Check for updates

Communication Strategies in Human-Autonomy Teams During Technological Failures

Human Factors
2024, Vol. 66(11) 2539–2555
© 2024 Human Factors
and Ergonomics Society
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00187208231222119
journals.sagepub.com/home/hfs
Sage

Julie L. Harrison[1], Shiwen Zhou[2], Matthew J. Scalia[2], David A. P. Grimm[1], Mustafa Demir[2], Nathan J. McNeese[3], Nancy J. Cooke[2], and Jamie C. Gorman[2]

Abstract

Objective: This study examines low-, medium-, and high-performing Human-Autonomy Teams' (HATs') communication strategies during various technological failures that impact routine communication strategies to adapt to the task environment.

Background: Teams must adapt their communication strategies during dynamic tasks, where more successful teams make more substantial adaptations. Adaptations in communication strategies may explain how successful HATs overcome technological failures. Further, technological failures of variable severity may alter communication strategies of HATs at different performance levels in their attempts to overcome each failure.

Method: HATs in a Remotely Piloted Aircraft System-Synthetic Task Environment (RPAS-STE), involving three team members, were tasked with photographing targets. Each triad had two randomly assigned participants in navigator and photographer roles, teaming with an experimenter who simulated an AI pilot in a Wizard of Oz paradigm. Teams encountered two different technological failures, automation and autonomy, where autonomy failures were more challenging to overcome.

Results: High-performing HATs calibrated their communication strategy to the complexity of the different failures better than medium- and low-performing teams. Further, HATs adjusted their communication strategies over time. Finally, only the most severe failures required teams to increase the efficiency of their communication.

Conclusion: HAT effectiveness under degraded conditions depends on the type of communication strategies enacted by the team. Previous findings from studies of all-human teams apply here; however, novel results suggest information requests are particularly important to HAT success during failures.

Application: Understanding the communication strategies of HATs under degraded conditions can inform training protocols to help HATs overcome failures.

[1]Georgia Institute of Technology, USA
[2]Arizona State University, USA
[3]Clemson University, USA

Received: February 3, 2023; accepted: November 29, 2023

Corresponding Author:
Julie L. Harrison, Georgia Institute of Technology, Technology Square Research Building, Room 228, 85 5th St NW, Atlanta, GA 30308, USA; e-mail: julieharrison@gatech.edu

Keywords
levels of automation, perturbations, remotely piloted aircraft systems, team communication

Introduction

As algorithmic technologies advance, humans in complex work environments are required to team with higher levels of artificial intelligence (AI). When embedded in complex environments, technology failures in work systems are inevitable, such that as the system grows in complexity, it can fail in increasingly unpredictable ways. Thus, to measure team effectiveness, we must assess how teams overcome technological failures (Cooke et al., 2007). Of particular importance is how teams overcome automation and autonomy failures to guard against automation brittleness, lack of transparency, and increased workload due to system monitoring (Shively et al., 2018). Automation is "physical technology, including mechanized or computerized systems," applied in a defined environment to aid execution of a process. Autonomy refers to "a state of being" for mechanical agents with a degree of independence, discretion, and adaptability to their environment (Kaber, 2018, p. 407). Accordingly, efficiently overcoming failures associated with these technologies is an important marker of team effectiveness. Due to the importance of communication for team coordination (Chow et al., 2000; Cooke et al., 2013), the current study considers which team communication strategies (between human teammates amongst each other and their artificial counterparts) contribute to effective teamwork in Human-Autonomy Teams (HATs) when technologies fail.

Human-Autonomy Teams are systems of interactive subcomponents, consisting of human and technological variables (Gorman et al., 2019). Increasingly, team tasks involve reliance on nonhuman autonomous teammates (Demir & Cooke, 2014; McNeese et al., 2018; O'Neill et al., 2022). However, the natural language processing abilities of AI agents remain somewhat limited, which places a constraint on team coordination (Demir et al., 2017; National Academies of Sciences, 2021). Although recent models at the time of this writing, such as Chat GPT, reflect significant progress in language abilities, research shows that their capabilities are still not at a caliber to replace human teammates (Tenhundfeld & ChatGPT, 2023; Wardat et al., 2023). Here, we consider which communication strategies in HATs contribute to success in overcoming automation and autonomy failures. The present study assesses these constructs in the context of a Remotely Piloted Aircraft System-Synthetic Task Environment (RPAS-STE), where one team member is assumed by a Wizard of Oz (WoZ; Riek, 2012) "AI" teammate that fails in experimentally controlled ways.

Implicit Coordination

Team cognition—"the cognitive processes or activities that occur at a team level" (Cooke et al., 2013, p. 256)—allows for implicit coordination strategies (e.g., anticipation of information needs) to develop among team members (Cannon-Bowers & Salas, 1990; DeChurch & Mesmer-Magnus, 2010; Entin & Serfaty, 1999; Orasanu, 1990). Implicit coordination—a team's ability to coordinate across members without relying on overt communication—is important to team success during periods of high workload as it allows the team to reduce the cognitive resources spent on communication overhead (MacMillan et al., 2004). Though communication is a necessary component of team coordination, it requires time, effort, and attention; communication overhead is the cost of these resources demanded by team communication (MacMillian et al., 2004). Given that autonomous agents do not yet communicate at a level sufficient to fully replace a human teammate (Tenhundfeld & ChatGPT, 2023; Wardat et al., 2023), implicit coordination is relevant to HATs.

Implicit coordination strategies have been shown to distinguish high- from low-performing teams (Entin & Serfaty, 1999). By relying on implicit coordination strategies, high-reliability teams reduce communication overhead and shift attentional resources to other aspects of the task. Teams can adapt their communication strategy by pushing information (i.e., unprompted information relaying) more than pulling information (i.e., information requests from one teammate to another that are reciprocated

with information sharing; Entin & Serfaty, 1999; MacMillan et al., 2004). Increases in pushing, relative to pulling, signal the development of implicit coordination processes, such that successful teams anticipate needs, thereby engaging in more efficient communication when task load increases (Entin & Serfaty, 1999). The present study assesses the impact of pushing and pulling on HATs' ability to overcome unanticipated automation and autonomy failures. We posit that the rates of these specific communication strategies will differ during failure periods.

In studying observable team behaviors, the Theory of Interactive Team Cognition (ITC; Cooke et al., 2013) posits that we are directly studying team coordination and team cognition. The theory, which we adopt in the present study, assumes that real-time interactions amongst team members and the task environment are team cognition. In accordance with these assumptions, communication artifacts have historically been assessed to understand coordination processes in all-human teams (e.g., MacMillan et al., 2004) as well as HATs (e.g., Demir et al., 2017). Given the limited communicative abilities of many artificial agents, it is not surprising that HATs have been shown to push and pull information less than all-human teams (Demir et al., 2017), where like in all-human teams, in HATs, pushing and pulling behaviors are noted for *all* members of the team, including the autonomous teammates. This reduction in information exchange in HATs contributes to performance deficits due to weaker team coordination and reinforces the importance of pushing information in team coordination, as pushing across all teams—all-human and HATs—was more positively associated with team performance than pulling (Demir et al., 2017). These findings complement prior work with all-human teams, wherein high-performing teams exhibited more pushing than low-performing teams, allowing the high-performing teams to communicate and coordinate more efficiently while maintaining effectiveness (MacMillan et al., 2004).

