International Journal of Social Robotics (2022) 14:1117–1136
https://doi.org/10.1007/s12369-021-00834-1

Towards Human–Robot Teaming: Tradeoffs of Explanation-Based Communication Strategies in a Virtual Search and Rescue Task

Erin K. Chiou[1] · Mustafa Demir[1] · Verica Buchanan[1] · Christopher C. Corral[1] · Mica R. Endsley[2] · Glenn J. Lematta[1] · Nancy J. Cooke[1] · Nathan J. McNeese[3]

Accepted: 16 September 2021 / Published online: 19 November 2021

Abstract

Autonomous robots have the potential to play a critical role in urban search and rescue (USAR) by allowing human counterparts of a response team to remain in remote, stable locations while the robots execute more dangerous work in the field. However, challenges remain in developing robot capabilities suitable for teaming with humans. Communicating effectively is one of these challenges, especially if plan deviations during field operations require robot explanation. A virtual USAR team task experiment was conducted in Minecraft with a confederate acting as the remote robot. Four explanation-based communication conditions were tested: (1) always explain–the robot automatically provided explanations for any off-plan behaviors, (2) explain if asked–the robot provided an explanation only when the human counterpart requests it, (3) pull prime–the same as (2) but participants also experienced implicit training to pull information from the robot, and (4) never explain–a baseline condition in which the robot acknowledged requests but would not provide an explanation. Results indicate that the training in (3) generated more team communication than (1), but this did not improve team performance or shared situation awareness. Rather, team performance and shared situation awareness was best supported by a moderate level of explanations and the robot pushing information. These findings reinforce the importance of designing robot communication strategies that can reduce human workload, particularly communication overhead, in dynamic and time-constrained tasks.

Keywords Human–robot team · Search and rescue · Communication · Explanation · Situation awareness · Complex environment

1 Introduction

Using robots in urban search and rescue (USAR) has been ongoing for over a decade: at the 9/11 World Trade Center collapse [1]; the 2004 Mid Niigata and 2011 Tohoku earthquakes in Japan [2]; hurricanes Katrina, Wilma, and Rita in 2005 [3], and hurricane Harvey in 2018 [4]. In the wake of these catastrophes, robots were primarily used in reconnaissance missions to assess the disaster environment and search for victims in need of retrieval or rescue [5]. Today, USAR robots remain more as teleoperated tools under the careful control of highly trained robot wranglers. Therefore, though robots may be useful in enhancing USAR efforts and team capabilities, they still require significant human resources and human control to be effective [6].

A team can be defined as two or more interdependent agents that interact to achieve shared goals or tasks [7]. Under this broad definition, robots enabled by artificial intelligence (AI) may thus serve as "team members" to human counterparts, similar to human-animal teams in which human supervision remains a necessary part of team functioning [8]. A common reason for introducing automation (including some AI and robot applications) into existing work systems is to expand the operational capabilities of a system, given a limited human workforce. However, a common reason for teaming goes beyond scaling the workforce through the strategic distribution of work. Rather, teams are meant to surpass the aggregate abilities of working individuals that are choreographed to achieve certain outcomes, i.e., coordination [9]. Therefore, effective teams that include people

✉ Erin K. Chiou
erin.chiou@asu.edu

1 Human Systems Engineering, Arizona State University, Mesa, AZ, USA

2 SA Technologies, Mesa, AZ, USA

3 Human-Centered Computing, Clemson University, Clemson, SC, USA

require interactive communication and social processes to achieve emergent outcomes in complex task environments [10]–[13]. This idea of how teams function underlies the theory of interactive team cognition, which posits that a team's cognition is the team's interactive communication [10].

Interactive communication not only serves to transfer critical task-related information within a team, it also serves to build and repair trust between teammates through explanation [14, 15]. Whereas information transfer describes the preliminary work of teams [10], explanations as a specific type of information transfer can affect perceptions of a teammate's ability to help achieve the team task, thereby affecting trust in the teammate [15–17]. These perceptions of the teammate may subsequently affect the team's ability to communicate effectively, thereby undermining the team's situation awareness and performance. Although there have been several studies on human-robot communication such as how people perceive robot communication, what are various qualities of human–robot communication, and even how robots should communicate with one another in front of people [18], our focus is on effective strategies for human–robot team communication in complex task environments – particularly as technology is developed to exhibit increasing autonomy [19].

Previous studies of teaming in command-and-control environments indicate that team communication can predict team performance and shared situation awareness, whether in human–human teams [10, 20] or human–machine teams [21, 22]. In USAR environments previous studies have also addressed the importance of team communication, including team communication training and its impact on team performance and situation awareness [23, 24]. Other studies have outlined the technological feasability of human-robot communication [25], and developed coding schemes for human-robot communication to study how team tasks are accomplished [26]. Although this prior work has been essential for understanding and improving current human-robot USAR teams, much of this work has focused on field applications in which existing capabilities are considered (i.e., current human–robot USAR teams primarily involve human–human verbal communication and human–robot non-verbal communication). One limitation of such studies is that they provide little insight on the design of increasingly autonomous robots that may allow them to function as more capable teammates. Furthermore, few studies have examined how robot explanations – and how various strategies for delivering those explanations–affect team performance and situation awareness in complex team task environments such as USAR.

This study addresses these gaps by imagining future robots with autonomous navigation capabilities [27], to explore how such robots might serve as better team players to their human counterparts in dynamic USAR environments, through the use of various explanation-based communication strategies and training. We tested three explanation-based communication strategies and the effects of a communication training condition on a human–robot team's performance and situation awareness. The primary motivation of this study is to explore the tradeoffs of these various explanation-based communication strategies to inform the design of future robots and their ability to team effectively with human counterparts.

1.1 Maintaining Situation Awareness Through Team Communication

Studies have shown that team communication is critical for team performance [28, 29], especially when team communication involves transferring task-related information between teammates [30]. In dynamic environments, communication also informs teammates about any task relevant changes thereby maintaining the team's situation awareness. Although situation awareness can develop at an individual and team level, this study focuses on a within-team concept of situation awareness known as *shared situation awareness*, due to its association with team task interdependencies that require team communication and coordination in dynamic settings. Shared situation awareness differs from both individual situation awareness and team situation awareness. Whereas team situation awareness focuses on each team member possessing the situation awareness requirements for their respective responsibilities, shared situation awareness refers to the common awareness of the situation that is shared between teammates [31]. Practically speaking, in some cases shared situation awareness has been considered a sub-component of overall team situation awareness [32].

In previous studies, shared situation awareness (also referred to as team situation awareness) has been found to contribute to team performance [32], [33], and its positive association with team performance has been found in human-robot teams as well as human-human teams [34–37]. Findings of team communication supporting shared situation awareness in human-human teams is more common [38], whereas team communication that supports shared situation awareness in human-robot teams has been more challenging to achieve.

In a USAR competition for human-robot teams, all instances of robot or test arena damage were attributed to low shared situation awareness [39]. Low shared situation awareness of robot's location also negatively affected human-robot team performance in locating and extracting victims in a physically simulated USAR environment [40]. These observations highlight the importance of maintaining shared situation awareness to avoid critical errors and to more effectively accomplish USAR team tasks. As such, shared situation awareness in human-robot teams for USAR refers to *the shared perception of the robots' location, surroundings, and status; the comprehension of their meaning; and the pro-*

jection of how the robot will behave in the near future [41, 42].

In previous studies involving a simulated USAR environment, teams that had differing access to information in a virtual environment, and a restricted ability to communicate, had lower team performance compared to teams with equal access to information and fewer restrictions on their communication [43, 44]. Teams with more restricted ability to communicate also had higher variability in their situation awareness that included a lower range. Possible reasons for the higher variability could be that the communication restrictions resulted in a stronger effect of individual differences on communication behaviors. Another reason could be that having the fewer restrictions resulted in more effective communicating that improved shared situation awareness, but with an upper limit to the level of awareness that could be achieved. With fewer restrictions on communication, the act of communicating may have led to missing other important information or events occurring in the task environment. Furthermore, human teammates on teams with more restricted communication experienced higher cognitive workload. Given this challenge of balancing the costs and benefits of communication in human–robot teams, it thus seems crucial to further explore what are effective robot communication strategies that support shared situation awareness and team performance [45].

To maintain shared situation awareness, anticipating future needs and front-loading relevant information in team communication may help [37]. Proactive team communication during downtime avoids the risks of reactive or just-in-time communication, which can increase workload and cause cascading effects in dynamic task environments, to the detriment of team performance [46]. In addition to proactive communication, information *elaboration* has been found to have a stronger relationship with team performance compared to communication frequency [30]. Information elaboration refers to communication that encompasses clarifying information, including explanations. Because the impact of clarifying information (e.g., status updates and status confirmations) on maintaining situation awareness in human–robot teams has been more extensively studied [13, 23, 26, 47], our study focuses on explanations more specifically, due to their purported but under-explored impact on human–robot teaming [15, 48, 49].

1.2 Explanations

Explanations refer to explicit communication that provides a *contrastive* reason behind a decision or action occurring relative to a counterpart's understanding [15]. This type of communication between teammates, which often instantiates as responses to "why" questions, can help improve human teams' shared situation awareness [38, 50], and trust in machine learning agents [51]. Situation awareness and trust are both important qualities in teams for successfully operating in dynamic and complex task environments.

Explanations help maintain shared situation awareness in dynamic situations, including awareness of team members' shifting purpose, process, and performance information, which are types of information that affect trust. Without the ability to explain actions while performing team tasks, misunderstandings between teammates can arise and lead to future surprises or additional work to resolve breakdowns in teaming. These misunderstandings can lead to disengagement or overreliance on a teammate, affecting a team's shared situation awareness and ability to adapt to unexpected events. Given the importance of shared situation awareness and trust in dynamic teaming [14, 52], a team's ability to perform effectively in a dynamic task environment may be undermined without explanations.

A lack of natural language processing or shared mental models with robot teammates can introduce unexpected challenges during deployment, especially without a way to resolve those challenges in the moment, compared to how human teammates are able to interact with and adapt to other people [53]. Currently, robots designed to operate like autonomous teammates essentially follow pre-planned routines, akin to performing choreography, or they require significant resources to operate (e.g., a dedicated robot operator). Such team compositions that rely on human adaptability to address unexpected issues may still be useful but will have limited value in USAR compared to teams with more dynamic abilities. As robots become increasingly autonomous, the flexibility of human–robot teams will be increasingly crucial for highly complex environments with limited resources and changing conditions. The potential for robot teammates to function beyond choreography toward facilitating emergent team interactions will thus require the ability to provide explanations to their human counterparts in the wake of unexpected events or behaviors.

2 Current Study

Our primary purpose was to investigate the effects of different explanation-based communication strategies on team performance and shared situation awareness in a dynamic USAR team task environment. First, we describe the microworld testing environment used to conduct the study.

