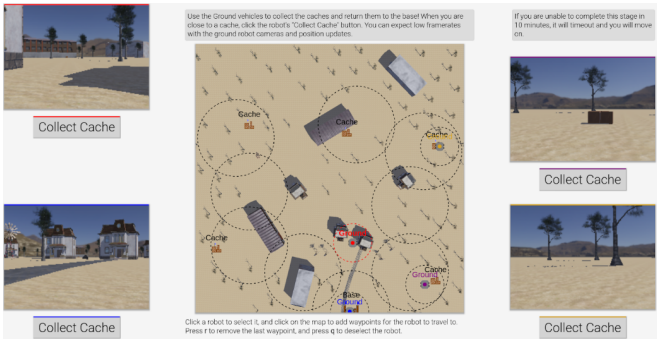

Fig. 4. The *controlling multiple robots for item retrieval* user interface. The user controls the ground robots by point-and-clicking waypoints on the center overhead map. Frontal camera feeds from each of the four robots are displayed around the overhead map.

tasked with retrieving each of the five caches and returning them to the robot base. To begin the scenario, the participant is given a full map of the environment labeled with cache locations, and a suitable relay network that extends to the caches. The participant will only control the retrieval robots and will not be able to change the network.

This scenario simulates a lower level of ground robot control. While in the *creating an ad-hoc robot network* scenario ground robots ignored small obstacles, in this scenario they will always travel directly to the next waypoint. As a result, the operator is required to closely supervise each robot and carefully arrange waypoints to avoid ground obstacles. The operator must also keep each robot within the boundaries of the relay network; a robot that exceeds the boundaries of the relay network will become disconnected, just as in the first scenario. When a robot approaches a cache, the user retrieves it by pressing a button, and then returns the robot to the base. The scenario ends when all caches have been returned or 10 minutes have elapsed. The user is given four ground robots to control.

Metrics: The participant is scored by the cumulative distance traveled by their robots, and the number of cache interactions (pick up and returns) the robots carried out, via the function $score = \frac{distance\_traveled}{\#interactions}$. Scores from participants with no cache interactions were discarded. A lower score is better.

VI. RESULTS

We conducted twenty trials for each of the four pretest-scenario pairings. Fig. 5 shows the distribution of participant scores for each pretest. No participants scored perfectly on either pretest, and both pretests show substantial variability via their interquartile ranges (IQR; situational awareness: .24, network inference: .15). These two attributes are important as they verify the utility of these pretests: if a pretest had many perfect scores, or if a pretest's variability was narrow, that pretest would be unable to meaningfully discern between participant abilities. As both pretests are grounded in prior work from cognitive science, we believe that they are succeeding in measuring their intended cognitive traits.

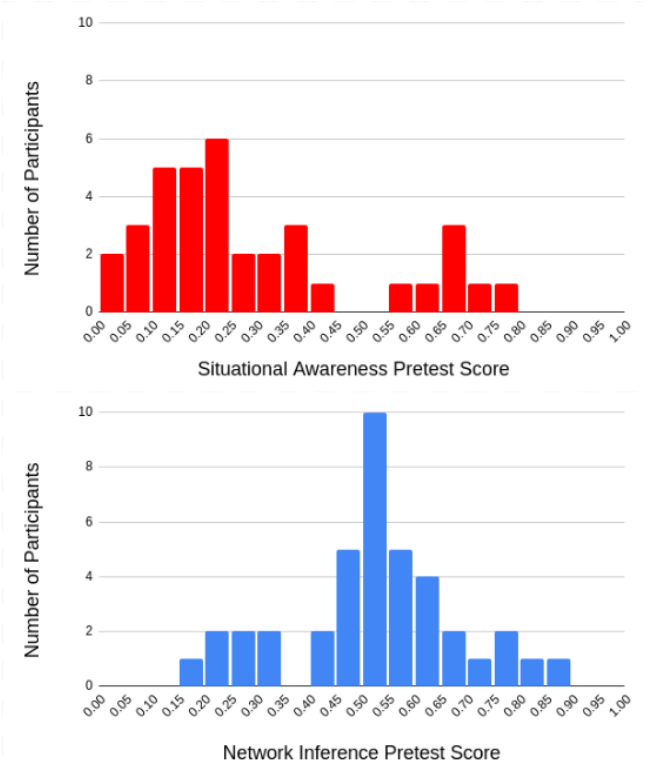

Fig. 5. Histogram of the situational awareness and network inference pretest scores.

Fig. 6 plots each pretest score relative to each scenario. Trend lines were fit to each plot, with correlation coefficients shown on the bottom right of each figure. Spearman's correlation coefficients were used because the score distributions for both HRT tasks are skewed.

