A Virtual Testbed for the Multidisciplinary Evaluation of Human-Agent Teaming Dynamics

Audrey L. Aldridge[1 [0000−0003−3733−4736]], Christopher Hudson[1 [0000−0002−6216−8234]], Karl Smink[1 [0009−0005−6750−1162]], Andrew R. Buck[2 [0000−0002−8892−3269]], Derek T. Anderson[2 [0000−0001−5888−3617]], Victor Paul[3], Rachel Anderson[3], Drew Hoelscher[3], Mary Quinn[4], Matus Pleva[5 [0000−0003−4380−0801]], Cindy L. Bethel[1 [0000−0001−9036−3275]], and Daniel W. Carruth[1 [0000−0003−0707−9252]]

[1] Mississippi State University, Mississippi State, MS 39762, USA
ala214@msstate.edu, chudson@cavs.msstate.edu, ks2925@msstate.edu, cbethel@cse.msstate.edu, dwc2@cavs.msstate.edu
[2] University of Missouri, Columbia, MO 65211, USA
buckar@missouri.edu, andersondt@missouri.edu
[3] U.S. Army DEVCOM-Ground Vehicle Systems Center, Warren MI 48397, USA
victor.j.paul2.civ@army.mil, rachel.e.anderson46.civ@army.mil, andrew.l.hoelscher.civ@army.mil
[4] Leidos, Inc., Chantilly, VA 20151, USA
mary.m.quinn@leidos.com
[5] Technical University of Košice, Košice 042 00, Slovak Republic
matus.pleva@tuke.sk

Abstract. While intelligent agents offer substantial potential when integrated into human teams, challenges persist in achieving effective collaboration and information sharing with these systems. A virtual testbed of modular design was developed to enable rapid implementation of different teaming behaviors, comprehensive observation and measurement of individual and team performance, and flexibility of team composition, task objectives, and environment configurations. Using the testbed, an initial study demonstrated that in a non-hierarchical collaborative task, sharing information among teammates improved task performance and the ease with which to work with autonomous teammates, while reducing participants' mental workload, as measured by the NASA-TLX survey.

Keywords: human autonomy teaming · robotics · mental workload.

1 Introduction

Robots, artificial intelligence (AI), and autonomous agents have the potential to increase efficiency and reduce hazardous risks for humans when integrated into teams. As experts explore incorporating multiple intelligent systems into human teams, they recognize the significant capabilities these multi-agent teams can offer. However, challenges remain in effectively communicating and sharing

information with these systems and in collaborating with them in ways comparable to interacting with other humans. These can include discrepancies in task understanding or a team's limited agility in dynamic environments.

Information sharing within a team plays a pivotal role in shaping team dynamics and fostering effective collaboration, as poor communication can result in misaligned goals and hinder team performance. Conversely, effective information exchange can promote seamless collaboration and enhance efficiency and performance in human-agent teams. As autonomous agents are increasingly integrated into teams, it is crucial to understand how information sharing affects interactions between humans and autonomous agents as well as how it supports the development of a common or shared understanding (i.e., shared mental model (SMM)) of task and team functions).

This paper investigates teaming dynamics through a study using a novel configurable virtual testbed. In the testbed, a human and two autonomous agents collaborate on a maze-based search task, working together to find keys and unlock doors. Although the human could technically complete the task independently, collaboration was expected to enhance efficiency and effectiveness. The autonomous agents dynamically assessed trust in their teammates based on prior interactions and observed task performance. The study examined how three levels of information availability (no personal or shared information, personal but no shared information, and shared information) influenced team dynamics, task performance, and participants' mental workload, offering insights into the critical role of information sharing in human-agent collaboration.

2 Background and Related Work

To investigate the effects of differing levels of information sharing on human-agent teams, a virtual modular testbed tailored for flexibility and adaptability was developed. The virtual testbed supports rapid implementation of diverse teaming behaviors, enables unrestricted observation and measurement of individual and team performance, and allows extensive modifications to team composition, task objectives, and the environment. It also facilitates research into the inter-relational properties of human-agent teaming. A review of existing human-agent testing environments revealed significant gaps: Few testbeds are designed to support the type of collaborative, pre-planned coordination expected in real-world operations. Many focus solely on autonomy or multi-agent teaming, often excluding human interaction. Current testbeds typically feature either a single human or agent and enforce rigid, hierarchical team roles where humans always lead [2]. Other testbeds for human-agent teaming studies were designed to test communication (explanations) for coordination tasks [4], human factors including trust in autonomous aerial vehicles for air combat [8, 13], communication with AI versus with a human in interdependent tasks [13], and potential measures for effective work [7].

As the potential for humans and intelligent agents to work together rapidly grows, it becomes crucial for researchers to address the challenges of human-

agent teaming. Rather than focusing on one aspect, such as autonomous navigation or human situation awareness, the virtual testbed envisioned in this work supports collaboration and promotes interaction from all teammates, creating an environment where multiple facets of human-agent teaming can be studied together.

3 Creation of A Virtual Testbed for Conducting Human-Agent Teaming Studies

The testbed was developed using Unreal Engine (UE 5.1.1) to investigate teaming dynamics between humans and autonomous agents in a collaborative, non-hierarchical, multi-agent task. The testbed is based on a simplified version of a complex search-and-rescue scenario. In the simplified version, a human and two autonomous agents navigate a maze to find a target. The maze consists of strategically placed walls and doors. Some doors must be opened to reach the goal while others are optional. To navigate through the maze, the agents must seek out unique keys which unlock corresponding doors. To complete the task, the team must locate and reach the target within eight minutes. Efficiently completing the task requires the human participant work with two virtual robotic teammates. The human should collaborate and rely on them to find and swap keys as necessary to open doors. In addition to the interactions with the human teammate, the two virtual robots may collaborate with each other. A human-agent interface with a visual display, or 'minimap', of the environment is included in the virtual testbed. This interface communicates visual information regarding the status of the task and the environment. It also provides mechanisms for collaborating with teammates to find and retrieve keys.