2.1 Simulated Search and Rescue Team Task Environment

A virtual USAR team task environment [54] developed in Minecraft (version 1.11.2) and based on [43] was rebuilt to include dynamic events [55]. These dynamic events were

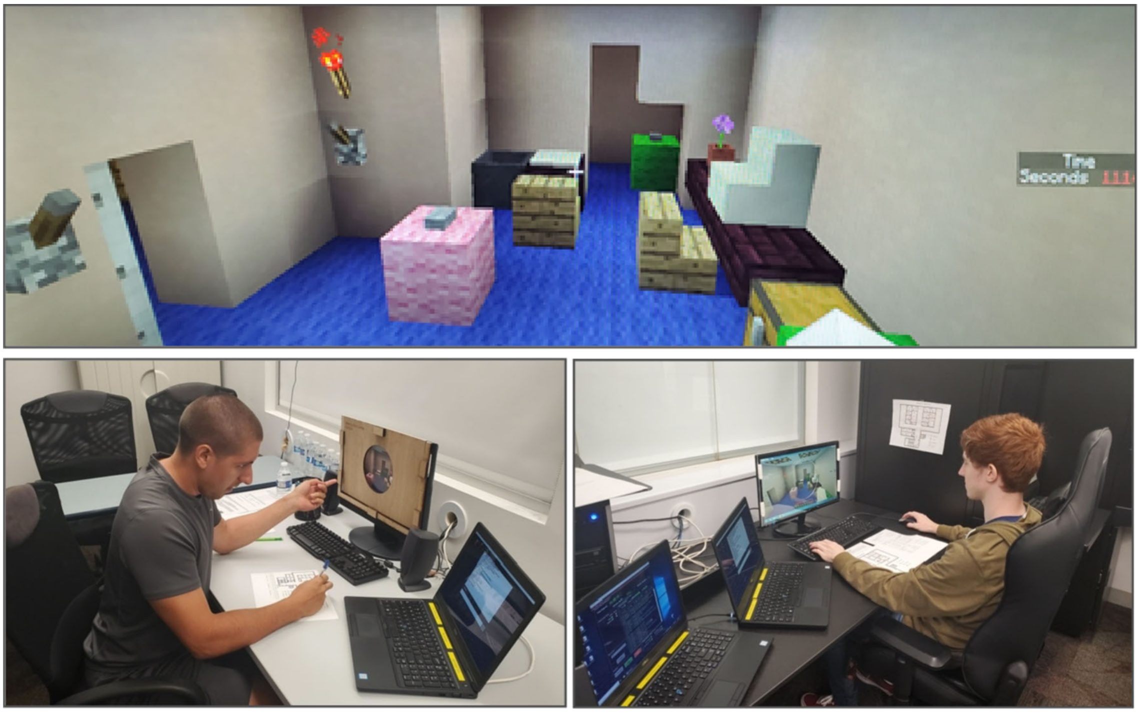

Fig. 1 Screenshot of the Minecraft simulated USAR task environment (top), and a reenacted setup of the participant station (bottom left) and the "robot" station (bottom right)

introduced to motivate the need for explanations–to justify plan deviations in real time. The USAR team involved two heterogeneous team members – a robot and a human – who interacted virtually. The robot's job was to autonomously navigate a partially collapsed building and to locate potential victims in the field, and the remote navigator's job was to monitor the robot's live video feed (a first-person view) to track the location of the robot while documenting the location of victims on a map of the building. The navigator was provided with marking utensils and a paper map (i.e., a floor plan) of the building pre-collapse and was also tasked with documenting any changes to the map (see Fig. 1). The objective of the team was to complete a reconnaissance mission (i.e., identifying victim locations and changes to the building) before sending in a recovery team to retrieve the victims safely.

The "robot" in our study was, in fact, a highly-trained researcher controlling the navigation of a virtual "robot" in Minecraft. This "wizard-of-oz" approach in user-centered research has been used in human-autonomy teaming research and allows researchers to compare technology designs that would be cost-prohibitive to develop fully [27, 56]. Wizard-of-oz is also used to anticipate future designs and test how people may respond to those designs. Hereafter, we refer to the researcher-controller as "the robot." Study participants were assigned the navigator role and told that they would be completing two USAR missions with an autonomous robot in a virtual task environment.

In addition to sharing a video feed [24], the robot interacted with the navigator via text chat communication. Participants were not limited to what they could ask or say to the robot through the text chat, but they were provided information about the robot's capabilities in advance. The robot was limited to providing explanations that were relevant to the task environment and to respond to the participant using a series of pre-scripted phrases. For example, in response to questions that were not relevant to the task environment the robot would write, "sorry, I am unable to help you with that." Comments that were not relevant to the task were ignored by the robot. The robot was designed not to have complete natural language abilities, in part to maintain the perception that participants were interacting with an autonomous robot rather than another human teammate, and also to ensure some control across study conditions.

2.2 Study Design

2.2.1 Human–Robot Team Communication

Although this study focuses on explanation-based communication strategies, other forms of team communication such as status updates and status confirmations are necessary to complete the team task. Table 1 shows example text chats between

Table 1 Examples of human–robot text chat communication during the USAR task

Example	Sender	Content	Type
1	Robot	205 is next	A status update absent a request (pushing information). The robot is communicating which room it will enter next
2	Navigator	Why the deviation	A "why" question that indicates a request for an explanation
	Robot	201 and 203 were skipped because we already searched them	An explanation is provided regarding a deviation observed via video feed that motivated the request
3	Robot	207 is next	A status update absent a request (pushing information)
4	Robot	All clear	A status update absent a request (pushing information). The robot is communicating the status of the room it was in – it has completed scanning and no victims were found
5	Navigator	How many in 207	A "how many" question in this task environment was considered a status update request (pulling information)
6	Navigator	Where are you is this room 207	A status update request (pulling information)
7	Robot	Ready to move on to the next hallway?	A status update request (pulling information)

the robot and human navigator that generally occurred across conditions, and an example robot explanation provided in response to a request depending on the study condition (addressed in the next section).

2.2.2 Selection of the Study Conditions and Hypotheses

Within our USAR team task environment, we were interested in which communication strategy would best support the following team effectiveness metrics: team performance (proportion of correctly triaged victims), shared situation awareness (accuracy of the annotated map), trust in the robot (high trust would be appropriate because the robot was designed to be autonomous and did not make mistakes), and workload (although some level of workload may indicate task engagement, interactions with the robot should not significantly increase workload).

Four conditions (between subjects effect) were tested: (1) *always explain*—the robot proactively explained plan deviations, and provided other explanations if asked (2) *explain if asked*—the robot provided explanations only when asked, (3) *pull prime*—the robot provided explanations only when asked *and* participants received some training to ask the robot questions, and (4) *never explain*—the robot did not provide explanations, but did acknowledge requests for explanations. We describe our reasoning behind each of these conditions in more detail below.

Based on prior work, we expected that the *always explain* condition would result in the best outcomes for team performance and shared situation awareness [57]. Proactive communication can indicate an ability to anticipate a teammate's needs [10], and reduces communication overhead that results from teammates needing to request information. Requesting information generally involves at least five steps: first determining the information needed, forming a request, making the request, monitoring for a response, and then processing the information. This is more cognitive work than simply receiving the right information at the right time, which generally involves just two steps – monitoring for new information and then processing the information.

However, it is also possible that *always explain* could result in information overload, requiring vigilant monitoring and filtering of an information stream. This information overload could further undermine the intended benefits of an *always explain* strategy which are to: reduce workload, increase shared situation awareness in a timely manner, and to foster trust in the robot teammate. Furthermore, the *always explain* condition may lead a teammate to adopt a more passive role in team communication, thereby failing to identify a need to request an explanation when one might be needed. Therefore, as an alternative to *always explain,* we also tested an *explain if asked* condition in which the robot would provide an explanation only if the participant requested one.

The *never explain* condition served as a baseline in which no explanations were provided, although the robot still communicated general status updates with the navigator in all conditions. We expected that shared situation awareness, and in particular team trust, would be the lowest in the *never explain* condition due to the purported role that explanations play in team communication. To avoid conflating perceptions of the robot's general ability to communicate with the robot's ability to provide explanations, the robot would politely [58] acknowledge requests for explanations in this condition without providing one by saying, "sorry, I am unable to help you with that."

Initially, we elected to test the above three conditions only. However, during pilot testing we found that most participants would not initiate requests for information in the *explain if* *asked* condition, thus undermining the condition. Therefore, we developed and introduced a fourth condition, called *pull prime*, in which participants were "primed" to initiate communication with the robot. Our use of the term "priming" refers to the presentation of a stimulus that is thought to impact subsequent behaviors, borrowing loosely from the more established notion of semantic priming that involves the rapid and successive presentation of the initial stimulus. In the *pull prime* condition, the researcher prompted participants to initiate an interactive training sequence by asking the robot a question. In the other three conditions the robot would initiate the interactive training sequence, which involved scripted communication about the navigation task with participants. Other than this difference in who initiated the training sequence, participants in the *pull prime* condition experienced the same communication strategy as *explain if asked*.

We expected that priming participants would lead to more requests for explanations (asking more "why" questions) during subsequent interactions with the robot, without explicitly telling them to do so. The reasoning behind this is based on social exchange theory and the primacy effect. Prior work in social exchange theory has found that the structure of repeated interactions between a person and an automated agent, such as who initiates requests for resources, can affect subsequent teaming behaviors with an agent [59]. In addition, the primacy effect states that a person's first impressions of a robot impacts their subsequent behaviors with the robot [60]. Therefore, we expected that introducing a social structure in which participants initiated interactions as part of their first impression of the robot would lead to participants asking more questions (and requesting more explanations) than in the other conditions. As a result of asking more questions, we also expected that the *pull prime* teams would be more effective than in the *explain if asked* condition, in which participants were not primed to ask the robot questions.

In summary, *always explain* was expected to have the highest shared situation awareness and team performance, high trust in the robot, but also the highest workload. *Pull prime* (which was *explain if asked* with participant priming) was expected to have high shared situation awareness and team performance, the highest trust in the robot due to a greater number of exchanges with the robot [12], and less workload than *always explain* due to actively requesting information when needed rather than constantly monitoring the text chat. *Explain if asked* was expected to have less shared situation awareness, lower team performance, less workload, and the same levels of trust in the robot compared to the previous two conditions, contingent on the resulting number of explanations requested. *Never explain* was expected to have the lowest shared situation awareness, team performance, workload, and trust in the robot.

2.2.3 Designing a Task-Relevant Need for Explanation-Based Communication

Before the experimental trials, as part of the interactive training, participants engaged in a short planning session with the robot in which the robot first established the planned route it would take, based on a map of the building, including the order of the targeted rooms. Two types of explanations were designed to occur as a result of deviations to this plan: (1) when the robot encountered an unexpected opening to another room and used it to access and search the target room, and (2) when the robot encountered areas that were blocked by debris and needed to deviate from the planned route. When it encountered an opening, the robot would explain that it found a different route to the target room through an opening in the wall. For blockages, the robot would explain that it could not enter the target room because it was blocked. There were four openings and four blockages in each of the two missions that participants experienced. Additionally, the second mission introduced two dynamic elements, (1) a dynamic wall collapse that triggered when the robot traversed its planned route and (2) the presence of critical victims that needed to be reached within a certain time period.

Critical victims were introduced to add temporal urgency and difficulty to the second mission. Orange blocks in Minecraft were used to represent these victims; these blocks would remain orange for the first 10 min of the mission, and then turn pink to indicate their expiry. Ideally, the human–robot team would prioritize searching for these critical victims within the first 10 min of the mission to minimize casualties. At the same time, the team's overall goal was still to accurately document structural differences between the pre-collapsed building and the actual environment, and to correctly mark the location and type of victim (e.g., alive, critical, deceased) on their mission map. To achieve this, the navigator and the robot needed to coordinate and communicate with one another in a timely manner.

3 Method

3.1 Sample Size and Participants

An a priori power analysis was conducted using G*Power3 [61] to test the difference between the means of 4-conditions by 2-missions using an *F-test*, with a medium effect size ($\eta_p^2 = 0.06$; [62]), and an alpha (∝) of 0.05. According to the result, a total sample of 60 participants with four equal-sized groups of $n = 15$ was required to achieve a power of 0.90.