The correlation coefficients and visual trends between the pretests and HRT scenarios supported two of our hypotheses. Specifically:

- H1 The *situational awareness* and *controlling multiple robots for item retrieval* pairing resulted in a weak correlation ($r_s = .149$) and relatively little score improvement. Thus the data rejects **H1**.
- H2 The *network inference* pretest appears to have a moderate correlation to the *creating an ad-hoc robot network* scenario ($r_s = .687$). Network inference pretest scores accounted for a roughly 3x performance difference in the scenario. Therefore the data supports **H2**.
- H3 The *situational awareness* pretest had a weak correlation to the *creating an ad-hoc robot network* scenario ($r_s = .292$). The pretest provided little indication of score improvement and there are no apparent trends in the data. Thus, there is evidence to support **H3**.
- H4 The *network inference* pretest had a moderate correlation to the *controlling multiple robots for item retrieval* scenario ($r_s = .324$). There also appears to be a positive trend in the relationship. Thus the data refutes **H4**.

VII. SUMMARY AND DISCUSSION

In summary, our work introduces cognitive pretests that measure human traits relating to situational awareness and mental models of latent network structures. Additionally, we developed two simulated HRT tasks that are common in real-world HRT scenarios – forming an ad-hoc network and teleoperation of multiple robots. Results from our user study demonstrate that our *network inference* pretest results in scores that meaningfully correspond to performance on two HRT tasks, while our *situational awareness* pretest does not meaningfully indicate HRT task performance on either task.

Importantly, we find that the cognitive pretests are not interchangeable. They do not simply measure cognitive ability in general, but instead capture innate abilities critical to specific operational roles. This is demonstrated by the fact that the relative slopes and correlation coefficients of each pretest vary by the HRT task applied to.

Our findings also provide further evidence that human operators have innate differences in their performances on

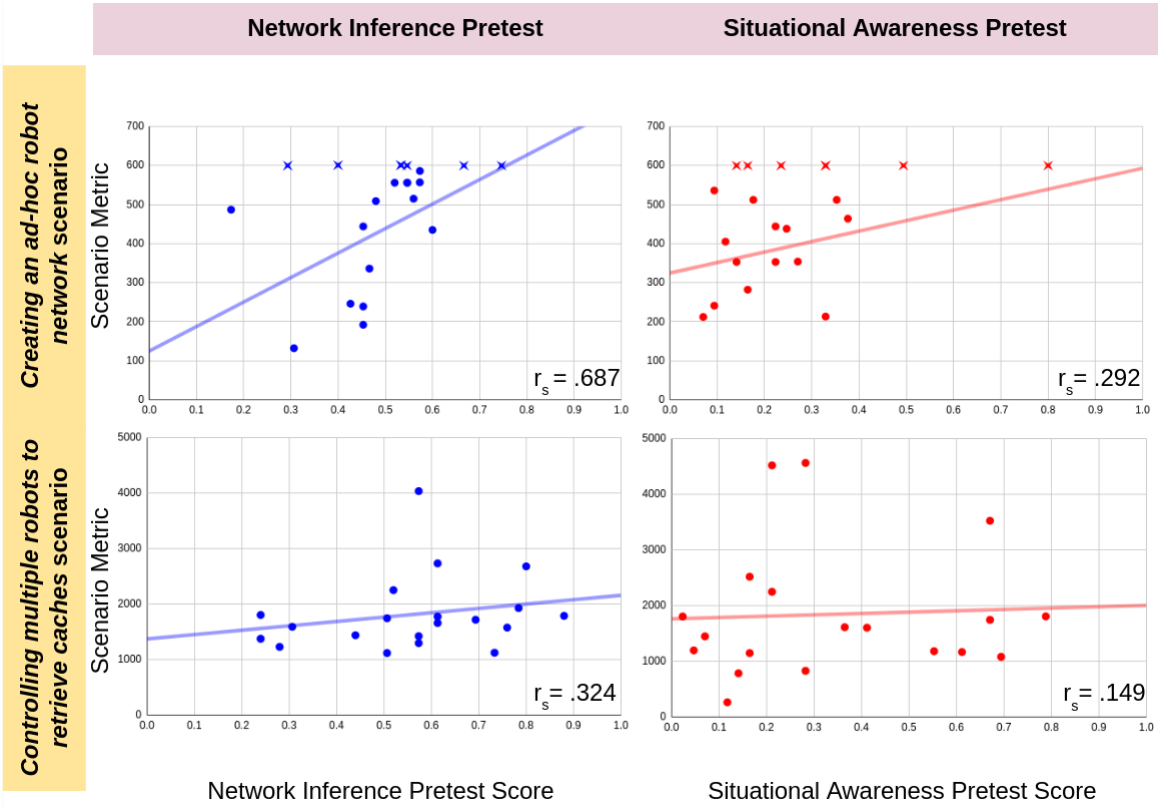

Fig. 6. Results for each pretest-scenario pairing. Lower scores are better for each metric.

