Check for updates

# Dynamical Measurement of Team Resilience

**David A.P. Grimm**, Georgia Institute of Technology, GA, USA, **Jamie C. Gorman, Nancy J. Cooke** and **Mustafa Demir**, Arizona State University, AZ, USA, and **Nathan J. McNeese**, Clemson University, SC, USA

Resilient teams overcome sudden, dynamic changes by enacting rapid, adaptive responses that maintain system effectiveness. We analyzed two experiments on human-autonomy teams (HATs) operating a simulated remotely piloted aircraft system (RPAS) and correlated dynamical measures of resilience with measures of team performance. Across both experiments, HATs experienced automation and autonomy failures, using a Wizard of Oz paradigm. Team performance was measured in multiple ways, using a mission-level performance score, a target processing efficiency score, a failure overcome score, and a ground truth resilience score. Novel dynamical systems metrics of resilience measured the timing of system reorganization in response to failures across RPAS layers, including vehicle, controls, communications layers, and the system overall. Time to achieve extreme values of reorganization and novelty of reorganization were consistently correlated with target processing efficiency and ground truth resilience across both studies. Correlations with mission-level performance and the overcome score were apparent but less consistent. Across both studies, teams displayed greater system reorganization during failures compared to routine task conditions. The second experiment revealed differential effects of team training focused on coordination coaching and trust calibration. These results inform the measurement and training of resilience in HATs using objective, real-time resilience analysis.

**Keywords:** adaptation, coordination, relaxation time, resilience engineering, team cognition

Address correspondence to David A.P. Grimm, School of Psychology, Georgia Institute of Technology, 654 Cherry St NW, Atlanta, GA 30332-0365, USA.
Email: david.grimm@gatech.edu

*Journal of Cognitive Engineering and Decision Making*
Vol. 17, No. 4, December 2023, pp. 351–382
DOI:10.1177/15553434231199729
Article reuse guidelines: sagepub.com/journals-permissions
Copyright © 2023, Human Factors and Ergonomics Society.

## Introduction

Effective teams efficiently coordinate heterogeneous and shared resources to accomplish shared and valued goals (Salas et al., 2008). In this context, a resilient team responds to undesirable conditions and challenges, such as system failures, by rapidly reorganizing its resources to maintain high levels of team performance (Alliger et al., 2015; Morgan et al., 2017). It is theorized that resilient teams accomplish this by rapidly recognizing, designing, and implementing changes to ward off novel impediments to team effectiveness outside their current areas of capable performance (Hoffman & Hancock, 2017). Lack of team resilience is exemplified in the 1996 Mount Everest climbing disaster, in which eight climbers died while climbing Mount Everest. A lack of team learning—including ill-defined purpose, vague leadership, and poor sensemaking—were key contributors to this disaster, but this disaster was at least partially attributable to a breakdown of team coordination (Kayes, 2004). Lack of resilience and coordination was also observed in the delayed response to Hurricane Katrina (Leonard & Howitt, 2006), in which a more rapid system reorganization may have sped up relief and the subsequent recovery of those impacted by the storm surge (Colten et al., 2008). In contrast, rapid reorganization of system resources has been associated with timely and effective responses (e.g., military-civilian evacuation efforts following 9/11; Boin & Bynander, 2015). In line with Hoffman and Hancock (2017), we propose that more resilient teams exhibit faster detection of impending catastrophes, implementation of required changes (i.e., reorganization behavior), and are thus able to recover from novel threats more rapidly.

Emergency response in aviation and other power system failures are long-standing concerns in resilience research (e.g., Woods et al., 1988). We