By abstracting the task, the environment and task complexity are reduced, minimizing potential confounding factors and focusing on the elements of primary interest to the research. As such, assumptions were imposed on the task space. The target in the environment (an injured person) is assumed to be stationary. The environment conditions are dynamic in that keys move locations as they are picked up and dropped; therefore, in this study, participants may lose track of the keys when they are moved by a teammate. As more information is shared through the human-agent interface in each study condition, participants gain awareness of dynamic key locations. Finally, although the environments are sometimes unknown (key-door assignments, key locations) and partially dynamic (key locations), it was assumed that there would be no unexpected events that would critically derail the teams.

3.1 Search Task in the Maze Environment

The simplified search task required the human-robot teams to locate and reach a target (an injured person, represented by a trophy object). Participants maneuvered through the environment in first-person view (Fig. 1) using the 'w',

'a', 's', 'd' keys (commonly used in computer games) or the arrow keys. Additionally, participants had a birds-eye-view of the entire maze (referred to as the minimap) in a human-agent interface, as seen on the right of Fig. 1. The environment contains 25 unique key-door pairs, where each key unlocks a specific door. Teammates can hold up to two keys at a time, and once a door is unlocked, both the door and key are removed from the environment. Collaboration between participants and the virtual robots is essential to finding and swapping keys to efficiently progress through the maze. A team is considered successful when the human teammate reaches the target. If the virtual robots locate the target, they mark its location on the minimap visible through the human-agent interface.

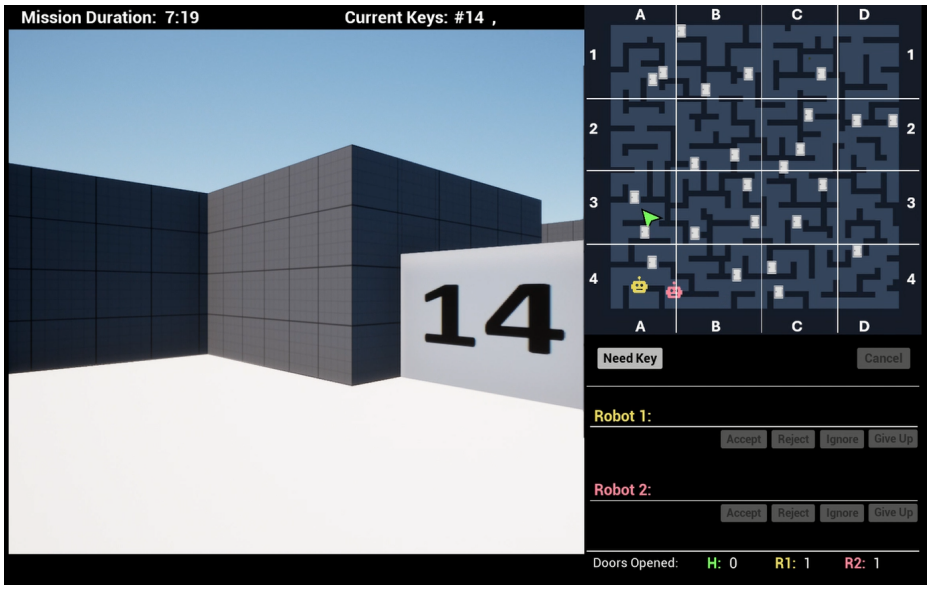

Fig. 1: First-person view of the maze environment showing Door 14 (left) with the human-agent interface (right).

3.2 Communication

Collaboration requires some form of communication between agents. In the testbed, a simulated communication channel allows the human and autonomous agents to share information. The communication channel is housed in the human-agent interface. It includes message bars where requests and responses are displayed and buttons for requesting and answering requests. In this communication channel, a human user can use the buttons on the interface to request help finding a key when standing next to the corresponding door. Each request stays

active in the interface for a maximum of 10 seconds or until one or both teammates respond. If the request times out, it is removed from the interface. Each teammate can only have one request posted at a time. Additionally, the human user can cancel a request or give up on the robots' requests for help finding a key.

3.3 Agent Autonomy and Navigation

The virtual robots operate autonomously without interference or override control from the human. UE5's built-in pathfinding algorithm, which is based on the A* algorithm, was used to compute an optimal path for the agents to reach their next destination. Zones were placed throughout the environment to act as waypoints to which the robots would travel. The robots may execute four action behaviors. They can *Explore* the maze environment to find the target. This behavior is based on Random Search, where a point on a map is "randomly" selected to which a robot navigates using the autonomous navigation behavior. Random Search was modified to prioritize the less traveled zones, eliminating the potential for the robots to ping back and forth "randomly" between two areas. The modified Random Search also ensures that the robots traverse as much of the environment as is accessible via unlocked doors.

At any time a starting location and end location are known, such as the location of a key, door, or teammate, the robots perform the *Search* behavior. *Search* uses UE5's A*-like pathfinding algorithm to move to the known object. This behavior is a directed search with a specific object as the end-goal. When a robot possesses a key that another teammate has requested, the robot exhibits the *Deliver* behavior. In the *Deliver* behavior, the robot begins to move to the teammate's location with the intent of delivering the key. Teammates could wait at their current location or also begin moving toward the robot to retrieve the key. For this navigation, the robots use UE5's A*-like pathfinding algorithm rather than the modified navigation which prioritizes zones. Lastly, the robots sometimes exhibit the *Wait* behavior, which was created to help the robots break free of a deadlock situation. In this behavior, the robot idles for three to five seconds before resetting its behaviors and starting a new action behavior.

3.4 Agent Protocols and Collaboration Behaviors

In addition to interacting with the human teammate, the two virtual robots collaborate with and rely on each other to find and swap keys to open doors. To do so, the robots can *Request* help to find a key and can respond to teammates' requests with an *Accept*, *Reject*, or *Ignore* message. Each of these behaviors is tied to a set of trust behaviors that help the robots determine how to respond to teammates' requests, ideally giving the sense of a cooperative and reliable teammate. The robots have some 'understanding' of what it means to be reliable teammates. Four trust behaviors dictate how a robot responds to teammates' requests: Implicit (blind trust), Untrust (uncertain trust), Distrust (lack of trust),

and Mistrust (misplaced trust). In this study, Mistrust is used as a state for rebuilding trust. These four levels of trust can fluctuate during a simulation based on a set of reciprocal cooperation rules and the resulting grudge that is formed when teammates do not help each other.