The present study builds upon existing work by not only assessing the pushing and pulling patterns of HATs but also looking at these communication strategies during failure. In dynamic environments, teams must adapt their interactions to meet the coordination demands of the situation (Cooke et al., 2013; Gorman et al., 2006). HATs that more effectively adapt their coordination patterns in response to situational changes are more successful than those that demonstrate rigidity during novel periods (Grimm et al., 2018; Song et al., 2022). One way in which teams may change their behavior to respond to novelty is via updates to their communication strategy. Specifically, increases in information sharing may distinguish higher-performing from lower-performing HATs (Song et al., 2022). In the present study, we empirically explore this hypothesis to understand how high-, medium-, and low-performing HATs adjust their pushing and pulling strategies in response to technological failures.

The Current Study

In the RPAS-STE, three team members—a navigator, photographer, and "AI" pilot—coordinated to fly a simulated RPA to capture reconnaissance photos of targets. The task was comprised of several missions with automation and autonomy failures injected for specified targets. Automation failures affected the displays of the pilot and photographer, whereas autonomy failures impacted the ability of the "AI" agent to accurately respond to the developing situation. Human team members needed to recognize failures, then coordinate to overcome them in a limited amount of time. High-, medium-, and low-performing teams were differentiated through a cluster analysis, subsequently explained.

Research Question #1 (RQ1). Do pushing and pulling communication strategies for the three team performance clusters significantly differ between failure types?

Hypothesis #1 (H1). As failures grew in complexity, we expected higher performing teams to increase their pushing communications. Lower-performing teams were expected to exhibit a relatively low rate of pushing communications for all failure conditions. Pulling communications are less tied to performance (Demir et al., 2017); however, we expected less pulling to make way for more pushing in higher-performing teams as failures grew in complexity.

Research Question #2 (RQ2). Do teams learn to adjust their pushing and pulling strategies in response to failures over time?

Hypothesis #2 (H2). Teams would increase pushing over time as they learned to communicate more efficiently. Given the general increase in pushing communications, we predicted an associated reduction in pulling over time.

Research Question #3 (RQ3). How do changes in failure severity impact pushing and pulling communications?

Hypothesis #3 (H3). In addition to changes in pushing/pulling over time (H2), we expected that pushing/pulling would be impacted by failure complexity. As the failures grew in complexity, we expected pushing to increase and pulling to decrease during the most severe failures to increase communication efficiency.

Method

Participants

Twenty-two teams (44 participants) were recruited from Arizona State University and surrounding areas. Each team had two participants, a photographer and navigator. The pilot was an experimenter trained to mimic behaviors of an AI agent. Normal or corrected-to-normal vision and fluency in English were required. Ages ranged from 18 to 36 (M_{age} = 23, SD_{age} = 3.90, 47% male). Each team participated in two seven-hour sessions and was compensated $10/hour. This research complied with the American Psychological Association Code of Ethics and was approved by the Cognitive Engineering Research Institute's Institutional Review Board. Participants consented to participate.

Apparatus

This study utilized the Cognitive Engineering Research on Team Tasks-RPAS-STE (CERTT-RPAS-STE; Cooke & Shope, 2004), which has three team member roles: (1) *navigator*, who provided airspeed, altitude, and radius information to the pilot, and created and updated the flight plan; (2) *pilot*, who controlled aircraft specifications and negotiated with the photographer regarding altitude and airspeed; and (3) *photographer*, who asked the navigator for the effective radius of targets, operated the camera, and relayed the quality of captured photos to the team. Each team member had access to a limited amount of information which required them to communicate to complete their individual- and team-level tasks: (1) *navigator* (Figure 1), had access to the map, waypoint information, and to view the status of the RPA (i.e., airspeed and altitude); (2) *pilot* (Figure 2), had access to RPA controls and to view the current route; (3) *photographer* (Figure 3), had access to the camera and to view the status of the RPA. Teammates communicated via text chat.

Teams flew a simulated RPA through a series of Restricted Operating Zones (ROZs) to take photographs of target waypoints while avoiding hazard waypoints. All waypoints had airspeed, altitude, and effective radius restrictions that had to be met for the CERTT-RPAS-STE to recognize that the RPA was in the effective radius of a waypoint. ROZs were marked by "entry" and "exit" waypoints with two to three target waypoints in between. Teams first flew the RPA through an "entry" waypoint by meeting set airspeed, altitude, and radius restrictions, then flew to target waypoints within the same ROZ.

Once the RPA was in the radius of a target, the photographer could take a photograph. The photographer had to negotiate the altitude and airspeed with the pilot to match the settings required by their camera. Meanwhile, the pilot would adjust the altitude and airspeed to match the photographer's camera settings. After a target photograph was taken the team would fly the RPA to the next target waypoint, the "exit" waypoint, and then to the next ROZ. This cycle continued until all the targets were photographed or the mission time expired.

In this WoZ paradigm (e.g., Cooke et al., 2020a, 2020b; Kelley, 1983), the navigator and photographer were seated in the same room separated by a partition and instructed that the pilot was an AI agent, while the pilot (trained experimenter) was located in a separate room following a script to act as if they were an AI agent. We utilized a WoZ paradigm to introduce failures in a controlled fashion. The pilot was modeled after a synthetic pilot teammate developed by Ball et al. (2010) that utilized ACT-R cognitive modeling architecture (Anderson, 2007) to simulate human cognition via text chat. The WoZ pilot therefore had a repository of dialog for the task,

Figure 1. Navigator Role Screens. *Note.* Map and route options (left), RPA status screen (middle), and text-chat system (right).

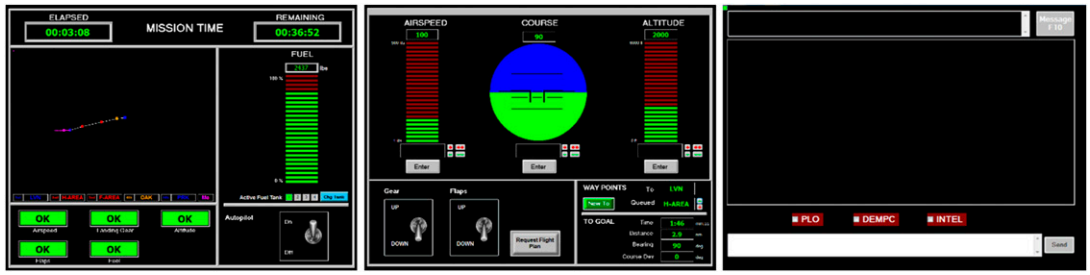

Figure 2. Pilot Role Screens. *Note.* RPA status screen (left), RPA controls (middle), and text-chat system (right).

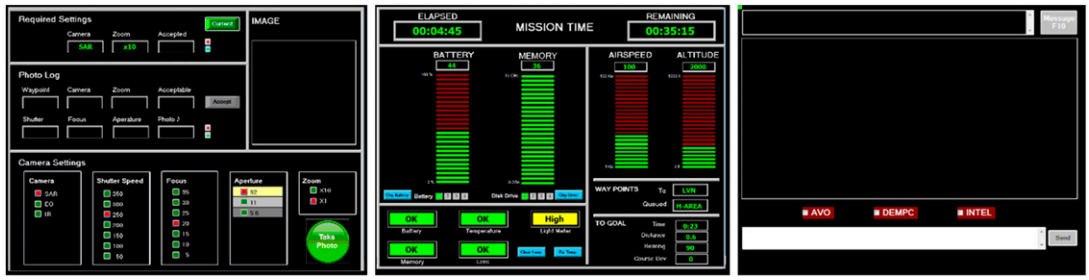

Figure 3. Photographer Role Screens. *Note.* Camera controls (left), RPA status screen (middle), and text-chat system (right).

and, during routine conditions understood requests for information and knew when to ask for information (McNeese et al., 2018).