build on this research by applying an objective, data-driven approach to measuring team resilience with the potential for real-time analysis that can provide training, feedback, and identify critical sources of system reorganization underlying resilience. We focus on human-autonomy teaming (HAT), which is defined as teams in which humans work with technological agents that are intelligent and autonomous enough to be considered a teammate (McNeese et al., 2018). The study of HATs increasingly applies to safety-critical domains, including urban search and rescue (Krujiff et al., 2014), uninhabited aerial systems (McNeese et al., 2018), cyberspace operations (Tambe et al., 1999), and self-driving autonomous vehicles (Campbell et al., 2010). By enabling flexible, adaptive, and rapid team responses (Hoffman & Hancock, 2017; Hollnagel et al., 2007), a resilient HAT would be better equipped to rapidly overcome potential pitfalls associated with unpredictable challenges, such as automation and autonomy failures, cyberattacks, communication link failures, and system power outages. Many of these common pitfalls in HATs are associated with brittleness, lack of transparency, miscalibrated trust, and a lack of shared awareness (Shively et al., 2017). For example, although a human working with an autonomous agent may lack shared situation awareness with the agent, a resilient HAT would be more likely to overcome an error resulting from this lack of shared awareness by quickly reorganizing how it coordinates across system layers. These potential pitfalls associated with HATs make these types of teams suitable for studying resilience.

In this paper, we describe a method for measuring team resilience in response to technological system failures (i.e., automation and autonomy failures), system power-downs, communication outages, and cyberattacks using the concept of system reorganization (Stevens et al., 2016). Reorganization refers to how a team dynamically alters its patterns of interaction, including communication and coordination, across human and technological system layers to adapt and overcome system failures. By measuring reorganization in response to failure perturbations, we aim to create objective metrics for measuring resilience that correlate with established measures of team performance. In addition, by correlating resilience metrics with team performance, we hope to better understand the nature of resilience, in which faster reorganization is hypothesized to correlate with increased team effectiveness. Thus, the primary focus of this paper is to present novel dynamical systems metrics of team resilience and validate them across a series of HAT experiments.

Our method takes a systems approach to team resilience in a remotely piloted aircraft system (RPAS), wherein adaptive solutions must be organized across operators (humans and autonomous agents), user interfaces, and vehicle system layers to overcome failures. Table 1 provides conceptual definitions of system layers as well as other terminology used in the current studies. Table 1 also outlines the theoretical interplay between stability, entropy, and reorganization, such that when an adaptive system is perturbed from equilibrium it has the capacity to reorganize component states to maintain order and function (or discover a new order and function; not explicitly investigated here), at the system level. Relaxation time (Table 1) and its components (described later) are key metrics for measuring the time course of this process.

## Systems Approach

Resilience engineering is relevant to the training and development of effective teams across a variety of settings. Resilience engineering emphasizes how sociotechnical systems of varying sizes, from teams to large organizations, are expected to encounter disturbances, errors, and perturbations, and how these systems flex and adapt to maintain peak performance (Hollnagel et al., 2007). In this light, the development of bottom-up, data-driven approaches to quantify and visualize team resilience that have the potential for real-time resilience analysis are a critical need. We will measure team resilience using metrics based in dynamical systems theory, with the goal of integrating real-time dynamical methods with concepts of team resilience in human factors and resilience engineering.

In resilience engineering, resilience is defined as the "systemic capacity to change

**TABLE 1:** Conceptual Definitions of Terms Used in the Current Studies (Operational Definitions are Provided Later Under Methods and Metrics)

| Terms | Conceptual definitions |
|---|---|
| Entropy and reorganization | Variety of system or layer states. Greater variety has higher entropy. In the current studies, entropy is measured within a (moving) window of time, where entropy fluctuations over time indexes increasing (more reorganization) and decreasing (less reorganization) variety of system states over time. This is called a reorganization time series (Gorman et al., 2020). |
| Failure complexity | Failures based on a single task element (e.g., a single automation or autonomy-related failure) are considered less complex than failures comprised of multiple task elements (e.g., combinations of automation and autonomy failures). |
| Ground truth resilience | An objective score that measures the change in system performance following a failure. In the current studies, it is measured as difference in performance on an RPA ground target when a failure is introduced and performance on the subsequent target. |
| Layered dynamics | A type of functional decomposition that groups all measurable sensor states according to the system layer that generates them. In the current studies, we model the RPAS HAT system using vehicle, communications, and controls layers, each of which is comprised of different sets of sensors; however, the approach is scalable to smaller or larger systems (e.g., Yin et al., 2022). Layered dynamics allows system reorganization to be assessed at the system-level, as well as within different system layers. |
| Relaxation time | The amount of time it takes a system or system layer to reorganize and stabilize following a failure perturbation. In the current studies there are three relaxation time components (initial, peak, end) corresponding to the components of a "resilience curve." |
| Resilience | The capacity of a sociotechnical system (e.g., RPAS HAT) to rapidly enact a response, adapt, and recover from conditions previously outside of its competence envelope (Hoffman & Hancock, 2017). |
| Robustness | The capacity of a system to overcome novel perturbations without degradation of performance. |