Thus, 60 participants from Arizona State University and its surrounding community were recruited for this study. Participants were required to be 18 years or older, fluent in English, have normal or corrected-to-normal hearing and vision (e.g., colorblind participants were excluded from this study), and comfortable using a standard computer mouse and keyboard. Participants were compensated $10 per hour for their time spent during the study, or roughly $15 total.

Study recruitment materials advertised a human–robot study in Minecraft. The resulting demographics collected at the end of the study show that participants were generally younger adults ($M_{age} = 22$, $SD_{age} = 7.21$), skewed male (70% identified as "Male"), and had some prior experience playing Minecraft. When asked, "I am experienced with playing Minecraft" on a scale of zero (strongly disagree) to seven (strongly agree), participants rated $M = 4.53$. However, there was some variation in Minecraft experience across study conditions, after random assignment (*never explain* $M = 4.6$, $SD = 2.41$; *always explain* $M = 3.27$, $SD = 2.4$; *explain if asked* $M = 5.73$, $SD = 1.94$; *pull prime* $M = 4.53$, $SD = 2.23$). This is addressed briefly in Sect. 5.3.

Table 2 Summary of the study procedure

Step	Description	Duration
1	Recruitment, briefing, and informed consent ($N = 60$ participants)	Pre-study
2	Random assignment into one of four study conditions: *always explain*, *explain if asked*, *pull prime*, *never explain*	Pre-study
3	Training via PowerPoint tutorial (~ 10 min) and in the task environment with the robot (~ 10 min)	20 min
4	Mission 1	20 min
5	Workload and trust questionnaires administered	~ 15 min
6	Mission 2	20 min
7	Workload, trust, and demographics questionnaires administered	~ 15 min
8	Debriefing	Post-study

3.2 Equipment

Four computers running Windows OS were used to conduct the study. One desktop computer was used to show PowerPoint slides as part of the training session, and to complete the missions in Minecraft. The missions were recorded using Snagit, a screen capture and recording software program (techsmith.com/screen-capture.html). The desktop was connected to two monitors with mirrored displays and two keyboards. The monitors were connected through a small opening in a wall separating the robot/researcher from the navigator/participant such that the robot and navigator each only had access to one of the connected monitors. Three laptop computers were also used to enable the robot and navigator to communicate with one another. Each team member used one laptop to access the chat system interface, and the third was used to host the chat server.

Figure 1 shows an example view of the virtual task environment (top), and the experimental setup between the navigator (participant, bottom left) and the robot (wizard-of-oz researcher, bottom right). The participant's field of view was purposefully restricted using a wood frame overlaid on their monitor screen, to simulate a task environment in which remote responders have a narrower field of vision, and to increase participants' required vigilance and dependence on the robot to complete their mapping task.

Minecraft by Mojang (minecraft.net and mojang.com) was selected as the platform for the team task environment due to its flexibility as well as the ease with which it can be controlled and modified. Following their training session, maps of the building floorplan pre-collapse were provided to participants. Following each experimental mission with the robot in Minecraft, participants were asked to complete several questionnaires, which are described in more detail below. Additional paper materials used were for consenting and debriefing participants.

3.3 Procedure

After consenting to participate in the study, each participant was randomly assigned to one of four conditions: *always explain*, *explain if asked*, *pull prime*, or *never explain*. All participants reviewed a voiced-over tutorial created in PowerPoint (~ 10 min), completed interactive training with the robot in Minecraft while using the chat system (~ 10 min), completed two experimental missions with the robot in Minecraft (20 min each), and were administered questionnaires (~ 15 min). In total, the session lasted roughly 1.5 h per participant. Table 2 outlines the overall study procedure, which is described in further detail below.

The PowerPoint tutorial informed participants about the team task and mission, individual roles, the robot's capabilities (including how and what could be communicated to the robot), and how team performance during the task would be scored. Images of common robots used in USAR (e.g., robot rovers with traction propulsion, a camera, manipulator, and end effector attached) were shown as part of the training to provide additional context that participants would be interacting with a robot. Following the tutorial, participants completed an interactive training with the robot in Minecraft to achieve three training goals: (1) to practice observing the robot navigating the environment, (2) to experience communicating with the robot, and (3) to practice marking up the mission map. No dynamic events were experienced during the interactive training. Next, participants completed two missions with the robot in Minecraft with increasing task complexity from Mission 1 to Mission 2. This increasing task complexity was to increase the workload between mis-

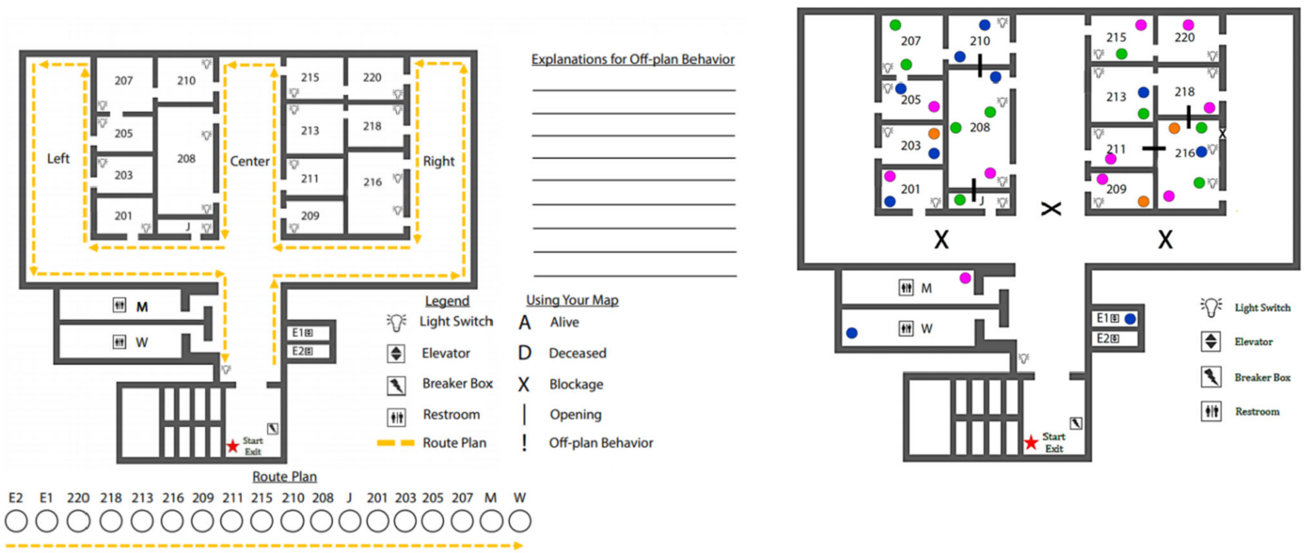

Fig. 2 Map provided to participants before the start of a mission (left) and the answer key map for Mission 2 (right), which was used by researchers to code participant map data. The colored circles indicate different types of victims, the "X" indicate blocked pathways, and "|" indicate new openings

sions by adding collapses, additional victims, and fires in the task environment.

In each mission, participants were responsible for marking the location of victims and any changes to the map (i.e., potential dangers, blockages, or openings). Meanwhile, the robot searched for victims in the building while relaying its location to the participant through the text chat. As part of their role as navigator, participants received a map of the building pre-collapse and were asked to use the map to log any found victims' location and changes to the building that would impede a rescue mission's effort to retrieve the victims. The map included the robot's planned path and location of light switches, elevators, and restrooms (see Fig. 2). The participant could reference a pre-planned search route annotated on the map to help track the robot's movement on the ground. The location of light switches was important for map orientation because as the robot entered a room, it would turn the lights on (see Fig. 2). Inconsistencies between the map and the actual virtual environment included: missing walls, additional walls, and misplaced doorways to simulate a damaged building.

After completing the first mission, participants were administered a modified NASA-TLX workload assessment and a questionnaire that assessed trust in the robot teammate. Following the second mission, participants were administered the same NASA-TLX workload assessment and trust questionnaire, as well as questionnaires for demographics, including previous experience with Minecraft. The robot's activities within Minecraft were obtained via a task data logging system constructed in Minecraft.

3.4 Outcome Measures and Data Processing

To address our research questions, the following measures of team effectiveness were assessed: team performance, shared situation awareness, trust in the robot, workload, and team communication. Three types of team communication were coded from the text chat data – explanations, pulling information, and pushing information (the latter two exclude explanations). We further describe these measures below.

3.4.1 Team Performance

Team performance in this task was measured by the proportion of correctly identified victims (location and status). Although documenting victims' location and status was the primary goal of the navigator, the navigator needed to rely on the robot to achieve it. The number of victims identified were counted, divided by the total number of victims in the collapsed structure, and then multiplied by 100 to arrive at the team performance score for each mission.

3.4.2 Shared Situation Awareness

Shared situation awareness was measured by the navigator's proportion of map accuracy in each mission. To calculate this, we summed the correctly annotated collapses and openings on the map, divided by the total number of imposed collapses and openings, then multiplied by 100 to obtain a percentage. This value made up the mission-level score for shared situation awareness.

3.4.3 Trust in the Robot

Trust in the robot was measured using a questionnaire adapted from Hoffman et al. [49] that focuses on trust in explainable AI. This eight-item questionnaire was based on Cahour and Forzy [63]; Jian, Bisantz, and Drury [64]; and Madsen and Gregor [65]. The resulting composite questionnaire emphasizes that trust should be viewed as a process when evaluating explainable AI systems and thus, emphasizes a repeatable evaluation of trust after explanations are provided. Trust ratings for each item used a 5-point Likert-type scale (1 = "I agree strongly", 2 = "I agree", 3 = " Neutral", 4 = "I agree", and 5 = "I disagree strongly"), containing seven positive items (e.g., "I am confident in the Robot. I feel that it works well") and one negative item ("I am wary of the robot.") To calculate the final trust score for each participant, we totaled the ratings from the positively valenced items and subtracted the rating from the negatively valanced item.

3.4.4 Workload

Workload was measured using the NASA-TLX [66] with two modifications – only the questionnaire items were used, and the response was limited to a 5-point scale to mirror the numeric range used in the trust questionnaire, which was administered at the same time (see Table 2). Past studies that used only the questionnaire portion of the NASA-TLX have also moved away from the 21-point scale [67]. For administrative simplicity and considering user experience, the same numeric range for all questionnaire items was used. In this study, the six items from the NASA-TLX were summed to form each participant's workload score.

3.4.5 Communication Behaviors: Explanation, Pushing Information, Pulling Information

The three communication measures were gleaned from the robot and participant text chat data. We considered the amount of explanations provided (robot only, due to its specific role in the task), the amount of information pulled (i.e., questions that would not be classified as requests for explanations, from the robot and participant), and the amount of information pushed (e.g., status updates, from the robot and participant). These three communication measures allowed us to assess participants' communication relative to the robot's (e.g., what were some resulting effects on team communication, and did our *pull prime* training lead to more questions or requests for explanations). To obtain these team communication measures, each message within the text chat data was transformed into a row within a datasheet. Two trained experimenters coded the rows independently, and Cohen's κ was run to assess inter-rater reliability.

Table 3 Example coding of explanations

Sender	Message Text	Explanation
Navigator	Why did you skip E0?	0
Robot	Because we already searched it	1
Navigator	What about room 212?	0
Robot	We already searched room 212	1

An *explanation* was identified (coded as 1 instead of 0) when the robot provided explanations after the navigator asked a "why" question or an implied "why" question, such as "what about" questions, as in Table 3. To avoid double-counting explanations, requests for historical information, such as requests to repeat previously answered information in the same context, were not considered explanations but clarifications. In coding the explanations, there was almost perfect agreement between raters [$\kappa = 0.936$ (95% CI, 0.885 to 0.987)]. Therefore, we took the average of the two raters' codes and totaled them across each mission before further analyses.