Using a within-subjects manipulation, each team encountered two major types of technological perturbations: (1) automation failures involving role-specific display failures, which required the operator to obtain this information by communicating with another teammate and (2) autonomy failures involving "AI" pilot failures to either comprehend or anticipate the developing situation. Teams had to overcome failures by pushing or pulling the correct information in nonroutine ways to take a good photograph within a time limit (Table 1).

Procedure

The experiment consisted of ten 40-min missions across two seven-hour sessions with a one-to-two-week interval between. Each participant completed a

Table 1. Description of Automation and Autonomy Failures With Associated Team Solutions.

		Description of Failure	Team Solution
Autonomy	Type I	*Comprehension Failure I*: The AI agent repeatedly asks the same question to a human team member who sent it information.	To correct the AI agent's behavior, the human team member provides information regarding the next target waypoint.
	Type II	*Anticipation Failure*: The AI agent moves onto the next waypoint without giving the photographer enough time to take a photo (Cooke et al., 2020a, 2020b).	Either the navigator or photographer needs to display an awareness of the AI agent's limitations by becoming aware of the AI agent's abnormal behavior and direct the agent back to the missed target.
	Type III	*Comprehension Failure II*: The AI agent does not perform the required actions for the target waypoint due to its misunderstanding of information from human team members. For example, the AI agent may repeatedly adjust altitude for the wrong target.	Either the navigator can provide information for the next target waypoint, or the photographer can provide the altitude information for the current target waypoint again. If the teammate repeats the correct information, the AI agent would correct the altitude (Cooke et al., 2020a, 2020b).
Automation	Type I	A problem occurs on the photographer's screen, which prohibits the photographer to see the RPA status. This included current and next waypoint information, time, distance, bearing and course deviation.	The photographer can acquire the RPA status information from the pilot.
	Type II	A problem occurs on the pilot's screen, which prohibits the pilot to see the current altitude and airspeed of the RPA. In the context of a WoZ study, the participants are led to believe that the pilot cannot see the current altitude/airspeed.	The pilot can contact the photographer, since the photographer's screen shows this information.
	Type III	A more severe problem occurs on the pilot's screen, which prohibits the pilot to see the current altitude and airspeed, remaining time, distance, and bearing to the current target waypoint.	The pilot needs to communicate with both the photographer and the navigator to acquire the correct information about the upcoming waypoint.

Note. All failures were implemented according to a predetermined schedule. Increasing type numbers indicate increases in failure complexity. For example, the Type III Automation Failure involves more screens going out than either the Type I or Type II Automation failures. All failures could have been overcome by implementing either pushing or pulling communication strategies. Table 1 was adapted from Grimm et al. (2018).

30-min PowerPoint training. Each team then completed a 30-min hands-on training. Experimenters ensured participants could perform the task before beginning the experimental missions. There were 11–20 targets per mission. One automation and one autonomy failure were applied to two specified target waypoints during Missions 2–10.

Measures

Mission- and target-level measures of team performance, team process ratings, and team situation awareness were collected. To focus on communication as it relates to team performance, performance clustering considered: (1) *Number of failures overcome:* If a team overcame any failure within a mission, they received a 1; if they did not overcome a failure, they scored 0. We summed this variable across the nine failure missions. (2) *Mission level performance score:* a weighted composite of time spent in warning and alarm states, number of missed targets, and rate of acceptable photographs per minute. Each team began with a maximum score of 1,000, and points were deducted depending on the

final values of mission parameters (Cooke et al., 2007). (3) *Target Processing Efficiency (TPE)*: TPE is based on the time spent inside target radii, with higher scores corresponding to greater efficiency. The maximum score per target was 1000. Points were deducted if a good photo was not captured (Cooke et al., 2007).

Teams were clustered into three groups (high, medium, and low) based on their average mission-level performance score, TPE, and number of failures overcome using *K-means clustering* (Hastie et al., 2009), as reported in McNeese et al. (2019). We use the same clustered teams they report. As *K-means clustering* groups the data into a prespecified number of groups, the "Elbow Method" was chosen to determine the optimal number of groups, *k*. This method entails plotting the within groups sum of squares for multiple values of *k* and finding the point at which the marginal drops, forming an "elbow" (Sarstedt & Mooi, 2014). Figure 4 below shows an "elbow" at $k = 2$ and $k = 4$. However, the values of the within groups sum of squares taper off after $k = 4$ suggesting the "elbow" to be $k = 4$ and the optimal number of clusters, *k*, to be three. Teams were then clustered into high-, medium-, and low-performing groups, with eight, seven, and seven teams, respectively, in each cluster.

Table 2 outlines the five communication strategies used to code pushing and pulling communications for human and "AI" teammates, derived from the broader spectrum of eight verbal behaviors associated with team coordination (Demir et al., 2016). For the purposes of this study, we selected only the five behaviors classified as either pushing or pulling of information.

Communications were independently coded by two experimenters. Inter-rater reliability was evaluated for agreement with Cohen's κ. Pushing codes, $\kappa = 0.844$, 95% CI [0.838, 0.850], and pulling codes, $\kappa = 0.861$, 95% CI [0.855, 0.867], indicating high agreement. To control for the total amount of time a HAT spent in a given failure or nominal (no failure) scenario, we divided pushing and pulling counts by the time (in minutes) the HAT spent addressing the scenario to generate pushing and pulling *rates*.

Overcoming failures required teams to coordinate using nonroutine patterns. We focus our analyses on differences in communication strategies between the various failures and levels of team performance, as pushing is hypothesized to be more advantageous than pulling information. Although the experimental manipulation and the focus of our analysis potentially share some variance due to the centrality of communication in both, all failures could have theoretically been overcome with either pushing *or* pulling strategies.

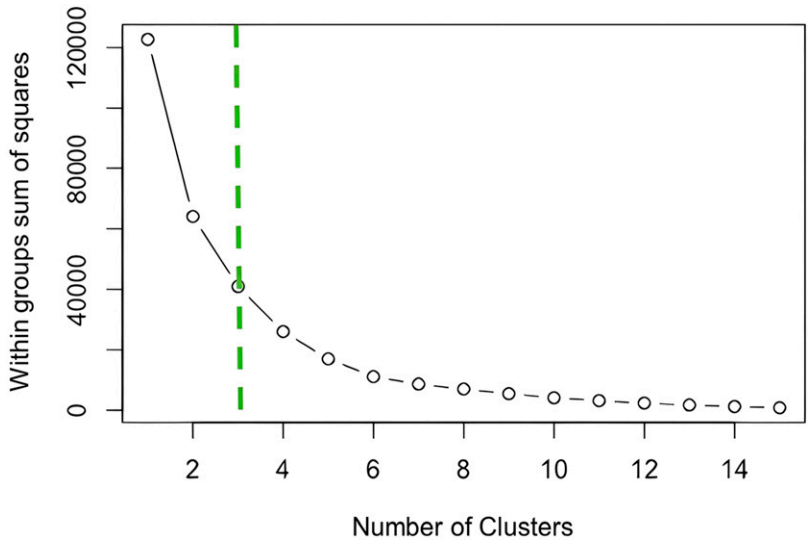

Figure 4. Within Groups Sum of Squares as a Function of the Number of Clusters. *Note*. Figure adapted from McNeese et al. (2019).

Table 2. Pushing and Pulling Communications.