4 Human-Agent Teaming Study

This study uses the virtual testbed to investigate how team performance and participants' mental workload are impacted by information availability. This was a within-subjects study, where counterbalanced randomization of the conditions determined the order in which participants were exposed to the different levels of information availability during the simulations. Based on an a priori power analysis for a repeated measures within factors study design using an effect size of 0.25, a significance level (α) of 0.05, and a power of 80%, the total sample size was calculated to be n = 55. To account for unexpected issues and circumstances, the target sample size was set to 60. Participants were recruited from Mississippi State University's Psychology Research Program and received either a $20 gift card or 2 research credits for their psychology class as compensation for their time. The study protocols (IRB-24-223) were reviewed and approved as Exempt by Mississippi State University's Human Research Protection Program (HRPP) and Institutional Review Board (IRB).

4.1 Conditions

Three study conditions defined the levels of available information. All conditions contained a minimap that displayed walls, door placement, and agent locations. Agent locations were always updated in real-time during the simulations.

In **Condition 1**, known as the "No Info" condition, the minimap contained minimal information and did not communicate the identity of the doors, key locations, or door-key pairings. This meant participants had to remember where they saw specific keys and doors. To create an equivalent cognitive state in the robots, they were only capable of storing the locations of two doors and two keys. Fig. 2 displays the minimap of available information from each teammate's perspective for Condition 1. This condition is similar to many state-of-the-art devices with GPS tracking.

Condition 2, referred to as "Personal Info", contained the same information as in Condition 1. However, the minimap visually represented each teammate's individual mental model of the environment. Fig. 3 displays the minimap of available information from each teammate's perspective for Condition 2 to demonstrate how each teammate can only access information based on their experience in the environment. This includes the locations of teammates, doors, and keys as well as door-key matches, indicated by the matching color and number. In this condition, the minimap used Fog of War [11] to show areas that had not been searched by the participant. As the participant moved through the environment, the "fog" unveiled doors, keys, and their unique identifiers as these objects

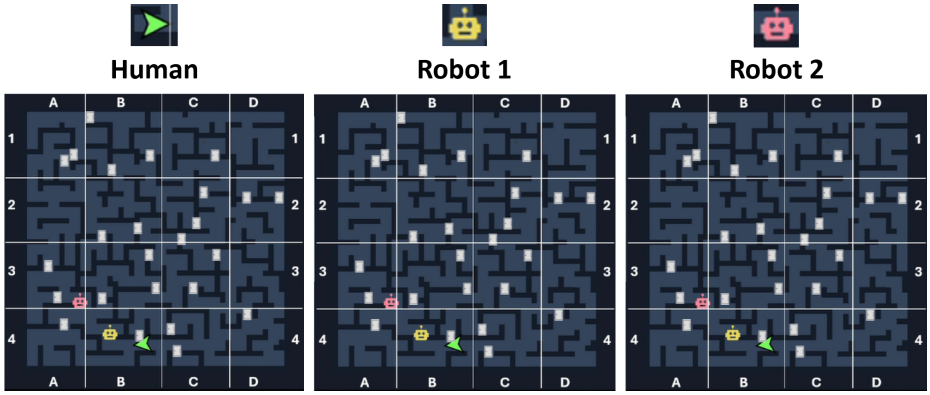

Fig. 2: Each teammate's perspective of the information available on the minimap in Condition 1.

were seen from the participant's first-person view. The robots had their own version of a memory model they could reference with the same information as the minimap for the participant but without the visual display. No information was shared between teammates, other than teammate locations. This meant that if a teammate saw a key and marked its location in their personal mental model (or minimap), their mental model or understanding of that key's location did not get updated when the key was picked up by a fellow teammate. This condition represents teammates operating from their own (personal) mental model that is not shared with other teammates.

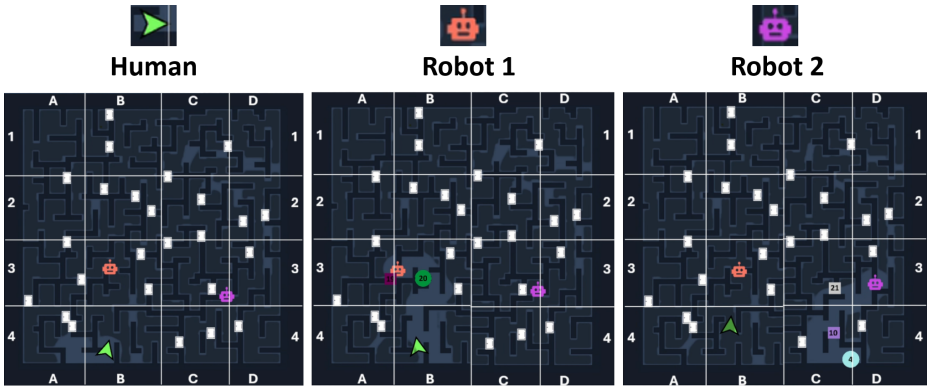

Fig. 3: Each teammate's perspective of the information available on the minimap in Condition 2.

In **Condition 3**, referred to as the "Shared Info" condition, the minimap contained the combined information stored across all three teammates' personal mental models to reflect a team's common understanding or SMM. Once again, this included the locations of teammates, doors, and keys as well as door-key pairs, indicated by the matching color and number, but as they were uncovered by each teammate. This means that as teammates moved through areas previously searched by another teammate (displayed as areas cleared of fog), the shared minimap reflected the current state of that part of the maze. Specifically, all teammates had access to the information uncovered by every other teammate. Fig. 4 displays the minimap of available information from each teammate's perspective for Condition 3 to demonstrate that everyone had access to the same information. This condition represents a team operating from their common understanding (or SMM) of a task and environment.