Pulling information was identified as any requests for information from any team member that was not semantically a "why" question. An example of this would be seeking confirmation, such as the navigator asking, "Did we search room E0?" Note that this definition does not overlap with an explanation, because explanations were coded as information provided in response to a "why" question (retroactive, contrastive information). Again, there was almost perfect agreement between the two raters [$\kappa = 0.973$ (95% CI, 0.969 to 0.977)], so we took the average of the two raters' codes totaled across each mission before further analyses.

Pushing information was identified when any team member sent information that did not follow a question. Therefore, potential "anticipatory explanations" (information provided in anticipation of a need to reconcile contrastive understanding) would have been captured in this measure, in addition to other task-relevant information provided before a request (or in the absence of a request). This could include status updates that would help maintain shared situation awareness but are not considered explanations because they do not provide directly observed contrastive information. There was also almost perfect agreement between the raters [$\kappa = 0.882$ (95% CI, 0.876 to 0.888)]. Therefore, we took the average of the two raters' codes and summed them across each mission before further analyses.

3.5 Data Analysis Approach

To assess how team performance, shared situation awareness, trust, workload, and robot explanations varied across conditions and missions, we conducted a 4 (number of conditions) × 2 (number of missions) split-plot analysis of variance

(ANOVA) for each of the five outcome measures. We chose to run separate ANOVAs instead of a multivariate analysis of variance (MANOVA) because of a significant moderate positive correlation between team performance and shared situation awareness, $r(118) = 2.11$, $p = 0.037$; and between trust and workload, $r(118) = 0.23$, $p = 0.013$.

Next, we ran two repeated measures three-factor mixed ANOVAs, to address how information pulling, and information pushing changed across conditions, missions, and roles (robot and navigator). We applied Fisher's Least Significance Difference (LSD) to compare each pair of conditions because four conditions control the family-wise alpha at the per contrast alpha. Additionally, Fisher's LSD assumes reasonably good homogeneity of variance [68, 69]. To conduct the ANOVAs, we used IBM SPSS Statistics 26 software [70].

Finally, we used stepwise regression (based on Akaike information criteria, AIC) to determine the best set of predictors for team performance and shared situation awareness from explanations, pushing information, pulling information, trust in the robot, and workload. We chose stepwise regression because it eliminates the multicollinearity issue by including an additional predictor variable (i.e., forward selection) and eliminating a predictor variable (i.e., backward elimination) already in the model [71]. We also chose AIC because our sample size was limited, and AIC places a moderate penalty on the number of predictor variables compared to Bayesian, which places a heavier penalty [72]. This analysis was conducted in R [73], using the MASS packages for stepwise regression [74] and lm-beta [75] for adding standardized regression coefficients.

Team performance and shared situation awareness were considered the primary outcome variables given the specific goals of the human–robot team. However, trust and workload were also considered as the literature indicates that they could be impacted by the communication strategies. The three communication behaviors (robot explanations, robot-navigator pushing information, and robot-navigator pulling information) were analyzed in an exploratory manner to cross-check the effects of the four conditions on the actual communication behaviors of the dyad.

Notably, participants' trust in the robot did not differ across conditions or missions; average trust in the robot was moderate ($M = 2.83$, $SD = 1.70$; $p > 0.05$) and was not related with the outcome variables ($p > 0.05$). Therefore, we omit reporting the trust measure in the results section and revisit the implications of this in the discussion section.

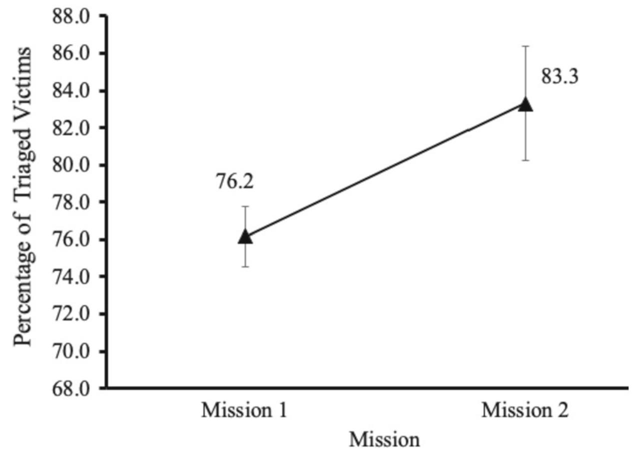

Fig. 3 Plot of mean team performance by mission (error bars are 95% confidence intervals)

4 Results

4.1 Split-Plot Analysis of Variance

The first split-plot ANOVA addresses how team performance (percent of correctly identified victims) differed across conditions and missions. While there was a significant mission main effect, $F(1, 56) = 26.5$, $p < 0.001$, $\eta_p^2 = 0.02$, the condition main effect and condition by mission interaction effect were not significant, $F(3, 56) = 0.44$, $p = 0.728$, $\eta_p^2 = 0.02$, and $F(3, 56) = 0.19$, $p = 0.902$, $\eta_p^2 = 0.01$, respectively. These results indicate that there may have been a team performance learning effect despite the additional challenges introduced in Mission 2 (Fig. 3).

The second split-plot ANOVA addresses how shared situation awareness (map accuracy) differed across conditions and missions. The findings indicate that all three effects were not statistically significant, the condition main effect, $F(3, 56) = 1.93$, $p = 0.136$, $\eta_p^2 = 0.09$, the mission main effect, $F(1, 56) = 0.25$, $p = 0.620$, $\eta_p^2 = 0.04$, and the interaction effect of condition by mission, $F(3, 56) = 0.66$, $p = 0.583$, $\eta_p^2 = 0.03$.

Although there was no significant condition main effect, pairwise comparisons (based on LSD—least significance difference) indicate that the *always explain* condition had better shared situation awareness than the *pull prime* condition ($p = 0.020$), but did not differ from the remaining two conditions, *never explain* or *explain if asked* (Fig. 4).

The third split-plot ANOVA addresses how participants' workload differed across the conditions and missions. Similar to team performance, there was a significant mission main effect, $F(3, 56) = 41.6$, $p < 0.001$, $\eta_p^2 = 0.43$, while the condition main effect and condition by mission interaction effect were not significant, $F(3, 56) = 0.95$, $p = 0.424$, $\eta_p^2 = 0.05$ and $F(3, 56) = 0.19$, $p = 0.906$, $\eta_p^2 = 0.01$, respectively.

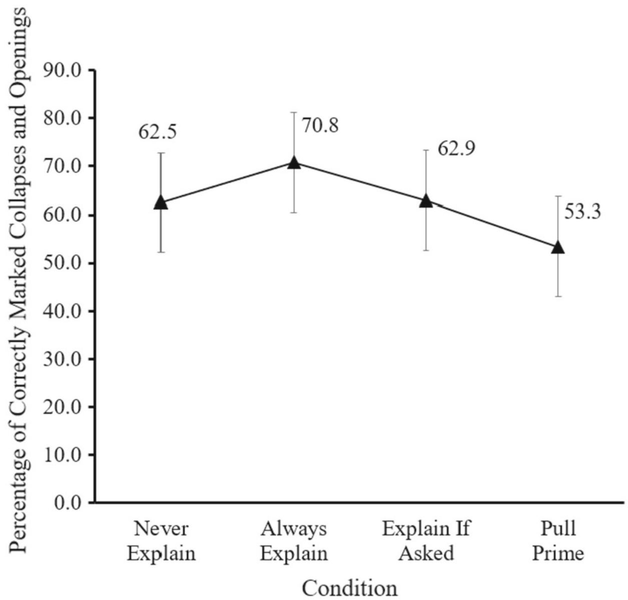

Fig. 4 Plot of mean shared situation awareness across conditions (error bars are 95% confidence intervals)

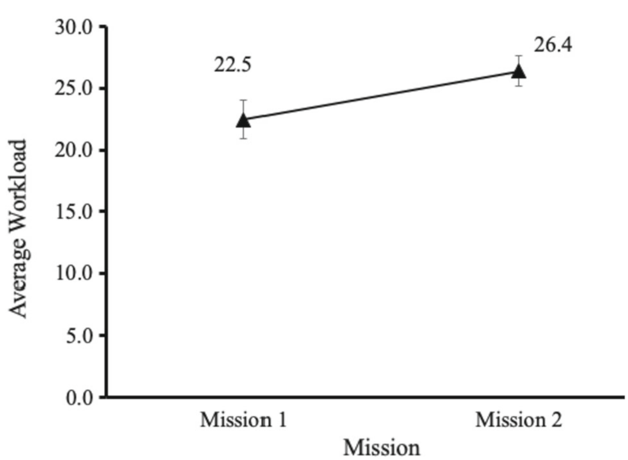

Fig. 5 Plot of participants' reported workload across missions, with 30 being the highest possible workload rating (means with 95% confidence intervals)

The pairwise comparisons (LSD) show that workload in Mission 2 was higher than in Mission 1 ($p < 0.001$; Fig. 5). This was somewhat expected as Mission 2 was designed to be more challenging than Mission 1, with dynamic events and an increased number of victims and roadblocks.

The fourth split-plot ANOVA addresses how the number of robot explanations differed across the conditions and missions. The condition main effect was statistically significant, $F(3, 56) = 4.28$, $p = 0.009$, $\eta_p^2 = 0.19$, while the mission main effect and the interaction effect of condition by mission were not statistically significant, $F(1, 56) = 1.44$, $p = 0.235$, $\eta_p^2 = 0.03$ and $F(3, 56) = 0.51$, $p = 0.678$, $\eta_p^2 = 0.03$, respectively.

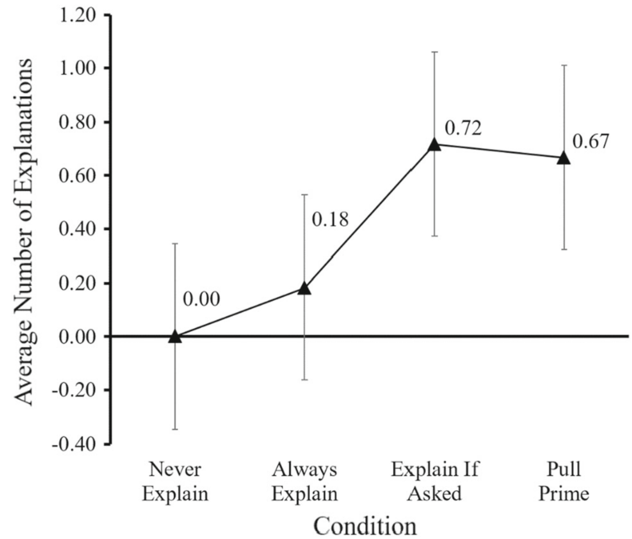

Fig. 6 Plot of robot explanations across conditions (means with 95% confidence intervals)

In accordance with the significant condition main effect, pairwise comparisons (LSD) indicate that participants in the *pull prime* and *explain if asked* conditions experienced more robot explanations than participants in the *always explain* ($p = 0.051$ and $p = 0.032$, respectively) and *never explain* conditions ($p = 0.008$ and $p = 0.005$, respectively; Fig. 6). Based on our coding scheme, this also indicates that participants in the *pull prime* and *explain if asked* conditions asked the robot more "why" questions than in the *always explain* condition.

4.2 Mixed Analysis of Variance

Next, we discuss the between-subjects effect (i.e., condition) on the information pulling and pushing behaviors. Given the way the team task was structured, only the robot's role demanded explanations. However, pulling and pushing of information generally needed to take place between the navigator and robot to carry out the team task.