Behavior	Type	Description
General status update	Push	Informing other team members about current status
Suggestion	Push	Making suggestions to the other team members
Planning ahead	Push	Anticipating next steps and creating rules for future encounters
Inquiry About the status of others	Pull	Inquiries about the status of other team member(s) seeking information on conditions, directions, ranges, or representations.
Repeated request	Pull	Requesting the same information or action from other team member(s)

Note. Table 2 was adapted from Demir et al. (2017).

Results

Team Communication as a Function of Performance Cluster and Failure Type

We ran separate 3 (Cluster: High, Medium, Low) × 3 (Failure: None, Automation, Autonomy) × 8 (Mission: 2–9) mixed Analysis of Variances (ANOVAs) on pushing and pulling rates. Because the within-subject data matrix resulted in missing communication data for some missions, we utilized a mixed ANOVA technique that allowed us to include all available data in the analyses (Enders, n.d). This technique analyzes the data as a multilevel model and uses the Satterthwaite approximation to calculate denominator degrees of freedom (UCLA, n.d). Due to the use of this technique, we measured effect sizes using Cohen's *d*, where $d = 0.20$, 0.50, and 0.80, indicate small, medium, and large effects, respectively (Cohen, 1977).

Pushing Communications. For pushing communications, there was a significant Mission main effect, $F(8, 545.15) = 14.17$, $p < .001$, Failure main effect, $F(2, 545.16) = 85.33$, $p < .001$, Mission × Failure interaction, $F(16, 545.10) = 4.34$, $p < .001$, and Cluster × Failure interaction, $F(4, 545.15) = 3.89$, $p = .004$.

To explore how failures differed by performance level (H1), we assessed the simple effects of the Cluster × Failure interaction by first examining the effect of Failure across different levels of Cluster (Least Significant Difference; LSD; Figure 5). Low-performing teams had a lower pushing rate during times of no failure ($M = 0.94$, $SD = 0.58$) compared to automation ($M = 1.42$, $SD = 0.79$, $p < .001$, $d = 0.70$) and autonomy failures ($M = 1.50$, $SD = 1.08$, $p < .001$, $d = 0.66$), and pushing rates did not differ between the two failure types. For medium- and high-performing teams, pushing differed across all levels of Failure, wherein autonomy failures had the greatest pushing rate ($M_{Med} = 1.57$, $SD_{Med} = 0.85$, $M_{High} = 2.04$, $SD_{High} = 0.86$), compared to automation failures ($M_{Med} = 1.43$, $SD_{Med} = 0.60$, $M_{High} = 1.59$, $SD_{High} = 0.55$), followed by times of no failure ($M_{Med} = 1.12$, $SD_{Med} = 0.62$, $M_{High} = 1.18$, $SD_{High} = 0.38$), all $p < .05$. In accordance with H1, results suggest high-performing teams adapted their pushing rates to match the increasing complexity of failure types, with high-performing teams increasing their pushing more than medium-performing teams from automation to autonomy failures (Figure 5).

Our investigation into H2 explored the Mission main effect, and a separate trend analysis conducted in R (R Core Team, 2021), to assess whether teams learned to increase pushing rates over time. Pairwise comparisons (LSD) provided support for this hypothesis, in that from Mission 2 through 4, $p < .05$ for all comparisons, from Mission 7 to 8, $p = .005$, and from Mission 9 to 10, $p < .001$, teams increased pushing behaviors (Figure 6). The trend analysis also supported H2, in that a significant linear trend was found over the nine missions, $F(8, 558) = 14.24$, $p < .001$, $\eta_p^2 = 0.17$. A significant cubic trend was also found, $F(8, 558) = 7.91$, $p = .005$, $h_p^2 = 0.10$, but this reflects the dip in pushing between Missions 4 and 5, which was likely the result of Mission 5 being the first mission of Session 2, for which teams had to briefly reacquire pushing strategies. When assessing trends in the two sessions separately, we found significant linear trends from both Missions 2–4, $F(2, 186) = 6.72$, $p < .001$, $h_p^2 = 0.07$, and from Missions 5–10, $F(5, 372) = 16.59$, $p < .001$,

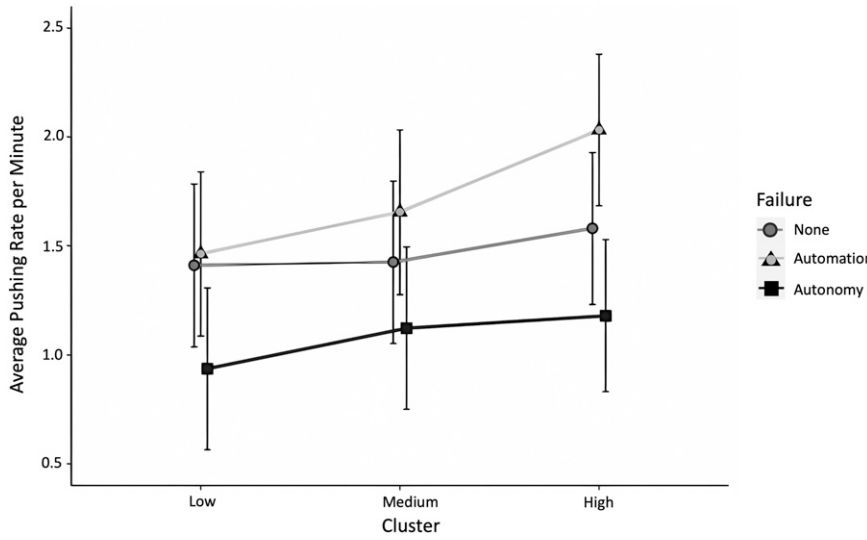

Figure 5. Performance Cluster by Failure Type Interaction on Pushing Communication Rate. *Note.* Error bars represent 95% confidence intervals (CIs).

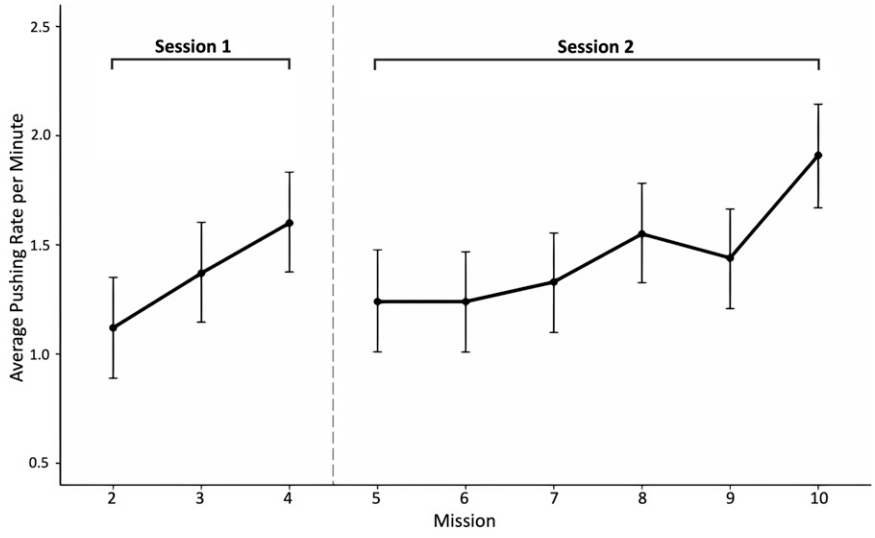

Figure 6. Average Rate of Pushing Behaviors over Missions 2-10. *Note.* Error bars represent 95% CIs.

$\eta_p^2 = 0.18$. We explore this finding further as it relates to H3.