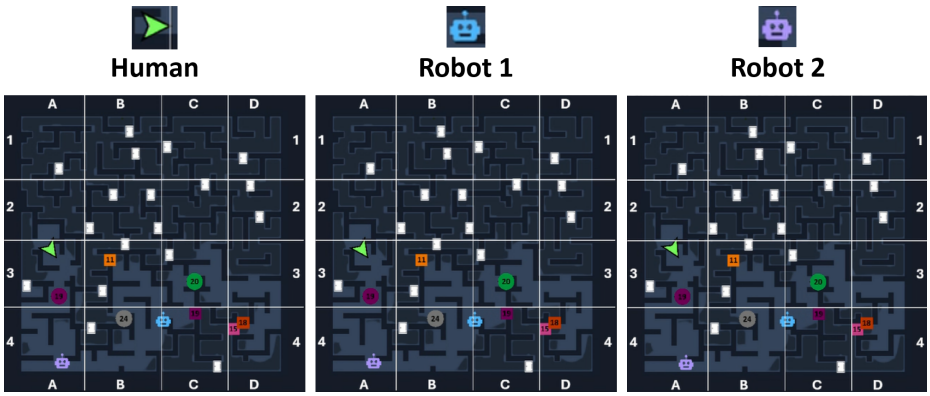

Fig. 4: Each teammate's perspective of the information available on the minimap in Condition 3.

4.2 Participants

A total of 61 students, ages ranging from 18 to 33 with an average age of 18.69, participated in the study. One participant withdrew from the study due to difficulty using the keys on the keyboard. Three participants were excluded from the data analysis due to 1) failure to follow instructions or 2) refusal to answer survey questions. Of the remaining participants, two reported having experience working on AI and autonomous systems. Two participants reported having worked with simulated environments. Fifteen participants reported playing more than three hours of computer/video games per week, and five of those reported playing five or more hours of computer/video games per week.

4.3 Study Protocol

At the start of a session, participants reviewed the task description and an informed consent document, with the option to ask questions before deciding to participate. After consenting, participants watched an introductory video and completed a 15-minute tutorial, which familiarized them with the search task, environment features, keyboard navigation, task mechanics, the human-agent interface, and the think-aloud method, where thoughts are verbalized while performing a task. The think-aloud method was included in the study to mimic the radio communication that might happen during a mission. Before starting each search task, participants viewed a brief video explaining the assigned study condition. Participants were given eight minutes to complete a search task in the simulated environment, after which they answered the Workload Profile and NASA-TLX surveys to measure their perceived mental workload. After answering survey questions, participants could take a short break before beginning the next task. At the end of the session, participants completed demographic questions and provided feedback on their experience with video games, autonomous vehicles, and AI agents. After a debriefing and final questions about their experience participating in the study, the session concluded. The total time commitment for the study was approximately 90 minutes.

4.4 Assessments

Overall team performance on the search task was evaluated based on success and duration. Due to the complexity of the maze and the difficult time constraint, secondary factors including the number of critical keys found and the number of doors opened were used in addition to assess performance. Although the main task was not to open as many of the doors as possible, the total number of doors opened by a team was included to check if participants became overly focused on opening all the doors instead of strategically choosing doors to open. Mental workload of the overall search task was measured using the Workload Profile [10] and the NASA Task Load Index (NASA-TLX) [6]. Additionally, the Workload Profile evaluated participants' mental workload regarding four sub-tasks. These included working with the robot teammates, understanding the information on the minimap, using the information on the minimap, and verbally answering questions. In rating the proportional resource demands of a task, "0" equates to no attentional demand, while "1" equates to maximum (or full) attentional demand [10, 14]. To score overall workload for each task, the ratings along the different dimensions were summed. To evaluate workload using the NASA-TLX, the raw workload scores were analyzed individually and averaged to yield a total workload score [5, 1].

5 Data Analysis and Results

A mixed-methods analysis of the results was conducted to determine how three conditions of information availability (C1 = No Information, C2 = Personal Information, C3 = Shared Information) impacted performance (duration, critical

keys found, number of doors opened) and mental workload (WP and NASA-TLX). The Shapiro-Wilk test concluded that duration, task success, number of critical keys found, the NASA-TLX, and the Workload Profile were not normally distributed, whereas the number of doors opened was normally distributed. A repeated measures univariate analysis of variance (ANOVA) was used to analyze the number of doors opened, with partial eta-squared determining the effect size and the Tukey HSD test comparing the conditions pairwise to determine which conditions were statistically significantly different. For the nonparametric data, the Friedman test, a nonparametric alternative to the repeated measures ANOVA, was used to evaluate the differences between conditions for each dependent variable. Kendall's W coefficient was used to determine effect size for significant results from the Friedman Test. Based on Cohen's interpretation of effect size [12, 3], Kendall's W coefficient is considered a small effect from 0.1 to < 0.3, moderate effect from 0.3 to < 0.5, and a large effect if $>= 0.5$. Pairwise testing, using the Wilcoxon signed rank test with Bonferroni's correction, was performed to determine which conditions were statistically significant for the nonparametric data.

5.1 Individual and Team Performance

Due to the small portion of participants who successfully found the target, additional measures of performance were included in the data analysis. As such, task performance is broken down into individual and team performance. Team performance consists of duration, success rate, critical keys found, and number of doors opened. Individual performance is described by the number of doors opened by a teammate. This measure was used to determine if participants strayed from focusing on the overall task: finding the target. Table 1 displays the average values for team performance measures across the three conditions (C1 = No Info, C2 = Personal Info, C3 = Shared Info).

Table 1: Mean values for team performance measures per information availability condition. Best scores shown in bold and enclosed in a rectangle.

Condition	**Success Rate**	**Duration (min:sec)**	**Critical Keys Found (3 total)**	**Doors Opened (25 total)**
No Info (C1)	3.51%	8:00	0.75	10.14
Personal Info (C2)	1.75%	7:59	1.88	**13.95**
Shared Info (C3)	**38.60%**	**7:11**	**2.21**	10.54

Task Success: Table 1 displays the rate at which participants successfully completed the maze in each condition. The Shapiro-Wilk test concluded a non-normal data distribution. The Friedman rank sum test found a statistically significant difference in task success for the three information availability conditions, X2(2) = 36.61, $P < 0.0001$, with a moderate effect (W = 0.32). Performing the Wilcoxon signed rank test revealed a statistically significant difference in the success rate of participants between Conditions 1 and 3 ($P < 0.0001$) and Conditions 2 and 3 ($P < 0.0001$).