4.2.1 Pulling of Information

The teams' pulling of information was analyzed via a repeated measure three-factor mixed ANOVA, with condition as the between-subjects factor, mission as a within-subjects factor, and role nested (within missions) as within team factors. There were significant main effects for condition [$F(3, 224) = 9.53$, $MSe = 106$, $p < 0.001$, $\eta_p^2 = 0.11$] and role (within mission) [$F(2, 224) = 5.88$, $MSe = 65.4$, $p = 0.003$, $\eta_p^2 = 0.05$], and a significant interaction effect for condition by role (within mission) [$F(6, 224) = 5.67$, $MSe = 63.2$, $p < 0.001$, $\eta_p^2 = 0.13$]. However, the mission main effect was not significant [$F(1, 224) = 0.35$, $MSe = 3.88$, $p =$

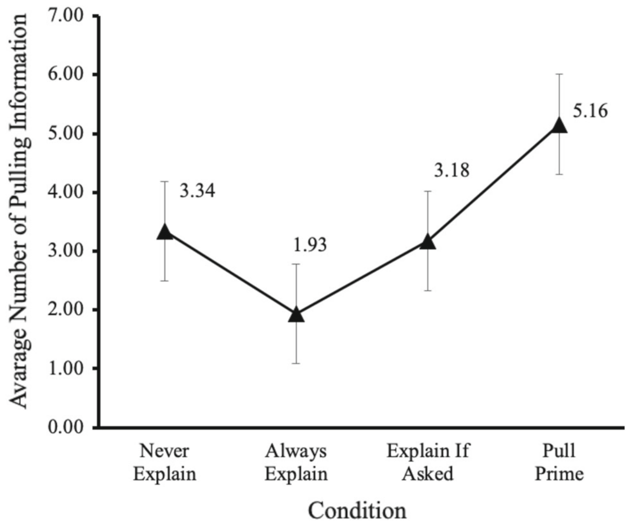

Fig. 7 Plot of information pulled by condition (means with 95% confidence intervals)

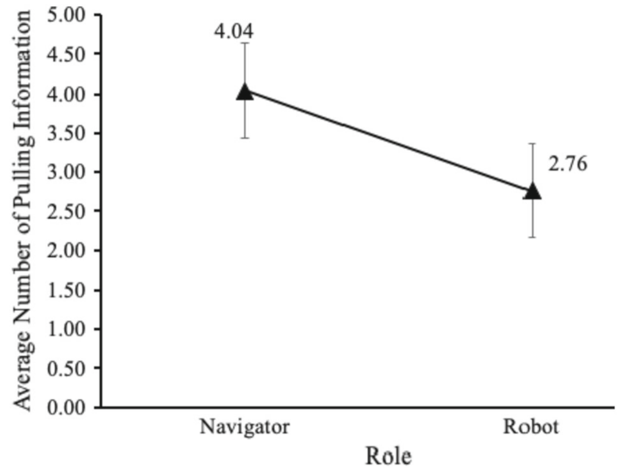

Fig. 8 Plot of information pulled by role (means with 95% confidence intervals)

0.555, $\eta_p^2 = 0.002$], nor was the condition by mission interaction effect [$F(3, 224) = 0.58$, $MSe = 6.41$, $p = 0.63$,$\eta_p^2 = 0.008$].

In accordance with the significant condition main effect, pairwise comparisons (LSD) indicate that the *pull prime* teams pulled more information than the *always explain* ($p < 0.001$), *explain if asked* ($p < 0.001$), and *never explain* teams ($p = 0.003$; see Fig. 7). Meanwhile, the *always explain* teams pulled significantly less information than the remaining three conditions (*never explain*, *explain if asked*, $p = 0.043$; and *always explain*, $p < 0.001$). Teams in the *explain if asked* and *never explain* conditions did not differ in the amount of information pulled ($p = 0.784$).

According to the significant role (within mission) main effect, navigators generally pulled more information from the robot than the robot pulled from the navigators (based on LSD, $p < 0.001$, see Fig. 8). This finding was somewhat expected, based on the team task design, as the navigator's role updating the map required pulling information from the robot.

The significant interaction effect of condition by role (within mission) and pairwise comparisons (LSD) further support the finding that *pull prime* navigators pulled the most information compared to the navigator in the other conditions ($p < 0.001$; Fig. 9) and also pulled more information from the robot than the robot pulled from them ($p < 0.001$). In contrast, the *always explain* navigators pulled the least amount of information compared to the navigator in the other conditions ($p < 0.001$), and also seemed to be the only condition in which the navigator pulled less information than the robot. The navigators in the remaining two conditions, *explain if asked* and *never explain*, did not differ in terms of the amount of information they pulled ($p = 0.892$). Across all four conditions, the robot pulled similar amounts of information. This indicates that our wizard-of-oz robots pulled information from participants relatively consistently across all conditions, to the benefit of experimental control.

4.2.2 Pushing Information

Pushing information refers to any information sent by a team member, before being asked for that information. Generally, it indicates anticipatory team communication. As in the previous analysis, pushing of information was analyzed via a repeated measure three-factor mixed Analysis of Variance (ANOVA) with condition as a between-teams manipulation, and mission with role nested as within-team factors.

Results show a significant condition main effect [$F(3, 224) = 3.86$, $MSe = 110$, $p = 0.010$,$\eta_p^2 = 0.05$], mission main effect [$F(1, 224) = 5.17$, $MSe = 148$, $p = 0.024$, $\eta_p^2 = 0.02$], and role (within mission) main effect [$F(2, 224) = 499$, $MSe = 14{,}276$, $p < 0.001$, $\eta_p^2 = 0.82$]. There were no significant interaction effects of condition by role (within a mission) [$F(6, 224) = 0.44$, $MSe = 12.5$, $p = 0.854$, $\eta_p^2 = 0.01$] or condition by mission [$F(3, 224) = 0.12$, $MSe = 3.32$, $p = 0.951$,$\eta_p^2 = 0.01$].

In accordance with the significant condition main effect, pairwise comparisons (LSD) indicate that teams in the *always explain* and *pull prime* conditions pushed similar amounts of information ($p = 0.62$), which was more information than teams in the *never explain* ($p = 0.027$ and $p = 0.007$, respectively) and *explain if asked* conditions ($p = 0.043$ and $p = 0.012$, respectively; Fig. 10, left), which also pushed similar amounts of information ($p = 0.845$).

In accordance with the significant mission main effect, pushing information increased from Mission 1 to 2 ($p = 0.024$; Fig. 10, right), likely due to the dynamic events involved in Mission 2.

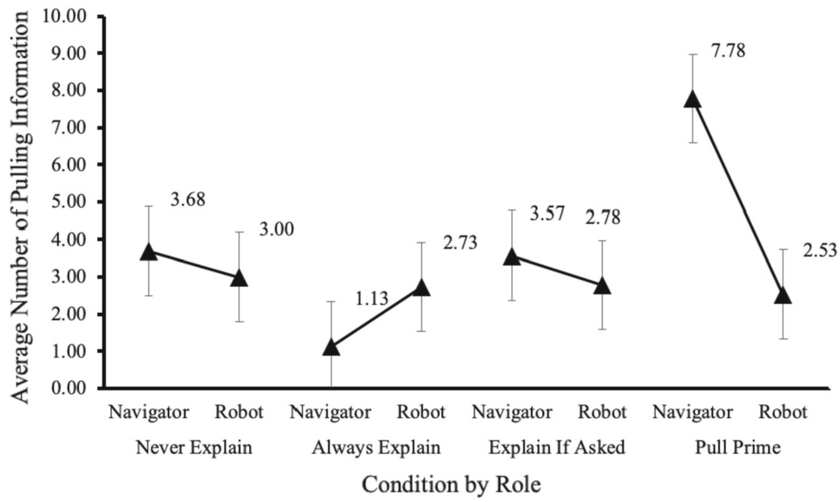

Fig. 9 Plot of information pulled by condition and role (means with 95% confidence intervals)

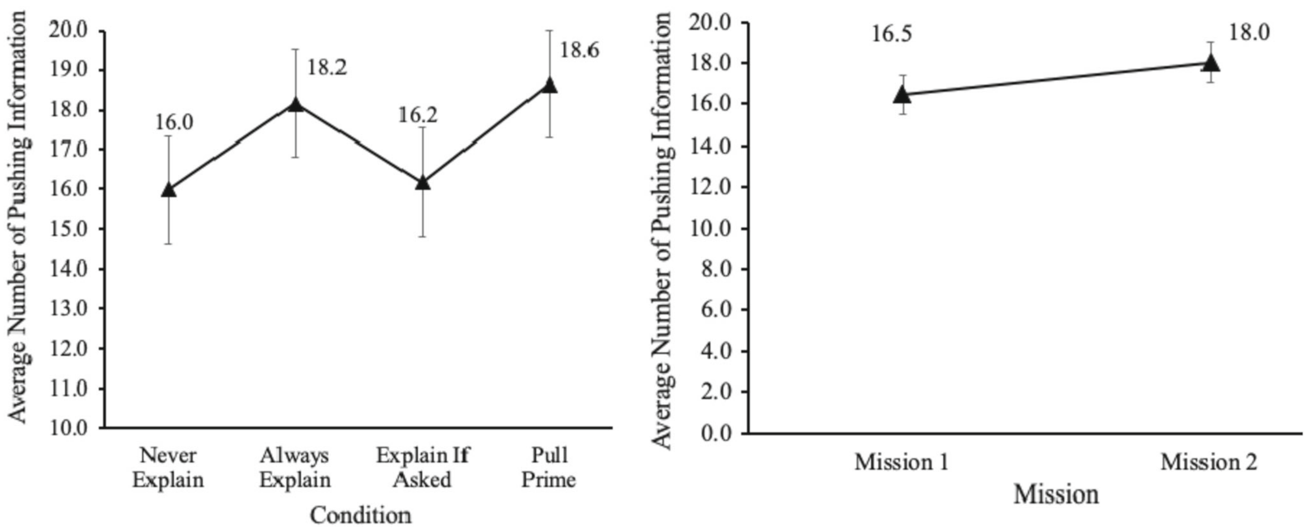

Fig. 10 Plot of information pushed by condition (left) and by mission (right); means with 95% confidence intervals

According to the significant role (within a mission) main effect, we can also confirm that the robot pushed more information than the navigator overall ($p<0.001$, Fig. 11), as was expected given their respective roles in the team task.

Even without a significant condition by role interaction effect, we still examined how each role differed across conditions. As seen in Fig. 12, the navigator did not differ in the amount of information they pushed across conditions. However, results indicate that the robot pushed significantly more information in the *always explain* ($p=0.021$) and *pull prime* conditions ($p=0.010$) compared to the *never explain* condition. This means that the robot's behaviors were inconsistent in pushing information across conditions, and a confounding factor because pushing information was unsolicited behavior.