In testing H3, pairwise comparisons (LSD) indicated that all three Failure conditions led to different pushing rates, $p < .001$. During autonomy failures ($M = 1.73$, $SD = 0.96$), teams generated higher pushing rates than during automation failures ($M = 1.48$, $SD = 0.65$, $d = 0.31$) and no failure ($M = 1.09$, $SD = 0.54$, $d = 0.84$). Automation failures generated higher pushing rates than times with no failures ($d = 0.67$). These findings indicate that as failures became more challenging to overcome, teams pushed more information to overcome them, providing support for H3.

To further assess H3, a simple effects analysis of the Mission × Failure interaction was conducted

to understand how pushing communications changed in response to different failure types (e.g., Type II vs. Type III autonomy; see Table 1). We assessed the effect of Failure at different levels of Mission as each mission had only one type of autonomy and automation failure. Pairwise comparisons (LSD) indicated that meaningful differences predominately occurred during autonomy failures (Figure 7). During times of no failure and automation failures, teams displayed relatively stable pushing rates. Pushing appears to be particularly important in addressing autonomy failures. As shown in Figure 3, pushing serially increased from Type I to Type II to Type III autonomy failures. In particular, the Type III comprehension failure likely forced human teammates to continually push information to the agent to enforce comprehension, while also sharing information with each other. This suggests that both failure complexity (e.g., automation vs. autonomy) and failure severity (e.g., comprehension autonomy failure with direct implications for task performance; autonomy failure Type III) drive pushing behaviors, in accordance with H3.

Pulling Communications. For pulling communications, there was a significant Failure main effect, $F(2, 545.51) = 109.98$, $p < .001$, Mission main effect, $F(8, 545.50) = 2.81$, $p = .005$, Cluster × Failure interaction, $F(4, 545.50) = 3.48$, $p = .008$, and Mission × Failure interaction, $F(16, 545.35) = 3.18$, $p < .001$.

To explore how failures differed by performance level (H1), we assessed the simple effects of the Cluster × Failure interaction by examining the effect of Failure at different levels of Cluster (LSD; Figure 8). Low-performing teams had a lower pulling rate during times of no failure ($M = 0.85$, $SD = 0.28$) compared to automation ($M = 1.35$, $SD = 0.64$), $p < .001$, $d = 1.00$, and autonomy failures ($M = 1.39$, $SD = 0.55$), $p < .001$, $d = 1.25$, although pulling rates did not differ between the two failure types. For medium- and high-performing teams, however, pulling differed across all levels of failure, where autonomy failures had the greatest rate of pulling ($M_{Med} = 1.26$, $SD_{Med} = 0.52$, $M_{High} = 1.63$, $SD_{High} = 0.94$), compared to automation failures ($M_{Med} = 1.04$, $SD_{Med} = 0.38$, $M_{High} = 1.22$, $SD_{High} = 0.44$), followed by times of no failure ($M_{Med} = 0.68$, $SD_{Med} = 0.28$, $M_{High} = 0.81$, $SD_{High} = 0.28$), $p \leq .01$ for all comparisons within clusters. We expected less pulling to make way for more pushing in higher performing teams as failures grew in complexity (H1); on the contrary, the pulling results essentially mirrored the pushing results. Thus, increases in failure complexity led to increased pushing *and* pulling communications.

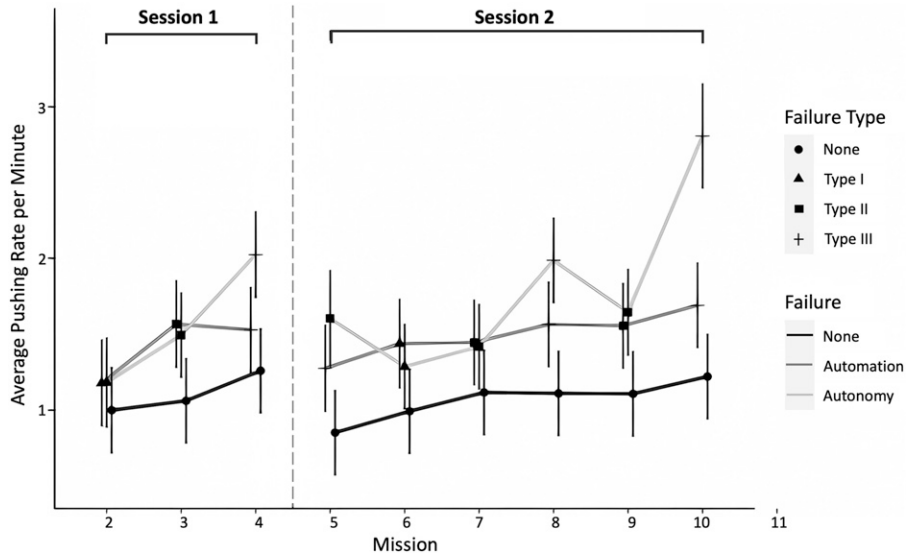

Figure 7. Average Rate of Pushing Behaviors over Mission, Split by Failure Type. *Notes.* Lines represent broad failure type (none, automation, autonomy), and markers represent different failure subtypes (I, II, or III). Error bars represent 95% CIs.

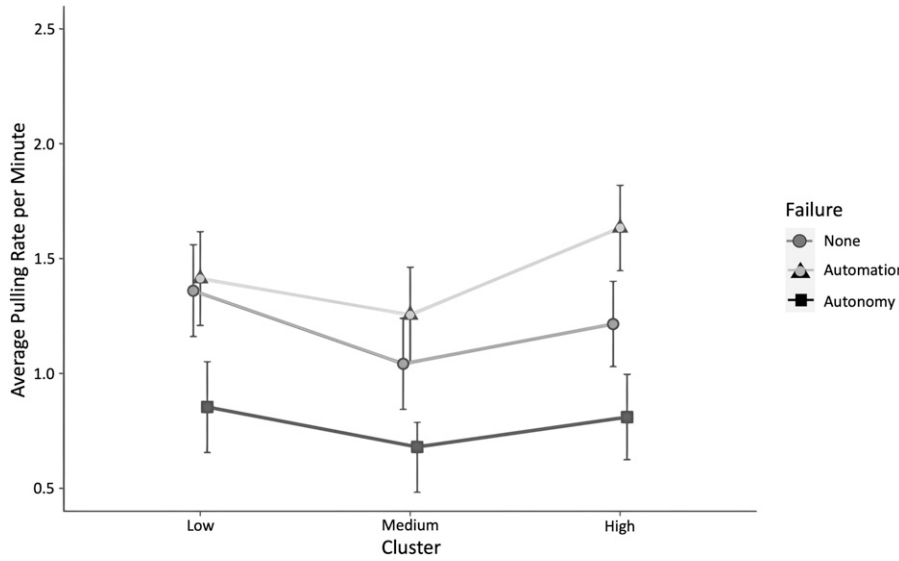

Figure 8. Performance Cluster by Failure Type Interaction on Pulling Rates. *Note.* Error bars represent 95% CIs.

Examining the simple effects of the Cluster × Failure interaction for the effect of Cluster at different levels of Failure (Figure 8) we found two significant differences. During automation failures, medium-performing teams (M = 1.04, SD = 0.38) had a lower pulling rate than low-performing teams (M = 1.35, SD = 0.64), $p < .05$, d = 0.58. During autonomy failures, high-performing teams (M = 1.63, SD = 0.94) had a higher pulling rate than medium-performing teams (M = 1.22, SD = 0.52), p = .009, d = 0.53. These results suggest that high pulling rates do not precisely correspond to high performance (Figure 4). Given that high-performing teams always pushed the most and low-performing teams always pushed the least, these failure-based differences in pulling behaviors reinforce pulling communications' less direct impact on performance compared to pushing (Demir et al., 2017).