HRT tasks. Predicting individual performance and making task assignment decisions accordingly is therefore critical for effective heterogeneous teaming. In the future, we envision simple, fast, and easy-to-deploy pretests such as ours being applied to improve the effectiveness of human-robot teaming.

REFERENCES

[1] S. Kohlbrecher, A. Romay, A. Stumpf, A. Gupta, O. Von Stryk, F. Bacim, D. A. Bowman, A. Goins, R. Balasubramanian, and D. C. Conner, "Human-robot teaming for rescue missions: Team vigir's approach to the 2013 darpa robotics challenge trials," *Journal of Field Robotics*, vol. 32, no. 3, pp. 352–377, 2015.

[2] R. Parasuraman, M. Barnes, K. Cosenzo, and S. Mulgund, "Adaptive automation for human-robot teaming in future command and control systems," 2007.

[3] T. Fong and I. Nourbakhsh, "Interaction challenges in human-robot space exploration," *Interactions*, vol. 12, no. 2, pp. 42–45, 2005.

[4] G. A. Korsah, A. Stentz, and M. B. Dias, "A comprehensive taxonomy for multi-robot task allocation," *The International Journal of Robotics Research*, vol. 32, no. 12, pp. 1495–1512, 2013.

[5] C. J. Shannon, D. C. Horney, K. F. Jackson, and J. P. How, "Human-autonomy teaming using flexible human performance models: An initial pilot study," pp. 211–224, 2017.

[6] H. Ravichandar, K. Shaw, and S. Chernova, "Strata: unified framework for task assignments in large teams of heterogeneous agents," *Autonomous Agents and Multi-Agent Systems*, vol. 34, pp. 1–25, 2020.

[7] S. Ponda, H.-L. Choi, and J. How, "Predictive planning for heterogeneous human-robot teams," in *AIAA Infotech@ Aerospace 2010*, 2010, p. 3349.

[8] J. Y. Chen and M. J. Barnes, "Human–agent teaming for multirobot control: A review of human factors issues," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 1, pp. 13–29, 2014.

[9] C. E. Harriott, A. E. Seiffert, S. T. Hayes, and J. A. Adams, "Biologically-inspired human-swarm interaction metrics," in *Proc. of the Human Factors and Ergonomics Society Annual Meeting*, vol. 58. SAGE Publications Sage CA: Los Angeles, CA, 2014, pp. 1471–1475.

[10] M. L. Cummings and C. E. Nehme, "Modeling the impact of workload in network centric supervisory control settings," in *2nd Annual Sustaining Performance Under Stress Symposium*, 2009.

[11] D. R. Olsen Jr and S. B. Wood, "Fan-out: Measuring human control of multiple robots," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2004, pp. 231–238.

[12] M. L. Cummings, C. E. Nehme, J. Crandall, and P. Mitchell, "Predicting operator capacity for supervisory control of multiple uavs," in *Innovations in Intelligent Machines-1*. Springer, 2007, pp. 11–37.

[13] P. Walker, S. Nunnally, M. Lewis, A. Kolling, N. Chakraborty, and K. Sycara, "Neglect benevolence in human control of swarms in the presence of latency," in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2012, pp. 3009–3014.

[14] M. Endsley, "Endsley, m.r.: Toward a theory of situation awareness in dynamic systems. human factors journal 37(1), 32-64," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 37, pp. 32–64, 03 1995.

[15] P. M. Salmon, N. A. Stanton, G. H. Walker, D. Jenkins, D. Ladva, L. Rafferty, and M. Young, "Measuring situation awareness in complex systems: Comparison of measures study," *International Journal of Industrial Ergonomics*, vol. 39, no. 3, pp. 490–500, 2009.

[16] L. Paletta, A. Dini, C. Murko, S. Yahyanejad, M. Schwarz, G. Lodron, S. Ladstätter, G. Paar, and R. Velik, "Towards real-time probabilistic evaluation of situation awareness from human gaze in human-robot interaction," in *Proc. of Human-Robot Interaction*, 2017, pp. 247–248.

[17] M. R. Endsley, "Situation awareness global assessment technique (sagat)," in *Proc. of IEEE national aerospace and electronics conference*. IEEE, 1988, pp. 789–795.

[18] C. W. Lynn, A. E. Kahn, N. Nyema, and D. S. Bassett, "Abstract representations of events arise from mental errors in learning and memory," *Nature communications*, vol. 11, no. 1, pp. 1–12, 2020.

[19] O. Michel, "Cyberbotics ltd. webots™: professional mobile robot simulation," *International Journal of Advanced Robotic Systems*, vol. 1, no. 1, p. 5, 2004.