As a visual summary of our findings, we plotted the mean frequencies of team communication behaviors across conditions (Fig. 13). This figure interpreted alongside the earlier statistical tests shows that although our robot was incon-

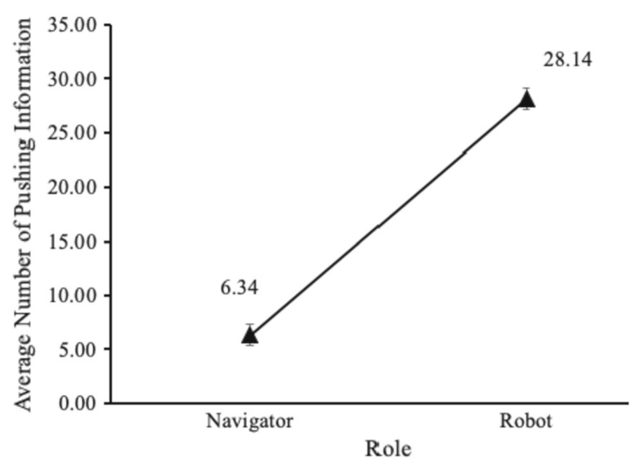

Fig. 11 Pushing information by role (means with 95% confidence intervals)

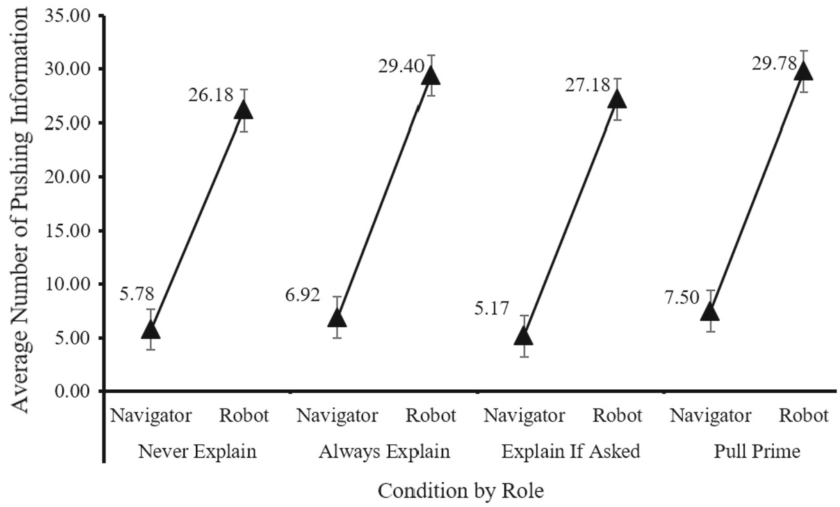

Fig. 12 Pushing information by condition and role (means with 95% confidence intervals)

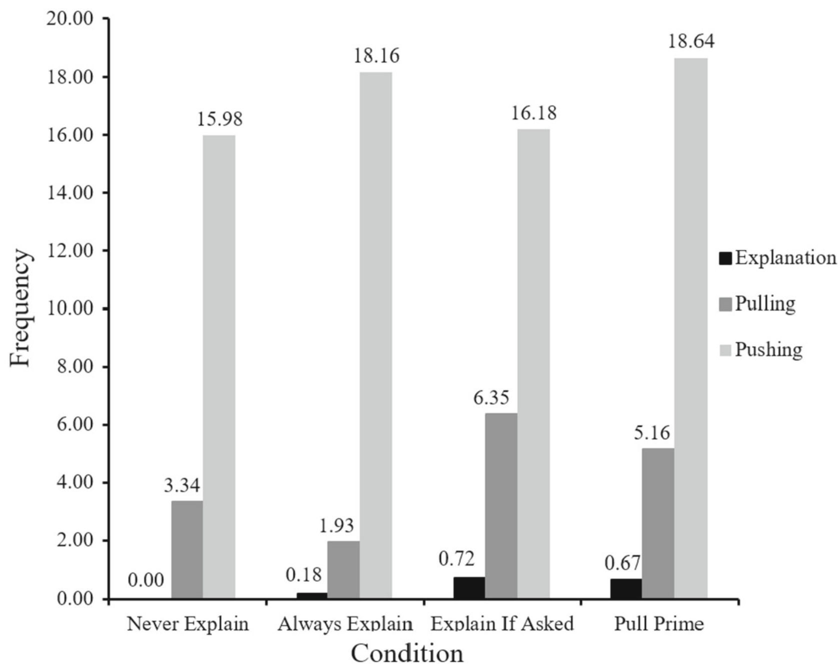

Fig. 13 Team communication behaviors across conditions (mean frequency of explanation, information pulling, and information pushing messages)

sistent in pushing information across conditions, the more critical variable to inspect is the number of explanations the robot provided. Given how we coded the data, the number of explanations provided also indicate the number of explanations requested.

As previously noted, participants in the *explain if asked* and *pull prime* conditions asked more "why" questions compared to *always explain*. Concurrently, *pull prime* did not lead to more "why" questions – rather, it led to other questions (i.e., pulling more information) such as requests for status updates and location. This additional communication in *pull prime* did not improve the team's shared situation awareness (e.g., Fig. 4).

Next, we discuss what variables could be the most useful predictors for team performance and shared situation awareness in this task.

4.3 Stepwise regression

4.3.1 Predicting performance

We applied stepwise regression (Akaike criteria—AIC) to predict team performance (i.e., the proportion of accurately triaged victims). The model accounted for 20.7% of the variance ($F(6, 113) = 4.92$, $p < 0.001$). According to these findings, a moderate level of robot explanations and pushing information were positively related to team performance in this task (based on the model's significant linear and

Table 4 Predicting team performance

Variable	Term	β	$SE\ \beta$	t	p
Robot explanations	Linear	0.46	0.22	2.08	0.039
	Quadratic	− 0.23	0.10	− 2.41	0.017
Robot pushing information	Linear	− 0.07	0.02	− 2.89	0.005
	Quadratic	0.01	0.00	2.94	0.004
Navigator pushing information	Linear	0.09	0.02	3.55	0.001
Robot pulling information	Linear	0.15	0.07	2.06	0.042

quadratic terms; Table 4). In addition, the model indicates that the navigator's pushing and the robot's pulling contributed to higher team performance.

4.3.2 Predicting Shared Situation Awareness.

We applied another stepwise regression (AIC) to predict shared situation awareness (i.e., the proportion of correctly identified collapses and openings on the map). The regression model accounted for 14.6% of the variance ($F(5, 114) = 3.89$, $p = 0.002$; Table 5). According to significant linear and quadratic terms, a moderate level of robot explanations and robot pushing information contributed to better shared situation awareness (Table 5).

5 Discussion

Our results indicate that in a simulated USAR human–robot team task environment, a robot that (1) communicates task-relevant information proactively, including potential explanations for unplanned behaviors while (2) allowing a human counterpart to request additional explanations as needed, strikes the best balance among the team communication strategies tested for supporting team performance and shared situation awareness. Teams in the *always explain* condition experienced a moderate level of robot explanations, which were coded as responses to the navigator's "why" questions. This moderate level of requests that received an explanation was associated with improved shared situation awareness relative to the other conditions, particularly *pull prime*, in which the robot would only respond to requests for explanations rather than provide this same information in advance.

Although the *pull prime* robot engaged in the same strategy as the *explain if asked* robot, the difference was that participants in *pull prime* were trained to initiate communication with the robot by asking it questions. However, this training did not lead to asking for more explanations. Instead, it led to asking for more status updates and other task-relevant types of information. These findings merit further discussion on the potential tradeoffs of communication in future human–robot USAR teams.

5.1 Tradeoffs of Communication Frequency and Explanations

Pull prime resulted in the highest frequency of team communication–primarily by the *robot pushing* and *navigator pulling* more information–which did not improve team performance relative to the other conditions. What contributed to higher team performance was the *navigator pushing* and the *robot pulling* information. This is interesting because the navigator's role was more suited to information pulling and the robot's role was more suited to information pushing. It seems that in dynamic task environments, even when task roles may demand a particular directionality of information (i.e., either pushing *or* pulling), having more bi-directional communication enhances team performance. This complements and supports prior research that found team performance is best predicted by team members spending equal amounts of time communicating during team meetings [76], and that a team's shared situation awareness in dynamic settings is best supported by moderate levels of communication (not too much and not too little) [37].

This leads to the insight that higher communication frequency does not necessarily translate to higher team performance and shared situation awareness, especially when the types of information (e.g., pushing and pulling) are highly unequal or highly unilateral (e.g., one team member is communicating much more than the other, or the communication roles are more rigid). A previous dynamical systems analysis of team behaviors found that by pushing and pulling information in a timely and constructive manner on a three-agent team, a single agent was able to affect team dynamics in a way that was positively associated with team performance and situation awareness. The study also found that, in a victim location task similar to this study, the best performing human dyads had moderately structured communication patterns (neither completely random, nor completely deterministic), suggesting that a balance between overly structured (i.e., rigid) and completely random (i.e., flexible) communication patterns results in the most effective team dynamics [77]. This also indicates, as did our study, that moderate explanations may provide the flexibility needed for teams in dynamic settings, and indicate team effectiveness.

Table 5 Predicting shared situation awareness

Variable	Term	β	$SE\ \beta$	t	p
Robot explanations	Linear	8.65	2.84	3.05	0.003
	Quadratic	− 3.80	1.22	− 3.11	0.002
Robot pushing information	Linear	0.54	0.26	2.10	0.038
	Quadratic	− 0.05	0.03	− 1.98	0.051
Navigator pulling information	Linear	0.37	0.26	1.46	0.147

In terms of design implications, these results suggest that the ability to ask for explanations and the relative quantity of explanations are important but do not fully determine a team's situation awareness and performance. Providing anticipatory explanations, as in the *always explain* condition, may better allow people to focus more on their other tasks at hand, such that they only need to request explanations due to unexpected needs not known by the robot in advance. Therefore, robots should be designed to communicate proactively, including identifying potential needs for explanations and then providing them, but to also allow participants to request for additional explanations where needed. This aligns with recent theoretical literature on trusting automation, which recommends designing not just for reliability but also for responsivity when it comes to AI-enabled robots meant for human teaming [12].

5.2 *Implications for* Trust and Workload

In this study, we also examined trust as a potential outcome of the various explanation-based strategies, but trust in the robot did not differ across the four conditions or two missions. A possible reason is that the number of explanations experienced in this study were relatively low compared to other types of communication, such as requesting and reporting status updates. Furthermore, participants may not have had time to reflectively calibrate their trust following robot explanations, due to the dynamic and demanding task environment, and relatively short exposure (under an hour) working with the robot. That participants reported high workload, or an average rating of 24.5 out of 30, generally supports this argument.

We also revisit the finding that trust in the robot teammate and perceived task workload were correlated. Although not direct outcomes of interest, it makes sense that perceptions of the robot's reliability (i.e., trustworthiness) would be related to perceptions of workload – the more workload one experiences, the more likely one may feel the need to rely on a teammate. However, this correlation was moderate, and trust did not increase between missions whereas workload did increase. It is possible that if communication is seen as a necessary part of teaming [23], then perhaps additional communication would simply crowd out other important task activities rather than add to the perceived task load. This is supported by participants rating higher workload in Mission 2 than Mission 1, a result of the dynamic events, rather than a result of the communication conditions. With a moderate number of explanations and an increase of information pushing (primarily by the robot), higher team performance was also achieved in Mission 2 despite the additional perceived workload.

A couple other reasons for the lack of differences in trust ratings relate to the explanations themselves and whether they were contrastive enough to impact trust. Dynamic events in the task environment allowed the robot to engage in off-plan behaviors that would ostensibly motivate a need for explanation. However, taking a shorter path due to a new opening, or rerouting due to a blockage may be non-controversial behaviors, especially if participants could mostly see what the robot saw through their shared video feed (e.g., noticing a blockage as the robot changed paths). Because the robot's behavior was an apparent response due to changes in the environment, the need for an explanation may have been limited to when participants happened to miss seeing the reason for the deviation in the video feed. Presuming the video feed facilitated a shared mental model of the robot's decision-making [24], this would make the need for explanations less critical. This is supported by the finding that the *always explain* condition did not significantly differ from the *never explain* situation in terms of predicting shared situation awareness, despite the binary distinction that one condition allowed for robot explanations and the other did not.