Our investigation into H2 explored the Mission main effect, and a separate trend analysis conducted in R (R Core Team, 2021), to assess whether teams adapted their communication strategy over time. Pairwise comparisons (LSD) revealed only a significant decrease in pulling during Mission 8 (M = 0.94, SD = 0.46) compared to all other missions ($p <$ .05; Figure 9). The dip in pulling at Mission 8 may be attributable to the challenge of overcoming both a Type III automation and autonomy failure. This was corroborated by Mission 4, where two Type III failures occurred resulting in a dip in pulling. The trend analysis revealed a significant linear trend, F(8, 558) = 4.15, p = .04, η_p^2 = 0.06, suggesting a slight decreasing trend over the nine missions, further supporting H2. When assessed separately for the two sessions, no significant trends were found.

In testing H3, pairwise comparisons (LSD) of the Failure main effect revealed the three Failure conditions had different pulling rates, $p < .001$ for all comparisons. Mirroring pushing, autonomy failures (M = 1.43, SD = 0.74) generated the highest pulling rate, compared to automation failures (M = 1.21, SD = 0.51), followed by times of no failure (M = 0.78, SD = 0.29). Contrary to expectations, pulling may also play a role in overcoming complex failures. This is likely due to the nature of autonomy comprehension failures, which may require human team members to repeatedly ask the pilot for information.

To further understand how pulling rates changed in response to different failures (H3), we conducted a simple effects analysis of the Mission × Failure interaction. Because each mission had only one type (e.g., Type I, Type II, and Type II; see Table 1) of autonomy and automation failure, we assessed the effect of Failure at different levels of Mission (LSD). As shown in Figure 10, for automation failures, pulling was relatively stable except for a peak at

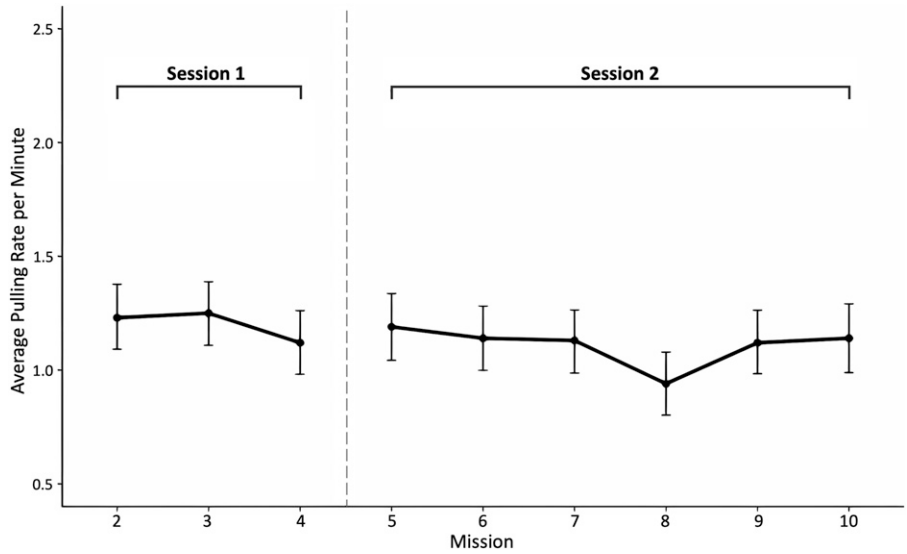

Figure 9. Average Rate of Pulling Behaviors over Missions 2-10. *Note.* Error bars represent 95% CIs.

Mission 3 (M = 1.57, SD = 0.60, $p < .05$), which significantly differed from all missions except Mission 6, ($p = .07$). For autonomy failures, a slight peak in pulling was observed during Mission 5 (M = 1.73, SD = 1.15), which differed from Missions 3, 4, 6, 7, and 8 ($p < .05$), and a valley during Mission 8 (M = 1.02, SD = 0.65), which differed from all missions ($p < .05$) except Mission 4 ($p = .08$). The Mission 4 and 8 valleys in pulling communications correspond to peaks in pushing that came in response to Type III automation and autonomy failures. Furthermore, though not statistically significant, we observed a downward trend in pulling over time during times of no failure, which partially supports H2. This pulling decline was met with a slight increase in pushing during routine times. Though pulling has a less systematic relationship with overcoming failures, teams do seem to adopt more efficient communication strategies as they become familiar with the task.

Control Analysis. We ran a control analysis to test whether our findings could be due to the overall frequency of communication rather than pushing or pulling specifically. To do this, we ran a 3 (Cluster) × 3 (Failure) × 9 (Mission) mixed ANOVA on the number of chat messages (i.e., overall communication frequency) for Missions 2–10. We were unable to replicate the findings of the prior analyses and did not find a Cluster effect ($p = .802$), indicating that the effects reported above are specific to pushing and pulling, not communication in general.

Discussion

Our hypotheses were partially supported, with the data revealing a more complex picture than anticipated. H1 predicted higher pushing rates for higher performing teams across the three failure types, wherein higher pushing rates would occur with more complex failures. Furthermore, we predicted an associated reduction in pulling during complex failures, as pushing is more efficient. Thus, pushing should be relied on during periods of high workload (MacMillan et al., 2004) and has been found to be more strongly associated with performance (Demir et al., 2017). The former was supported, wherein high-performing teams better calibrated their communication rates to the degree of failure complexity. Medium-performing teams managed to adapt their communication as the failure situation changed, but not with the fidelity of high-performing teams, and low-performing teams failed to update their communications to match the autonomy failures. What appeared to differentiate medium- and high-performing teams was how they handled pulling behaviors, which was unexpected. Though past literature suggests high-performing teams might trade pulling for

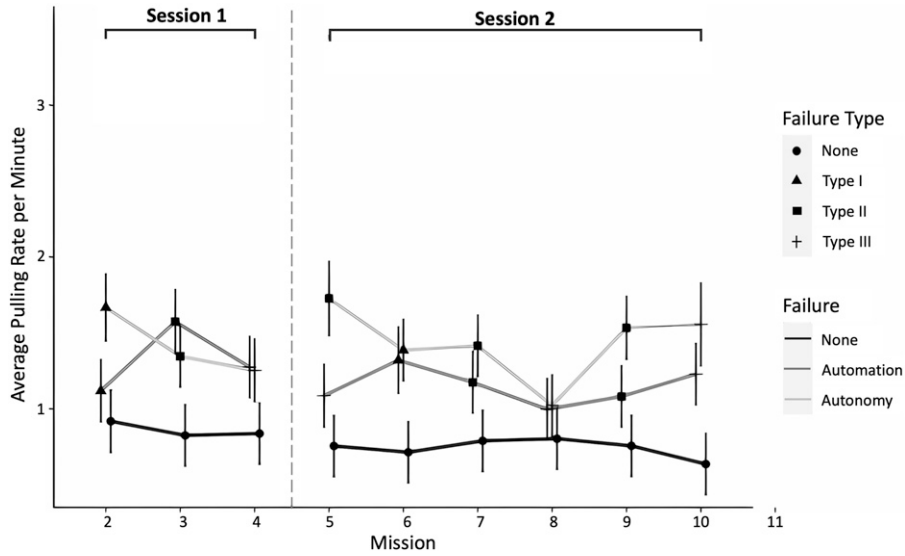

Figure 10. Average Rate of Pulling Behaviors over Mission. *Note.* Lines represent broad failure type (none, automation, autonomy), and markers represent different failure subtypes (I, II, or III). Error bars represent 95% CIs.

pushing communications during times of high stress (MacMillan et al., 2004), this only seemed to be the case for medium-performing teams. Perhaps due to increased bandwidth, high-performing teams matched the complexity of the failures through both increased pushing and pulling communications.