Task Duration: Task duration was measured as the number of minutes taken for a participant to reach the target in the maze environment, with a maximum of eight minutes (min) allowed. From the figure, it seems that Condition 1 and Condition 2 required nearly all participants to use the full eight minutes to search the maze for the target. Alternatively, Condition 3 saw a significant drop in task duration from Conditions 1 and 2, even though the median value for Condition 3 remained at eight minutes. The result of the Friedman rank sum test indicated a statistically significantly difference in duration across the three information availability conditions, X2(2) = 39.13, $P < 0.0001$, with a moderate effect size (W = 0.34). Using the pairwise Wilcoxon signed rank test between conditions revealed statistically significant differences in task duration between C1 and C3 ($P < 0.0001$) and C2 and C3 ($P = 0.0001$).

Number of Critical Keys Found: Like duration, the data for the number of critical keys found could not be normalized. The results in Table 1 reveal that participants found the fewest critical keys in Condition 1, with an increase in critical keys found in Condition 2 and a further increase in critical keys found in Condition 3. The result of the Friedman rank sum test indicated a statistically significant difference in the number of critical keys found across the three information availability conditions, X2(2) = 67.88, $P < 0.0001$, with a large effect size (W = 0.60). Using the pairwise Wilcoxon signed rank test between conditions revealed statistically significant differences in the number of critical keys found between C1 and C2 ($P < 0.0001$), C1 and C3 ($P < 0.0001$), and C2 and C3 ($P = 0.02$). Fig. 5 shows a violin plot of the number of critical keys found for the three conditions (C1 = No Info, C2 = Personal Info, C3 = Shared Info).

Number of Doors Opened: On average, teams opened three more doors in Condition 2 than they opened in Conditions 1 or 3. Because the Shapiro-Wilk test confirmed normal distribution for the number of doors opened, Fisher's repeated measures one-way ANOVA revealed that the number of doors opened was statistically significantly different across conditions, F(2, 112) = 23.02, $P < 0.0001$, $\eta_p^2 2 = 0.29$ (large effect). From the Tukey HSD test, pairwise differences between C1 and C2 ($P < 0.0001$) and C2 and C3 ($P < 0.0001$) are statistically significant.

For the individual performance results, i.e. the number of doors opened by each teammate, participants opened the most doors in Condition 3 (3.91), fol-

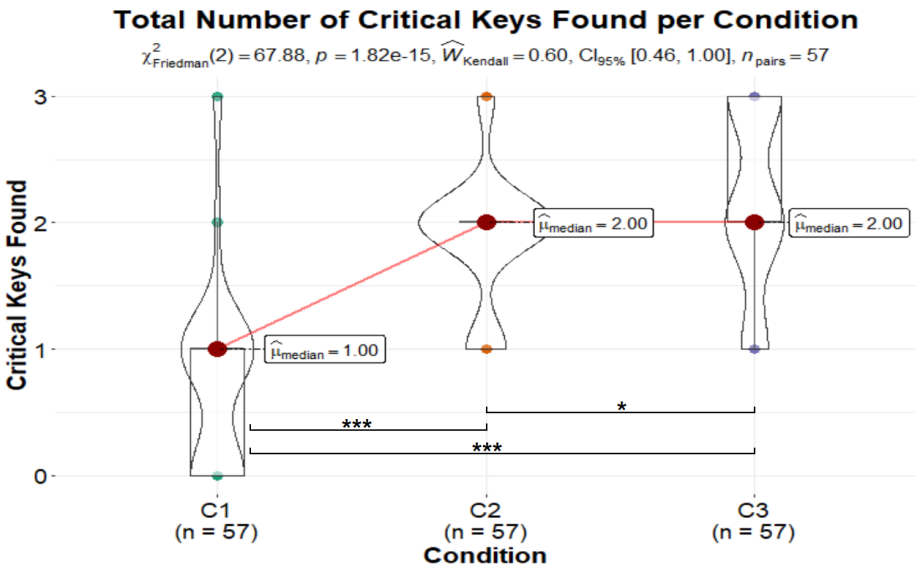

Fig. 5: Number of critical keys found across the three information availability conditions (C1, C2, C3). *P*-value scale: * (<0.05), ** (<0.01), *** (<0.001).

lowed by Condition 1 (3.53) and Condition 2 (3.20), respectively. Robot 1 opened the most doors in Condition 2 (6.36), followed by Condition 3 (3.57) then Condition 1 (3.13). Additionally, Robot 2 opened the most doors in Condition 2 (4.39), followed by Condition 1 (3.48) then Condition 3 (3.05). With these values all being relatively similar across the conditions, it can be deduced that participants did not get carried away with opening as many doors as possible as a result of having more information available.

5.2 Mental Workload

Both the Workload Profile and NASA-TLX were administered after participants completed each condition. The mean values for mental workload, as measured by the Workload Profile, are C1 = 4.03, C2 = 3.95, and **C3 = 3.75**. The mean values for mental workload, as measured by the NASA-TLX, are C1 = 53.80, C2 = 46.37, and **C3 = 39.91**. The values in bold reflect the best scores for each measure.

Workload Profile: In addition to the overall search task (OST), the Workload Profile was used to evaluate subjective mental workload for the following four sub-tasks: (ST1) working with robot teammates, (ST2) understanding information from the minimap, (ST3) using information on the minimap, and (ST4) verbally answering questions. Table 2 displays the mental workload scores for

the overall search task and each of the sub-tasks, as measured by the Workload Profile.

Table 2: Average mental workload scores, measured by the Workload Profile, for each condition. Best (lowest) scores for each task are shown in bold. Most demanding sub-task (ST) for each condition is enclosed in a rectangle.

Condition	**OST**	**ST1**	**ST2**	**ST3**	**ST4**
No Info (C1)	4.03	3.24	**3.01**	**3.17**	3.53
Personal Info (C2)	3.95	2.88	3.13	3.26	**3.25**
Shared Info (C3)	**3.75**	**2.79**	3.10	3.18	3.28

Note: **OST** = overall search task, **ST1** = working with robot teammates, **ST2** = understanding information from the minimap, **ST3** = using information on the minimap, **ST4** = verbally answering questions.

(OST) Overall Search Task: For the overall search task (Table 2), participants reported that searching the maze for the target in Condition 1 required the most amount of mental effort (4.03), compared to Condition 2 (3.95) and Condition 3 (3.75). Additionally, the results in Table 2 indicate that the overall search task required the most mental effort of all tasks in each information availability condition. The Friedman rank sum test did not indicate a statistically significant difference in the perceived mental effort required for the overall search task across the three conditions.