Last but not least, another reason for the lack of differences in trust ratings could be that our team task environment did not faithfully represent the type of task complexity that merits explanations, given that the team's goals were fairly well aligned. One definition of complexity focuses on the extent to which individual goals require tradeoffs with the team's shared goals [12], which differs from the type of team task environment that is defined more by the number of interdependent agents involved, or the dynamism of the task environment and interactions [20, 22, 37, 38, 77]–[79]. The latter definition of environmental complexity is essentially a coordination problem that benefits more from the strategic allocation of work or attention, rather than a cooperation problem that would benefit from the social effects of explanations [15]. As such, this study indicates that explanations may be more critical in task environments where there is higher potential for goal misalignment.

5.3 Future Directions

There were several limitations of this study, including the challenge of measuring anticipatory explanations. Because explanations are an inherently sequential concept – meaning a person must have first had a thought before an explanation can have its purported impact by providing a contrastive example to that initial thought – this makes designing for "anticipatory explanations" in dynamic or emergent task environments difficult. Anticipatory explanations are somewhat of an oxymoron because explanations can only have their effect in a post hoc way. Alternatively, there must be robust evidence that many people would interpret a particular situation in a similar way that would then require an explanation or specific set of explanations. Given the challenges of unobtrusively capturing what a person was thinking while engaging in this task environment, it is possible that our coded data underestimates the actual number of explanations that occurred in the *always explain* condition, because we were unable to capture information that the robot automatically provided that might have answered an unverbalized "why" question. We did not conduct a prior study to determine which scenarios in our task environment would require explanations for the *always explain* condition, and instead relied on the research team's assumption that explanations would be needed when the robot deviated from its original plan.

Another limitation of this study is the common challenge of translating results from a participant sample comprising mostly university students to experts in a particular domain area. The Minecraft environment was intentionally designed [43, 55] to meet our research objective of advancing theory on human–robot team communication strategies, and was also tailored to the participant sample we knew we were likely to draw. In other words, experience playing Minecraft, urban search and rescue, and human–robot teaming was not necessary to succeed in completing the team task. Nevertheless, there was some variation across the study conditions with respect to self-reported Minecraft experience (see Sect. 3.1). Therefore, it is possible that participants' perceptions of the robot, and their subsequent communication behaviors, were influenced by what they imagined or expected the robot to be going into the study. Follow up studies that focus specifically on the relationship between prior experience and team communication in the domain area of interest would be merited.

Nevertheless, we still believe that the results from this study have added insight – and lessons learned – for others conducting similar studies for future human–robot teaming in USAR environments. Determining when explanations are needed in real-world team task environments may be an important primer for future work in this area – an insight that is not currently addressed in human–robot USAR communication studies, given their necessary but limited focus on conducting studies with existing robot capabilities [26, 80].

6 Conclusion

The primary goal of this research was to characterize how different explanation-based communication strategies of a human–robot team impacted their team's effectiveness, which included assessments of team performance, shared situation awareness, trust, and workload in a simulated urban search and rescue task. Because team communication is typically an interdependent sequential process–meaning that behaviors of one teammate impact the behaviors of another teammate, which impact the behaviors of the other teammate, and so on – this study took a multifaceted, descriptive approach to the experimental data. Interpreted as a whole, our findings support the following practical implications about future human–robot team communication in USAR tasks:

(1) Priming people to initiate interactions with a responsive robot can lead to more frequent communication overall during subsequent exchanges – this contrasts with previous experience in which study participants tended to take on more passive communication roles while interacting with autonomous technology.
(2) More frequent communication does not necessarily support team performance, shared situation awareness, or trust – particularly when communication is too frequent and conflicts with human sensory modalities (i.e., visual, manual) that are relevant to performing other ongoing team task activities.
(3) To help support shared situation awareness, the function of a robot teammate should include understanding how to reduce communication overhead by pushing relevant information to its human counterpart, including anticipating potential situations that may need explanation, and then providing them.
(4) Positioning robots as high-performing, autonomous teammates in dynamic task environments demands the ability to communicate bi-directionally within the team, as well as more equally distributed and moderate levels of pushing and pulling information, to support unanticipated information needs and shared situation awareness.
(5) Explanations may not have much impact on trust in pure team coordination tasks, i.e., team tasks that do not involve potential goal tradeoffs that require inter-agent cooperation.

USAR operators typically deal with high levels of uncertainty, influxes of information, changing conditions, and critical time constraints. Although the high workload of our simulated team task environment bodes positively for its ecological validity, our study also illuminates several ongoing challenges for real-world application of human–robot teams working in complex and time-constrained environments. These challenges include (1) the need to obtain and to filter the quality of information exchanged [81], (2) maintaining shared situation awareness through communication without performance degradation in other tasks, and (3) moving beyond rigid roles and function allocation within human–robot teams toward designing for more dynamic and flexible teaming.

Acknowledgements The authors thank the many CERTT Lab students for their assistance with developing the task environment, data collection, and data pre-processing. The manuscript was improved thanks to comments from two anonymous reviewers.

Author Contributions E.K.C. and M.D. designed the study, analyzed the data, discussed and interpreted the results, wrote the manuscript; V.B. performed the study and wrote the manuscript; C.C.C. performed the study and wrote the manuscript; M.R.E. designed the study and wrote the manuscript; G.J.L. designed and performed the study, and wrote the manuscript; N.J.C. conceived and designed the study, discussed and interpreted the results; N.J.M designed the study, discussed and interpreted the results. Author order is listed primarily based on direct contributions to the final written document; all authors reviewed the manuscript.

Funding This work was supported by the Air Force Office of Scientific Research (FA9550-18–1-0067). Any findings or recommendations expressed in this material are those of the authors and do not necessarily reflect the views or policies of the research sponsor.

Declarations

Conflict of interest The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Data Availability The datasets generated and analyzed for the current study are not publicly available due to participant privacy protection measures, but are available from the project principal investigator (NC) upon reasonable request.

Ethics Approval This study was approved by Arizona State University's IRB for human subjects research in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki and its later amendments.

Consent to Participate All participants included in this study provided their informed consent.

Consent to Publish As part of the consent process, all participants were informed that their data in this study may be published in aggregate form.

References

1. Murphy RR (2004) Trial by fire. IEEE Robot Autom Mag 11(3):50–61. https://doi.org/10.1109/MRA.2004.1337826
2. Guizzo E (2011) Japan earthquake: More robots to the rescue. In: IEEE Spectrum. https://www.spectrum.ieee.org/japan-earthquakemore-robots-to-the-rescue. Accessed 27 Oct 2021
3. Murphy RR, Steimle E, Griffin C, Cullins C, Hall M, Pratt K (2008) Cooperative use of unmanned sea surface and micro aerial vehicles at Hurricane Wilma. J Field Robot 25(3):164–180. https://doi.org/10.1002/rob.20235
4. Fernandes O, Murphy R, Merrick D, Adams J, Hart L, Broder J (2019) Quantitative data analysis: small unmanned aerial systems at Hurricane Michael. 2019 IEEE int Symp Saf Secur Rescue Robot SSRR 2019:116–117. https://doi.org/10.1109/SSRR.2019.8848935
5. Liu Y, Nejat G (2013) Robotic urban search and rescue: a survey from the control perspective. J Intell Robot Syst Theory Appl 72(2):147–165. https://doi.org/10.1007/s10846-013-9822-x
6. Murphy RR (2004) Human-robot interaction in rescue robotics. IEEE Trans Syst Man Cybern Part C Appl Rev 34(2):138–153. https://doi.org/10.1109/TSMCC.2004.826267
7. Salas E, Cooke NJ, Rosen MA (2008) On teams, teamwork, and team performance: discoveries and developments. Hum Factors J Hum Factors Ergon Soc 50(3):540–547. https://doi.org/10.1518/001872008X288457
8. Phillips E, Ososky S, Swigert B, Jentsch F (2012) Human-animal teams as an analog for future human-robot teams. Proc Hum Factors Ergon Soc 56:1553–1557. https://doi.org/10.1177/1071181312561309
9. Malone T, Crowston K (1994) The interdisciplinary study of coordination. ACM Comput Surv 26(1):87–119. https://doi.org/10.1145/174666.174668
10. Cooke NJ, Gorman JC, Myers CW, Duran JL (2013) Interactive team cognition. Cogn Sci 37(2):255–285. https://doi.org/10.1111/cogs.12009
11. Salas E, Wilson K, Murphy C, King H, Salisbury M (2008) Communicating, coordinating, and cooperating when lives depend on it: tips for teamwork. Jt Comm J Qual Patient Saf Jt Comm Resour 34(6):333–341
12. Chiou EK and Lee JD (2021) Trusting automation: Designing for responsivity and resilience Hum. Factors 10/gjvcr2
13. Murphy RR, Burke JL (2005) Up from the rubble: lessons learned about HRI from search and rescue. Proc Hum Factors Ergon Soc Annu Meet 49(3):437–441
14. de Visser EJ, Pak R, Shaw TH (2018) From 'automation' to 'autonomy': the importance of trust repair in human–machine interaction. Ergonomics 61(10):1409–1427. https://doi.org/10.1080/00140139.2018.1457725
15. Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. Artif Intell 267:1–38. https://doi.org/10.1016/j.artint.2018.07.007
16. Chakraborti T, Sreedharan S, Grover S, Kambhampati S (2019) Plan explanations as model reconciliation -- An empirical study," in 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Mar 2019, pp 258–266 https://doi.org/10.1109/HRI.2019.8673193
17. Hancock PA et al (2011) A meta-analysis of factors affecting trust in human-robot interaction. Hum Factors J Hum Factors Ergon Soc 53(5):517–527. https://doi.org/10.1177/0018720811417254
18. Williams T, Briggs P, Scheutz M (2015) Covert robot-robot communication Human perceptions and implications for human-robot interaction. J Hum-Robot Interact 4(2):24–49