This finding is reminiscent of Ashby, (1956) Law of Requisite Variety, which posits that for acceptable performance of a controlled system, the variety demanded by the system must be matched by the controller's increased behavioral variety required to control the system. This is exhibited in the current study by the varying degree of communication rates across the three performance levels, for which high-performing teams exhibit increased pushing and pulling rates for all three failure states, whereas the other performance clusters engaged in a narrower range of communication rates. Additionally, the relationship between increased variety in pushing/pulling and performance cluster was contingent on failure type, with low-performing teams exhibiting increased pulling for automation failures and high-performing teams exhibiting increased pulling for autonomy failures. This suggests that both the *amount* of communication variety and the *type* of communication variety contribute to team effectiveness under degraded conditions.

H2 predicted that teams would increase pushing over time. We also hypothesized a reciprocal reduction in pulling, as teams learned to reduce communication overhead (MacMillan et al., 2004). Though teams tended to increase pushing over time, this was only matched with a downward trend in pulling during times of no failure. Thus, we infer the boundaries of the communication overhead framework may not extend to tasks involving failure perturbations. In the present study, the nature of the failure may have been a partial driver of communication demands, wherein effective teams may not have been able to reduce their reliance on pulling to reduce communication overhead. Though, theoretically, pushing was an equally viable strategy to addressing failures, pulling communications may have been a default strategy to overcome failures.

Lastly, H3 predicted that more severe failures would require teams to rely more on pushing than pulling. This tradeoff was observed in only the most complex failure environments involving both Type III automation and autonomy failures. It is possible that the Type III failures may have required teams to adjust the efficiency of their communications such that more attention could be paid to the task (MacMillan et al., 2004). In failures of lesser severity, there was little evidence of a systematic push-pull tradeoff. This points to an explanation of the

communication overhead framework that goes beyond prior theories, which attribute the emergence of efficient communication strategies to shared mental models (Entin & Serfaty, 1999; MacMillan et al., 2004). Here, we see that differences in pushing and pulling communications may also be context/task-dependent and likely driven by the changing constraints of the environment. In times of nonroutine team coordination, even with potential training for team coordination in place, shared mental models may not always capture the nuances of the present situation. Thus, in accordance with ITC, team coordination strategies reflect the momentary demands of the task environment, wherein task constraints are reflected in team communication strategies.

Limitations and Future Directions

The present study may have been limited in several ways. First, the task environment was itself a high workload situation, wherein, even without failures, some teams could not process all targets within the allotted time. Additionally, the experiment spanned two seven-hour sessions, which could have contributed to fatigue. These sources of fatigue may impact the generalizability of the results. Though HATs tend to work in complex environments, future research should explore how the task's workload directly impacted communication overhead, separate from the failure conditions.

Another limitation is the type of environmental variety mentioned in the previous section. The present study used a limited set of failures to enact experimental control. The failures the present study adopted may provide only a small number of potential perturbations that would occur in actual RPAS scenarios. Future research should explore a larger variety of failures to better generalize to field settings.

Finally, we expected teams to overcome autonomy and automation failures using specific pushing and pulling communications that were rooted primarily in failures involving the "AI" pilot. Assuming the different failures may constrain different pushing and pulling strategies that teams employ, the impact of communication failures on the HAT system could be further explored if they occurred in different roles on the team other than the pilot.

Conclusion

The current study contributes to a growing body of knowledge on HAT communication and team performance (e.g., Demir et al., 2017). Failures implemented in the present study were uniquely positioned to be studied using team communication, as their solution required novel communication patterns. Like prior work using all-human teams (e.g., MacMillan et al., 2004), the anticipation of team member needs via pushing communications was evident in high-performing HATs. However, we also saw a preponderance of pulling communications that contributed to HAT success in failure-laden contexts—a finding not previously observed in all-human teams. In designing autonomous agents, close attention should be paid to their ability to not only communicate effectively but also their agility in adapting their interaction patterns to meet the changing complexity of dynamic and often degraded environments in which they are deployed (Song et al., 2022).

Key Points

- Higher performing teams had higher rates of pushing behaviors across failure types, wherein higher rates of pushing occurred with higher complexity failures.
- Higher performing teams better calibrated their communication behavioral complexity to the degree of failure complexity compared to medium performing teams, and low performing teams failed to adjust their communications.
- It is not just the amount of communication variety but the type of communication variety that contributes to team effectiveness in degraded conditions.
- Teams tended to adjust their communication strategies in response to failures over time by increasing their pushing communications. However, this was not consistently matched with a decrease in pulling communications.
- Only the most severe autonomy comprehension failures required teams to confront the efficiency of their communications behaviors through which teams relied more on pushing communications than pulling communications.

Acknowledgments

This research is supported by ONR Award N000141712382 (Program Managers: Marc Steinberg, Micah Clark). We acknowledge the assistance of Paul Jorgeson and Steve Shope in modifying the CERTT-RPAS-STE, Terri Dunbar for assistance with data analysis, and Cody Radigan, Craig Johnson, Sophie He, Matthew Lin, Tanvi Tandolkar, Garrett Zabala, and Alexandra Wolff for data collection efforts.

ORCID iD

Shiwen Zhou https://orcid.org/0000-0002-2851-1940

References

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford University Press.

Ashby, W. R. (1956). *An introduction to cybernetics*. John Wiley and Sons.

Ball, J., Myers, C., Heiberg, A., Cooke, N., Matessa, M., Freiman, M., & Rodgers, S. (2010). The synthetic teammate project. *Computational & Mathematical Organization Theory*, *16*(3), 271–299. https://doi.org/10.1007/s10588-010-9065-3

Cannon-Bowers, J. A. & Salas, E. (1990). Cognitive psychology and team training: Shared mental models in complex systems. In Paper presented at the 5th annual Conference of the Society for Industrial and Organizational Psychology. Miami, FL.

Chow, R., Christoffersen, K., & Woods, D. D. (2000). A model of communication in support of distributed anomaly response and replanning. In Proceedings of the Human Factors And Ergonomics Society - Annual Meeting (Vol. *44*, No. 1, pp. 34–37). Sage CA, Sage Publications.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. Academic Press.

Cooke, N., Demir, M., McNeese, N., Gorman, J., & Myers, C. (2020a). *Human–autonomy teaming in remotely piloted aircraft systems operations under degraded conditions*. Cognitive Engineering Research Institute.

Cooke, N. J., Demir, M., & Huang, L. (2020b). *A framework for human-autonomy teaming research*. In 22nd International Conference on Human-Computer Interaction, Copenhagen, Denmark.

Cooke, N. J., Gorman, J., Pedersen, H., Winner, J., Duran, J., Taylor, A., Amazeen, P. A., Andrews, D. H., & Rowe, L. (2007). *Acquisition and retention of team coordination in command-and-control*. Cognitive Engineering Research Inst Mesa Az.