(ST1) Working with Robot Teammates: When working with robot teammates, the Workload Profile results (Table 2) show that Condition 1 required the highest average mental effort (C1 = 3.24), followed by Condition 2 (C2 = 2.88) and then Condition 3 (C3 = 2.79). The Friedman rank sum test indicated a statistically significant difference in the perceived mental effort required for working with robot teammates across the three information availability conditions, X2(2) = 10.20, $P = 0.006$, W = 0.09 (small effect). The pairwise Wilcoxon signed rank test found statistically significant differences between C1 and C2 ($P < 0.01$) and between C1 and C3 ($P = 0.01$). Fig. 6 shows a violin plot of the mental workload for Sub-Task 1 (ST1) 'working with robot teammates' across the three conditions.

(ST2) Understanding Information on the Minimap: The results from the Workload Profile in Table 2 show that the average mental effort required to understand

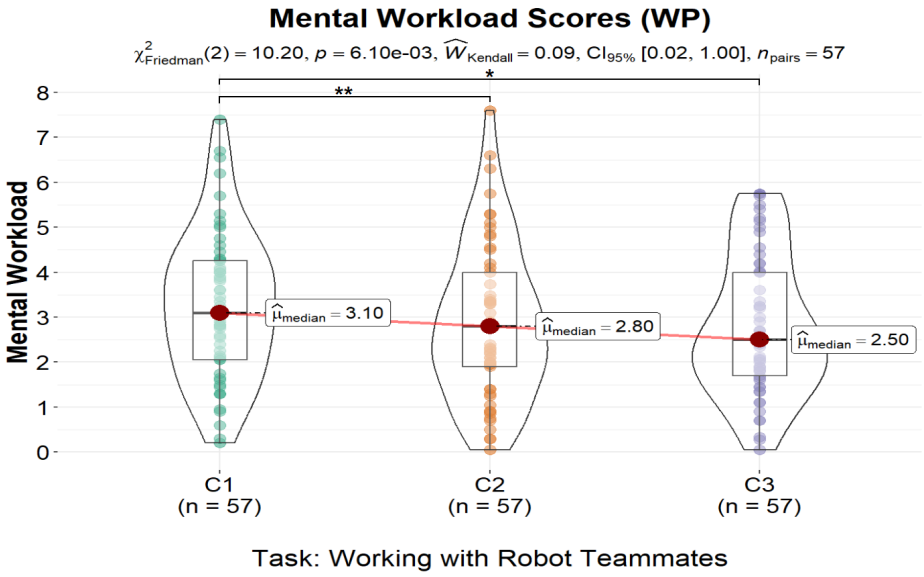

Fig. 6: Mental workload scores from Workload Profile (WP) for the *ST1: Working with Robot Teammates* for each information availability condition (C1, C2, C3). *P*-value scale: * (<0.05), ** (<0.01), *** (<0.001).

the information available on the minimap was greatest for Condition 2 (C2 = 3.13), less for Condition 3 (C3 = 3.10), and least for Condition 1 (C1 = 3.01). The Friedman rank sum test did not indicate a statistically significant difference in the perceived mental effort required for understanding the information on the minimap across the three information availability conditions.

(ST3) Using Information from the Minimap: In Table 2, the average mental effort required to use the information available on the minimap was greatest in Condition 2 (C2 = 3.26), less in Condition 3 (C3 = 3.18), and least in Condition 1 (C1 = 3.17). The Friedman rank sum test did not indicate a statistically significant difference in the perceived mental effort required for working with robot teammates across the three information availability conditions.

(ST4) Verbally Answering Questions: For the sub-task of verbally answering questions, the Workload Profile results indicate that Condition 1 (C1 = 3.53) required the highest average mental effort to perform the sub-task. Consequently, Conditions 2 and 3 (C2 = 3.25 and C3 = 3.28) required slightly less mental effort for the sub-task than Condition 1 required, although nearly the same amount between Conditions 2 and 3. The Friedman rank sum test did not indicate a statistically significant difference in the perceived mental effort required for the overall search task across the three information availability conditions.

NASA-TLX: On average, for the NASA-TLX results, Condition 1 (C1) required the most mental workload, followed by Condition 2 (C2), and Condition 3 (C3), respectively. After confirming a non-normal distribution with the Shapiro-Wilk test, the Friedman rank sum test revealed statistically significant differences in the NASA-TLX scores across the three information availability conditions, X2(2) = 34.86, $P < 0.0001$, W = 0.31 (moderate effect). The pairwise Wilcoxon signed rank test found statistically significant differences between C1 and C2 ($P < 0.0001$), between C1 and C3 ($P < 0.0001$), and between C2 and C3 ($P = 0.004$). Fig. 7 displays violin plots of the NASA-TLX data.

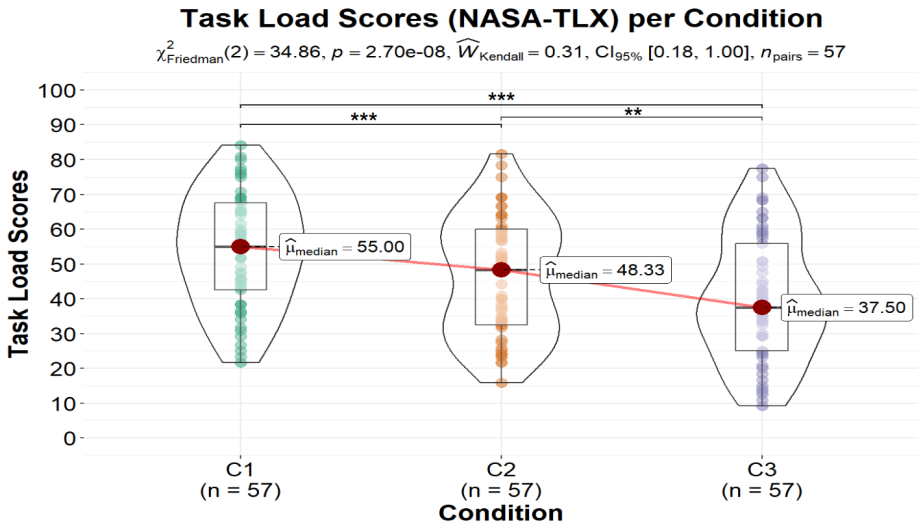

Fig. 7: NASA-TLX scores for the overall search task for each information availability condition (C1, C2, C3). P-value scale: * (<0.05), ** (<0.01), *** (<0.001).