19. Endsley MR (2017) From here to Autonomy: lessons learned from human-automation research. Hum Factors J Hum Factors Ergon Soc 59(1):5–27
20. Cooke NJ, Gorman JC, Duran JL, Taylor AR (2007) Team cognition in experienced command-and-control teams. J Exp Psychol Appl 13(3):146–157. https://doi.org/10.1037/1076-898X.13.3.146
21. Demir M, McNeese NJ, Cooke NJ (2016) Team communication behaviors of the human-automation teaming, in 2016 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), Mar. 2016, pp. 28–34 https://doi.org/10.1109/COGSIMA.2016.7497782
22. McNeese NJ, Demir M, Cooke NJ, Myers C (2018) Teaming with a synthetic teammate: insights into human-autonomy teaming. Hum Factors 60(2):262–273. https://doi.org/10.1177/0018720817743223
23. Burke JL and Murphy RR (2004) Human-robot interaction in USAR technical search: two heads are better than one, in RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759), Sep 2004, pp 307–312 https://doi.org/10.1109/ROMAN.2004.1374778
24. Burke J, Murphy R (2007) RSVP: an investigation of remote shared visual presence as common ground for human-robot teams, in 2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI), Mar 2007, pp 161–168 10/dh76b4
25. Nourbakhsh IR, Sycara K, Koes M, Yong M, Lewis M, Burion S (2005) Human-robot teaming for search and rescue. IEEE Pervasive Comput 4(1):72–79
26. Burke JL, Murphy RR, Riddle DR, Fincannon T (2004) Task performance metrics in human-robot interaction: taking a systems approach," UNIVERSITY OF SOUTH FLORIDA TAMPA CENTER FOR ROBOT-ASSISTED SEARCH AND RESCUE, Aug. 2004. Accessed: May 19, 2021. [Online]. Available: https://apps.dtic.mil/sti/citations/ADA516069
27. Cooke NJ, Demir M, Huang L (2020) A framework for human-autonomy team research. Lect Notes Comput Sci Subser Lect Notes Artif Intell Lect Notes Bioinforma. https://doi.org/10.1007/978-3-030-49183-3_11
28. MacMillan J, Entin EE, Serfaty D (2005) Communication overhead: the hidden cost of team cognition Team Cogn Underst Factors Drive Process Perform 61–82 https://doi.org/10.1037/10690-004
29. Marks MA, Mathieu JE, Zaccaro SJ (2001) A temporally based framework and taxonomy of team processes. Acad Manag Rev 26(3):356. https://doi.org/10.2307/259182
30. Marlow SL, Lacerenza CN, Paoletti J, Burke CS, Salas E (2018) Does team communication represent a one-size-fits-all approach?: a meta-analysis of team communication and performance. Organ Behav Hum Decis Process 144:145–170. https://doi.org/10.1016/j.obhdp.2017.08.001
31. Endsley MR, Jones WM (2001) A model of inter and intra team situation awareness: Implications for design, training and measurement. In McNeese M, Salas E, Endsley M (eds) New Trends in Cooperative Activities: Understanding System Dynamics in Complex Environments. Human Factors and Ergonomics Society 2001, Santa Monica CA, p 25
32. Crozier MS et al (2015) Use of human patient simulation and validation of the team situation awareness global assessment technique (TSAGAT): a multidisciplinary team assessment tool in trauma education. J Surg Edu 72(1):156–163. https://doi.org/10.1016/j.jsurg.2014.07.009
33. Gardner AK, Kosemund M, Martinez J (2017) Examining the feasibility and predictive validity of the SAGAT tool to assess situation awareness among medical trainees. Simul Healthc 12(1):17–21. https://doi.org/10.1097/SIH.0000000000000181
34. Bonney L, Davis-Sramek B, Cadotte ER (2016) 'Thinking' about business markets: a cognitive assessment of market awareness. J Bus Res 69(8):2641–2648. https://doi.org/10.1016/j.jbusres.2015.10.153
35. Cooke NJ, Kiekel PA, Helm EE (2001) Measuring team knowledge during skill acquisition of a complex task. Int J Cogn Ergon 5(3):297–315. https://doi.org/10.1207/s15327566ijce0503_10
36. Coolen E, Draaisma J, Loeffen J (2019) Measuring situation awareness and team effectiveness in pediatric acute care by using the situation global assessment technique. Eur J Pediatr 178:837–850. https://doi.org/10.1007/s00431-019-03358-z
37. Demir M, McNeese NJ, Cooke NJ (2017) Team situation awareness within the context of human-autonomy teaming. Cogn Syst Res 46:3–12. https://doi.org/10.1016/j.cogsys.2016.11.003
38. Gorman JC, Cooke NJ, Winner JL (2006) Measuring team situation awareness in decentralized command and control environments. Ergonomics 49(12–13):1312–1325. https://doi.org/10.1080/00140130600612788
39. Drury JL, Scholtz J, Yanco HA (2003) Awareness in human-robot interactions. IEEE Syst Man Cybern 1(October):912–918. https://doi.org/10.1109/ICSMC.2003.1243931
40. Riley JM, Strater LD, Sethumadhavan A, Davis F, Tharanathan A, Kokini C (2008) Performance and situation awareness effects in collaborative robot control with automation. Proc Hum Factors Ergon Soc 1:242–246. https://doi.org/10.1177/154193120805200410
41. Yanco HA and Drury J (2004) "Where am I?' Acquiring situation awareness using a remote robot platform," Conf Proc - IEEE Int Conf Syst Man Cybern 3: 2835–2840, 2004, doi: https://doi.org/10.1109/ICSMC.2004.1400762
42. Endsley MR, "Design and evaluation for situation awareness enhancement," in Proceedings of the Human Factors Society Annual Meeting, Oct. 1988, vol. 32, no. 2, pp. 97–101. doi: https://doi.org/10.1177/154193128803200221
43. Bartlett CE and Cooke NJ (2015) Human-robot teaming in urban search and rescue Proc Hum Factors Ergon Soc 2015: 250–254 https://doi.org/10.1177/1541931215591051
44. Demir M, McNeese NJ, Cooke NJ (2020) Understanding human-robot teams in light of all-human teams: aspects of team interaction and shared cognition. Int J Hum Comput Stud 140:102436. https://doi.org/10.1016/j.ijhcs.2020.102436
45. Gatsoulis Y, Virk GS, Dehghani-Sanij AA (2010) On the measurement of situation awareness for effective human-robot interaction in teleoperated systems. J Cogn Eng Decis Mak 4(1):69–98. https://doi.org/10.1518/155534310X495591
46. Sarter NB, Woods DD, and Billings CE (1997) Automation surprises," in Handbook of Human Factors & Ergonomics, 2nd ed, G Salvendy, Ed Wiley https://doi.org/10.1109/VTSA.2003.1252543
47. Burke JL, Murphy RR, Coovert MD, Riddle DL (2004) Moonlight in Miami: field study of human-robot interaction in the context of an urban search and rescue disaster response training exercise. Human-Computer Interact 19(1–2):85–116. https://doi.org/10.1080/07370024.2004.9667341
48. Chakraborti T, Sreedharan S, Grover S, and Kambhampati S (2018) Plan explanations as model reconciliation -- an empirical study
49. Hoffman RR, Mueller ST, Klein G, and Litman J (2018) Metrics for explainable AI: challenges and prospects pp 1–50
50. Salas E, Prince C, Baker DP, Shrestha L (1995) Situation awareness in team performance: implications for measurement and training. Hum Factors 37(1):123–136. https://doi.org/10.1518/001872095779049525
51. Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug 2016, vol 13–17-Augu, pp 1135–1144 https://doi.org/10.1145/2939672.2939778
52. E. J. de Visser, R. Pak, and M. A. Neerincx, "Trust development and repair in human-robot teams," *ACMIEEE Int. Conf. Hum.-*

Robot Interact., pp. 103–104, 2017, doi: https://doi.org/10.1145/3029798.3038409.
53. Groom V, Nass C (2007) Can robots be teammates?: benchmarks in human-robot teams. Interact Stud 8(3):285–301. https://doi.org/10.1075/gest.8.3.02str
54. Gonzalez C, Vanyukov P, Martin MK (2005) The use of microworlds to study dynamic decision making. Comput Hum Behav 21(2):273–286. https://doi.org/10.1016/j.chb.2004.02.014
55. Lematta GJ et al (2019) Developing human-robot team interdependence in a synthetic task environment. Proc Hum Factors Ergon Soc Annu Meet 63(1):1503–1507. https://doi.org/10.1177/1071181319631433
56. Kelley JF (1983) An empirical methodology for writing user-friendly natural language computer applications, in Proceedings of the SIGCHI conference on Human Factors in Computing Systems - CHI '83, Boston, Massachusetts, United States, pp 193–196 10/dcdp3z
57. Demir M and Cooke NJ (2014) Human teaming changes driven by expectations of a synthetic teammate Proc Hum Factors Ergon Soc 2014 pp 16–20 https://doi.org/10.1177/1541931214581004
58. Parasuraman R, Miller CA (2004) Trust and etiquette in high-criticality automated systems. Commun ACM 47(4):51–55
59. Chiou EK, Lee JD, Su T (2019) Negotiated and reciprocal exchange structures in human-agent cooperation. Comput Hum Behav 90(August):288–297. https://doi.org/10.1016/j.chb.2018.08.012
60. Xu J and Howard A (2018) The impact of first impressions on human- robot trust during problem-solving scenarios RO-MAN 2018 - 27th IEEE Int Symp Robot Hum Interact Commun pp 435–441, 2018 https://doi.org/10.1109/ROMAN.2018.8525669
61. Faul F, Erdfelder E, Lang A-G, Buchner A (2007) G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 39(2):175–191
62. J. Cohen, *Statistical Power Analysis for the Behavioral Sciences - 1st Edition*, 1st ed. Academic Press, 1977. Accessed: May 20, 2021. [Online]. Available: https://www.elsevier.com/books/statistical-power-analysis-for-the-behavioral-sciences/cohen/978-0-12-179060-8
63. Cahour B, Forzy JF (2009) Does projection into use improve trust and exploration? an example with a cruise control system. Saf Sci 47(9):1260–1270. https://doi.org/10.1016/j.ssci.2009.03.015
64. Jian J-Y, Bisantz AM, Drury CG (2000) Foundations for an empirically determined scale of trust in automated systems. Int J Cogn Ergon 4(1):53–71. https://doi.org/10.1207/S15327566IJCE0401_04
65. Madsen M and Gregor S "Measuring human-computer trust," in Proceedings of Eleventh Australasian Conference on Information Systems, 2000, pp 6–8
66. Hart SG and Staveland LE (1988) Development of NASA-TLX (Task Load Index): results of empirical and theoretical research," in Advances in Psychology, vol 52, C, P A Hancock and N Mashkati, Eds Amsterdam: North Holland Press, 1988, pp 139–183 https://doi.org/10.1016/S0166-4115(08)62386-9
67. Hart SG (2006) Nasa-task load index (NASA-TLX); 20 years later. Proc Hum Factors Ergon Soc Annu Meet 50(9):904–908. https://doi.org/10.1177/154193120605000909
68. Howell DC (2021) Statistical Methods for Psychology, 8th ed Wadsworth Publishing, 2011 Accessed: May 20 [Online] Available: /c/statistical-methods-for-psychology-8e-howell/9780357670996PF
69. Meier U (2006) A note on the power of Fisher's least significant difference procedure. Pharm Stat 5(4):253–263
70. "Downloading IBM SPSS Statistics 26," Dec. 07, 2020. https://www.ibm.com/support/pages/downloading-ibm-spss-statistics-26 (accessed May 20, 2021).
71. Weisberg S (2005) "Variable Selection," in Applied Linear Regression Wiley pp 211–232 https://doi.org/10.1002/0471704091.ch10
72. Berk RA (2008) Statistical learning from a regression perspective Springer New York https://doi.org/10.1007/978-0-387-77501-2
73. R: the R project for statistical computing https://www.r-project.org/ (accessed May 20, 2021)
74. Ripley B (2021) "Support functions and datasets for venables and Ripley's MASS [R package MASS version 7.3–54]," May 03, 2021. https://CRAN.R-project.org/package=MASS (accessed May 20, 2021)
75. Behrendt S (2014) "Add standardized regression coefficients to lm-Objects [R package lm.beta version 1.5–1]," Dec. 28, 2014. https://CRAN.R-project.org/package=lm.beta (accessed May 20, 2021)
76. Duhigg C (2016) What Google learned from its quest to build the perfect team, The New York Times Magazine, no. The Work Issue, Feb. 25
77. Demir M, Cooke NJ, Amazeen PG (2018) A conceptual model of team dynamical behaviors and performance in human-autonomy teaming. Cogn Syst Res 52:497–507
78. Gorman JC, Cooke NJ, Amazeen PG (2010) Training adaptive teams. Hum Factors J Hum Factors Ergon Soc 52(2):295–307
79. Gorman J, Amazeen P, Cooke N (2010) Team coordination dynamics. Nonlinear Dyn Psychol Life Sci 14:265–289
80. Casper J, Murphy RR (2003) Human-robot interactions during the robot-assisted urban search and rescue response at the World Trade Center. IEEE Trans Syst Man Cybern Part B Cybern 33(3):367–385
81. Arrow KJ (1974) The limits of organization. W. W. Norton & Company, USA