[i.e., reorganize] because of circumstances that push the system beyond the [current] boundaries of its competence envelope" (Hoffman & Hancock, 2017, pp. 565–566). The RPAS synthetic task environment is appropriate for analyzing team resilience because it allows for the controlled introduction of different types of technology failures, referred to as perturbations, which are external forces that require a system to reorganize to remain in or find a new stable state (Gorman, Cooke, & Amazeen, 2010). In terms of resilience engineering, perturbations force teams to operate beyond the boundaries of their initial training. We analyze team resilience in the context of failure perturbations that provide a test of a team's ability to adapt to and overcome different types of HAT failures.

Because teams in dynamic environments continuously self-organize new arrangements of parts as they adapt to the changing environment, we view teams as complex adaptive systems (Elliott & Kiel, 2022; McGrath et al., 2000). Therefore, our measures focus on the coordinated behavior that emerges from individual-level interactions, as opposed to the individual-level actions themselves (Amazeen & Amazeen, 2017). When examining the coordinated behavior of human and technological components of a system, resilience can be viewed as the ability for components to mutually adapt when encountering unexpected perturbations and quickly recover to maintain stable and effective system performance. Thus, resilience involves maintaining system performance

across human and technological components to maintain a stable trajectory directed toward accomplishing team goals ("teleological variation," Gorman et al., 2019; Thorén, 2014). The time course of a system to re-stabilize or stabilize in a new state following a perturbation is called relaxation time (Trotsky et al., 2012), which is an index of the system's ability to enact a response, adapt, and recover following a perturbation (Abraham & Shaw, 1992; Mermin, 1970). In the current studies, we use the concept of relaxation time to measure how long it takes a HAT to reorganize following autonomy, automation, and other system failures to identify the reorganization profiles across system layers that correspond to different types of failure perturbations.

### The Current Studies

Our relaxation time metrics of resilience are based on a nonlinear prediction algorithm (Kantz & Schreiber, 1997) and layered dynamics (Gorman et al., 2019). We used these algorithms to measure (a) how quickly a team reorganizes system behavior in response to a perturbation, (b) the novelty of the reorganization, and (c) which system layers (operator communications, controls, vehicle, system overall) reorganize in response to failure perturbations. To examine the association between these resilience metrics and maintaining team effectiveness, we correlated them with objective team performance measures, including a team performance outcome score, a processing efficiency score, and a binary score of whether the team overcame the failure. We also correlated the relaxation time resilience metrics with a ground truth resilience score, which measures the change in the efficiency of taking photos of ground targets (the primary goal of RPAS missions) during and immediately following a failure perturbation. Thus, the team performance metrics and ground truth resilience score provided a test of criterion validity for the relaxation time resilience metrics. The purpose of testing our resilience metrics across different RPAS HAT experiments was to understand how these measures react to automation and autonomy failure perturbations (Experiment 1) and failure perturbations of increasing complexity (Experiment 2), as well as their sensitivity to HAT training manipulations, which were hypothesized to differently impact response to either automation or autonomy failures, as described in the Experiment 2 Methods section. The next section outlines the general method used in both experiments; the details of the participants and procedures of each experiment are separately provided in later sections. Study hypotheses are heavily informed by the design of the dynamical systems resilience metrics; hence, specific hypotheses are presented after the General Method, Measures section.

## General Method

### Overview

Results are reported from two experiments conducted at the Cognitive Engineering Research Institute (CERI) at Arizona State University. The data were collected in the Cognitive Engineering Research on Team Tasks RPAS Synthetic Task Environment (CERTT-RPAS-STE), which simulates teamwork components of RPAS operations and allows for system-level evaluations of these components. The two experiments use the CERTT-RPAS-STE but differ with respect to between- and within-subjects manipulations.