Table 3 displays the individual workload scores for each dimension measured by the NASA-TLX. From the table, Mental, Temporal, Effort, and Frustration show gradual decreases from Condition 1 to Condition 2 to Condition 3. Additionally, Performance shows a gradual increase across the conditions. The Friedman rank sum test revealed statistically significant differences in Mental Demand (X2(2) = 20.58, $P < 0.0001$, W = 0.18 (small effect)), Temporal Demand (X2(2) = 21.09, $P < 0.0001$, W = 0.18 (small effect)), and Frustration (X2(2) = 17.84, $P = 0.0001$, W = 0.16 (small effect)). Pairwise testing using the Wilcoxon signed rank test indicated that C1,C2 ($P = 0.0004$) and C1,C3 ($P = 0.0005$) were statistically significantly different for Mental Demand. For Temporal Demand, statistically significantly differences were revealed between C1 and C2 ($P = 0.008$) and between C1 and C3 ($P = 0.0003$). Lastly, C1 and C3 had the statistically significant difference ($P = 0.003$) for Frustration.

Table 3: NASA-TLX scores per mental workload dimension and condition (C1, C2, C3). Best (lowest, except for Performance) scores are shown in bold. Most taxing dimensions (excluding Performance) for each condition are enclosed in a rectangle.

Cond	Mental	Phys	Temp	Perf	Effort	Frust
C1	66.32	15.70	67.19	20.88	58.16	36.32
C2	56.84	**12.72**	58.60	31.75	53.25	28.60
C3	**52.30**	15.44	**51.40**	**56.93**	**50.26**	**27.02**

Note: **Cond** = Condition, **Phys** = Physical, **Temp** = Temporal, **Perf** = Performance, **Frust** = Frustration.

6 Discussion

The virtual testbed developed for this study was shown to support investigations of non-hierarchical multi-agent collaboration in a search-and-rescue analogue. The study demonstrated the use of the testbed for understanding human-agent interactions and insight into the effects of information sharing on team and individual task performance. With this testbed, real-time communication and collaboration with autonomous agents such that trust and reliance among teammates is encouraged and team performance is improved can be investigated.

6.1 Task Performance

Regarding task duration, success rate, and the number of critical keys found, overall task performance of the human-robot teams improved as more information became available. Interestingly, the number of doors the teams opened in each condition did not statistically significantly increase as more information became available, meaning participants did not obsess over opening every door they saw. Although the slight increase in the number of doors opened by a team from Condition 1 to Condition 2 could signify this behavior, the individual performance breakdown revealed that the robots were the cause of the increase in the doors opened metric. These results demonstrate that participants did not become distracted by opening as many doors as possible as information availability increased.

The results pose an interesting question regarding what happened in Condition 2 such that a team was able to find significantly more critical keys in Condition 2 than in Condition 1, open significantly more doors, but were unable to find the target more frequently. Knowing that the robots were the cause for the statistically significant increase in doors opened, it is probable that the eight-minute time constraint was too restricting, and with a small addition of time, more participants might have found the target in Condition 2 since a larger

number of critical keys were found. Alternatively, looking at the average number of doors opened by each teammate it could be that the participants and robots were working more as individuals in a team rather than as teammates performing a collaborative task. After further analysis on the participants that found all three critical keys in Condition 2 but did not open the door, it was revealed that several of the participants were on their way to open the final door and find the target when the time limit was reached. However, it is difficult to determine whether adding time to the time constraint would result in a statistically significant increase in success rate from Condition 1 to Condition 2 to Condition 3.

6.2 Mental Workload

Interestingly, the Workload Profile indicated a reduction in mental workload as the information available increased, albeit without significant results between Conditions 2 and 3. Nonetheless, only the NASA-TLX revealed a significant gradual decrease in mental workload as more information became available to the teammates. These results could be consequential to participants not fully understanding the definitions of the different Workload Profile dimensions and how the dimensions might apply to the specific task in question. Alternatively, these results could be due to the specific dimensions of mental workload that each metric measures. For instance, the eight dimensions of the Workload Profile, provide a thorough and well-rounded metric of mental workload [10] and have shown to have a high diagnosticity [10, 14], while the NASA-TLX's six dimensions have been scrutinized for overshadowing researchers' mental workload demands of interest, namely those more directly involved with mental workload and attentional demand than with task difficulty [9, 10]. As such, the Workload Profile and NASA-TLX were included in this study to measure different aspects of mental workload. In addition to mental workload decreasing, it was expected for the focus of participants' mental workload to change in response to either a change in a participant's task focus or to a change in their task strategy. Reportedly, the Workload Profile has shown sensitivity and precision in measuring differences between specified tasks [10, 14].

The statistical tests performed on the Workload Profile scores for the four sub-tasks revealed that participants found it the least mentally demanding to work with the robots when the information on the minimap was being shared among all teammates (Condition 3). It is thought that no significant decrease in mental workload was seen between C2 to C3 because participants were influenced by the significant increase in the number of doors the robots opened in Condition 2, compared with Condition 3. From the think-aloud method the participants performed, it is clear they were surprised by how many doors the robots opened in Condition 2. Additionally, from the think-aloud method, participants noted more often that they were requesting help from the robots and accepting to help the robots compared with Condition 1. For Condition 3, the requests participants made changed from needing help finding a key they had not seen to requesting help fetching a key that a robot was near. Because all information was shared in

Condition 3, the robots more frequently successfully handed off keys participants requested. Experiencing more successful collaborations with the robots likely influenced the participants' perceived mental workload in regard to working with the robots in Condition 3.