Cooke, N. J., Gorman, J. C., Myers, C. W., & Duran, J. L. (2013). Interactive team cognition. *Cognitive Science*, *37*(2), 255–285. https://doi.org/10.1111/cogs.12009

Cooke, N. J. & Shope, S. M. (2004). Designing a synthetic task environment. In L. R. E. Schiflett, E. Salas, & M. D. Coovert (Eds.), *Scaled worlds: Development, validation, and application* (pp. 263–278). Ashgate Publishing.

DeChurch, L. A. & Mesmer-Magnus, J. R. (2010). Measuring shared team mental models: A meta-analysis. *Group Dynamics: Theory, research, and practice*, *14*(1), 1–14. https://doi.org/10.1037/a0017455

Demir, M. & Cooke, N. J. (2014). Human teaming changes driven by expectations of a synthetic teammate. In Proceedings of the Human Factors and Ergonomics Society - Annual Meeting (Vol. *58*, No. 1, pp. 16–20). Sage CA, Sage Publications.

Demir, M., McNeese, N. J., & Cooke, N. J. (2017). Team situation awareness within the context of human-autonomy teaming. *Cognitive Systems Research*, *46*, 3–12. https://doi.org/10.1016/j.cogsys.2016.11.003

Demir, M., McNeese, N. J., & Cooke, N. J. (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) (pp. 28–34). IEEE.

Enders, C. K. (n.d.). *Three-factor mixed ANOVA (one between- and two within-subjects factors)*. Applied Missing Data .com: Companion website for Applied Missing Data Analysis. https://www.appliedmissingdata.com/three-factor-mixed-anova.html

Entin, E. E. & Serfaty, D. (1999). Adaptive team coordination. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *41*(2), 312–325. https://doi.org/10.1518/001872099779591196

Gorman, J. C., Demir, M., Cooke, N. J., & Grimm, D. A. (2019). Evaluating socio-technical dynamics in a simulated remotely-piloted aircraft system: A layered dynamics approach. *Ergonomics*, *62*(5), 629–643. https://doi.org/10.1080/00140139.2018.1557750

Gorman, J. C., Cooke, N. J., & Winner, J. L. (2006). Measuring team situation awareness in decentralized command and control environ- ments. *Ergonomics*, *49*(12-13), 1312–1325. https://doi.org/10.1080/00140130600612788

Grimm, D., Demir, M., Gorman, J. C., & Cooke, N. J. (2018). Systems level evaluation of resilience in

human-autonomy teaming under degraded conditions. In *2018 resilience week (RWS)* (pp. 124–130). IEEE.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (vol. *2*, pp. 1–758). Springer.

Kaber, D. B. (2018). A conceptual framework of autonomous and automated agents. *Theoretical Issues in Ergonomics Science*, *19*(4), 406–430. https://doi.org/10.1080/1463922x.2017.1363314

Kelley, J. F. (1983). An empirical methodology for writing user-friendly natural language computer applications. Proceedings of ACM SIG-CHI '83*Human Factors in Computing S*ystems (pp. 193–196). ACM.

MacMillan, J., Entin, E. E., & Serfaty, D. (2004). Communication overhead: The hidden cost of team cognition. In E. Salas, & S. M. Fiore (Eds.), *Team cognition: Understanding the factors that drive process and performance* (pp. 61–82). American Psychological Association. https://doi.org/10.1037/10690-004

McNeese, N., Demir, M., Chiou, E., Cooke, N., & Yanikian, G. (2019). *Understanding the role of trust in human-autonomy teaming*. *Proceedings of the 52nd Hawaii International Conference on System Sciences*, *6*, 254–263. https://doi.org/10.24251/hicss.2019.032

McNeese, N. J., Demir, M., Cooke, N. J., & Myers, C. (2018). Teaming with a synthetic teammate: Insights into human-autonomy teaming. *Human Factors*, *60*(2), 262–273. https://doi.org/10.1177/0018720817743223

National Academies of Sciences. (2021). Engineering and medicine. In *Human-AI teaming: State of the art and research needs*. The National Academies Press. https://doi.org/10.17226/26355

O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2022). Human–autonomy teaming: A review and analysis of the empirical literature. *Human Factors*, *64*(5), 904–938. https://doi.org/10.1177/0018720820960865

Orasanu, J. M. (1990). *Shared mental models and crew decision making (CSL Report No. 46)*. Princeton University, Cognitive Science Laboratory.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Riek, L. D. (2012). Wizard of Oz studies in Hri: A systematic review and new reporting guidelines. *Journal of Human-Robot Interaction*, *1*(1), 119–136. https://doi.org/10.5898/jhri.1.1.riek

Sarstedt, M. & Mooi, E. (2014). Cluster analysis. In *A concise guide to market research* (pp. 273–324). Springer. https://doi.org/10.1007/978-3-642-53965-7_9

Shively, R. J., Lachter, J., Koteskey, R., & Brandt, S. L. (2018). Crew resource management for automated teammates (CRM-A). In *International conference on engineering Psychology and cognitive ergonomics* (pp. 215–229). Springer.

Song, B., Gyory, J. T., Zhang, G., Soria Zurita, N. F., Stump, G., Martin, J., Miller, S., Balon, C., Yukish, M., McComb, C., & Cagan, J. (2022). Decoding the agility of artificial intelligence-assisted human design teams. *Design Studies*, *79*(1), 101094. https://doi.org/10.1016/j.destud.2022.101094

Tenhundfeld, N. & ChatGPT. (2023). *Two birds with one stone: Writing a paper entitled "Chat GPT as a tool for studying human-AI interaction in the wild" with ChatGPT*. https://doi.org/10.13140/RG.2.2.25319.73123

UCLA: Statistical Consulting Group. (n.d.). *SPSS Mixed Command*. https://stats.idre.ucla.edu/spss/seminars/spss-mixed-command/

Wardat, Y., Tashtoush, M. A., AlAli, R., & Jarrah, A. M. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *Eurasia Journal of Mathematics, Science and Technology Education*, *19*(7), em2286. https://doi.org/10.29333/ejmste/13272

Author Biographies

Julie L. Harrison is an engineering psychology graduate student at the Georgia Institute of Technology. She received her MS in psychology from Georgia Institute of Technology in 2021.

Shiwen Zhou is a PhD student in the human systems engineering program at Arizona State University. She received her MS in psychology from Georgia Institute of Technology in 2022.

Matthew J. Scalia is a PhD student in human systems engineering at Arizona State University. He received his MS in psychology from Georgia Institute of Technology in 2022.

David A.P. Grimm is a PhD student in engineering psychology at Georgia Institute of Technology. He received his MS in psychology from the Georgia Institute of Technology in 2020.

Mustafa Demir is currently an assistant research professor and faulty associate working at Global Security Initiative and Ira. A. Fulton Schools of Engineering, respectively, at Arizona State University. He received his PhD in simulation, modeling, and applied cognitive science, focusing on team coordination dynamics in human–machine teaming from Arizona State University in Spring 2017.

Nathan McNeese is the College of Engineering, Computing and Applied Sciences Dean's Professor, an assistant professor of Human-Centered Computing, and the Director of the Team Research Analytics in Computational Environments (TRACE) Research Group within the School of Computing, Clemson University. His research interests and expertise include human-AI teaming and human-centered AI.

Nancy J. Cooke is a professor of Human Systems Engineering at Arizona State University and directs ASU's Center for Human, Artificial Intelligence, and Robot Teaming. Dr. Cooke studies individual and team cognition and its application to human, AI, and robot teaming and conducts empirical assessments of teams and teamwork.

Jamie C. Gorman is a professor in the Human Systems Engineering program at Arizona State University. He received his PhD in psychology from New Mexico State University in 2006.