The raw scores of the NASA-TLX dimensions (Table 3) indicate that sharing a team's common understanding among teammates (Condition 3) helped shift the load on Temporal Demand to achieve a more balanced mental workload across the other dimensions. By reducing Temporal Demand and Mental Demand, participants used less effort, experienced less frustration, and perceived better performance as more information became available to teammates. As [9, 10] have pointed out, the NASA-TLX includes measures of task difficulty rather than focusing on attentional demand, making the NASA-TLX and its dimensions a descriptive measure of how and why task performance improved. Making more information available to teammates not only reduced participants' mental workload, mainly Mental Demand and Temporal Demand, but also made the search task easier, seen as reduced Effort and Frustration with increased (perceived) Performance.

7 Conclusion and Future Work

Future teams will incorporate intelligent agents to more efficiently accomplish mission objectives and prioritize team safety. In these teams success will depend on effective collaboration that utilizes the strengths of both humans and agents, alongside a shared or common understanding of task and team functions. Establishing trust, fostering reliance, and maintaining a common understanding of task objectives should enhance team performance and reduce the mental demand of human teammates. A novel virtual testbed was created to begin multidisciplinary studies investigating multi-agent team collaboration, common understanding, and other teaming dynamics. Findings of an initial study using this testbed to evaluate how information availability impacted teaming dynamics revealed that as more information became available, a multi-agent team's performance on a search task improved. Participants' total mental workload decreased as the minimap reduced the cognitive load on working memory. Visibility into teammates' actions and results eliminated guesswork, enabling participants to rely on their collaborators and focus their efforts where they were most needed, further reducing the amount of frustration, effort, mental, and temporal demand required for improved performance. The amount of mental workload used on subtasks, as measured by the Workload Profile, shifted from remembering door and key locations while interpreting the robots intent and actions (Condition 1) to focusing on strategies using the information available to collaborate with the robot teammates, verbally answer real-time questions, and complete the overall search task (Condition 3). The testbed and task developed and demonstrated in this study can be adapted to look at numerous factors potentially affecting teaming between humans and autonomous agents. While this study investigated the effects of information sharing on team performance and mental workload

for human participants, future work could investigate the effects of failures in understanding of the task requirements (e.g., an autonomous agent collects keys but never uses them to open doors). In addition, the data collected during the task performance includes think-aloud data which could be analyzed to assess speech stress levels as a real-time assessment of human stress and potentially provide insight into trust in the autonomous agents.

Acknowledgement. The authors wish to acknowledge the technical and financial support of the Automotive Research Center (ARC) in accordance with Cooperative Agreement W56HZV-24-2-0001 U.S. Army DEVCOM Ground Vehicle Systems Center (GVSC) Warren, MI. Authors from TUKE wish to acknowledge the financial support to projects APVV-22-0414 & KEGA 049TUKE-4/2024. This material is based upon Cindy Bethel's work supported while serving at the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bolton, M.L., Biltekoff, E., Humphrey, L.: The mathematical meaninglessness of the nasa task load index: A level of measurement analysis. IEEE Transactions on Human-Machine Systems **53**(3), 590–599 (2023)
2. Chung, H., Holder, T., Shah, J., Yang, X.J.: Developing a team classification scheme for human-agent teaming. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 0 (2024). https://doi.org/10.1177/10711813241260387
3. Cohen, J.: Statistical power analysis for the behavioral sciences. Lawrence Erlbaum, Hillsdale, NJ, 2 edn. (1988)
4. Harbers, M., Bradshaw, J.M., Johnson, M., Feltovich, P., van den Bosch, K., Meyer, J.J.: Explanation and coordination in human-agent teams: A study in the bw4t testbed. In: 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. vol. 3, pp. 17–20 (2011)
5. Hart, S.G.: Nasa-task load index (nasa-tlx); 20 years later. Proceedings of the Human Factors and Ergonomics Society Annual Meeting **50**(9) (2006). https://doi.org/10.1177/154193120605000909
6. Hart, S.G., Staveland, L.E.: Development of nasa-tlx (task load index): Results of empirical and theoretical research. In: Hancock, P.A., Meshkati, N. (eds.) Human Mental Workload, Advances in Psychology, vol. 52, pp. 139–183. North-Holland (1988). https://doi.org/10.1016/S0166-4115(08)62386-9
7. Johnson, M., Bradshaw, J., Duran, D., Vignati, M., Feltovich, P.J., Jonker, C., Riemsdijk, M.: Rt4t: A reconfigurable testbed for joint human-agent-robot teamwork. In: Proceedings of the International Conference on Human-Robot Interaction (04 2015)
8. Lorenz, G., Ehrenstrom, J., Ullmann, T., Palmer, R., Tenhundfeld, N., de Visser, E., Donadio, B., Tossell, C.: Assessing control devices for the supervisory control of

autonomous wingmen. In: 2019 Systems and Information Engineering Design Symposium (SIEDS). pp. 1–6 (04 2019). https://doi.org/10.1109/SIEDS.2019.8735606
9. Mckendrick, R., Cherry, E.: A deeper look a the nasa tlx and where it falls short. Proceedings of the Human Factors and Ergonomics Society Annual Meeting (1), 44–48 (2018)
10. Rubio, S., Diaz, E., Martin, J., Puente, J.M.: Evaluation of subjective mental workload: A comparison of swat, nasa-tlx, and workload profile methods. Applied Psychology: An International Review **53**(1), 61–86 (2004). https://doi.org/10.1111/j.1464-0597.2004.00161.x
11. Shapiro, M.J.: 'the fog of war'. Security Dialogue **36**(2), 233–246 (2003). https://doi.org/10.1177/0967010605054651
12. Tomczak, M., Tomczak-Łukaszewska, E.: The need to report effect size estimates revisited. an overview of some recommended measures of effect size. Trends in Sport Sciences **01**(21), 19–25 (2014)
13. Tossell, C.C., Kim, B., Donadio, B., de Visser, E.J., Holec, R., Phillips, E.: Appropriately representing military tasks for human-machine teaming research. In: Stephanidis, C., Chen, J.Y.C., Fragomeni, G. (eds.) HCI International 2020 – Late Breaking Papers: Virtual and Augmented Reality. pp. 245–265. Springer International Publishing, Cham (2020)
14. Tsang, P.S., Velazquez, V.L.: Diagnosticity and multidimensional subjective workload ratings. Ergonomics (3), 358–381 (1996)