### Materials

The CERTT-RPAS-STE consists of seven hardware consoles (three for task roles, four for experimenters) in which participants and experimenters use a chat interface to communicate (Grimm et al., 2018; McNeese et al., 2018). The task consists of three team-member roles: (1) a navigator who creates the flight plan and sends waypoint restrictions (altitude, airspeed, waypoint name and type, effective radius) to the pilot and photographer; (2) a pilot who monitors and controls vehicle altitude, heading, and airspeed based on the flight plan, and maintains fuel, gears, and flaps settings; additionally, the pilot negotiates with the photographer to achieve required altitude and airspeed to enable successful photographs of target waypoints; and (3) a photographer who controls camera type and

settings, takes target photos, and communicates feedback of the target photo results to the navigator and pilot. Each team member has three screens, including a screen that displays role-specific information, a screen that presents RPA status (e.g., current target; speed; altitude; distance to target), and a chat interface screen. The goal of the team is to fly the RPA through a series of target waypoints (11–20 per mission) to take reconnaissance photos while meeting waypoint restrictions (i.e., acceptable speed/altitude) and to minimize warnings and alarms during a series of 40-min missions.

This research sought to understand resilience in HATs under degraded conditions, which is a term used to specifically refer to automation, autonomy, and malicious attack failures (Cooke et al., 2020). In the current studies, the navigator and photographer were informed that the pilot was an autonomous agent, although the autonomous agent was actually a trained experimenter. Known as the Wizard of Oz paradigm (WoZ; Kelley, 1983), this technique was used to introduce autonomy failures in a controlled manner rather than programming an autonomous agent that failed in controlled ways. Other than introducing autonomy failures, the WoZ pilot emulated the behavior of an actual autonomous agent pilot, known as the synthetic teammate (Ball et al., 2010). The synthetic teammate was developed using Adaptive Control of Thought-Rationale (ACT-R; Anderson et al., 1997) and interacts with human teammates through text chat and is responsible for all taskwork aspects of the pilot role. Prior work with the synthetic teammate revealed limitations of the agent's communication and coordination capabilities (McNeese et al., 2018; Scalia et al., 2022), which were replicated using the WoZ paradigm in the current studies. Therefore, participants (navigator and photographer) were given cheat sheets to assist in effective communication with the WoZ pilot.

## Measures

*Performance Metrics.* We measured team effectiveness using three performance scores. *Team Performance* was a mission-level outcome score, that emphasized the overall ability to successfully photograph targets while accounting for other mission parameters, including time spent in warning/alarm states, number of good photographs, missed targets, and fuel and battery consumption. Teams started each mission with a score of 1,000, and points were deducted based on those parameters. *Overcome* measured how many failures teams successfully overcame, defined as the team successfully photographing the target impacted by the failure. If the team overcame the failure, they received a 1, and if they failed to overcome the failure, they received a 0. Finally, *Target Processing Efficiency* (TPE) measured performance at the target level based on how much time the team spent in the effective target radius to take a photo (shorter times are more efficient). TPE was negatively scored, such that higher scores corresponded to greater efficiency (range = 0–1000). The closer the score to 1,000, the better the TPE; however, there was no *a priori* range regarded as optimally efficient TPE. Team performance and overcome are outcome-based measures, whereas TPE is a process-based measure, as it deducts points for inefficient team processing while in the target radius. Overcome was scored 1 if the team successfully obtained a good photo of the failure target and 0 if not; all other performance scores were generated automatically by the task software.

*Ground Truth Resilience Score.* The ground truth resilience score (GTRS) is a process-based measure of team resilience computed from TPE scores. GTRS measures the performance difference between TPE on the failure target and TPE on the subsequent target. Conceptually, GTRS measures both how much a team is initially impacted by a failure and how well a team recovers following the failure. This score is calculated as the difference between TPE on the failure target and TPE on the following (non-failure) target (Equation (1)).

$$GTRS = TS_{f+1} - TS_f \tag{1}$$

$GTRS$ = ground truth resilience score,
$TS_{f+1}$ = TPE on the target immediately following the failure target,
$TS_f$ = TPE on the failure target.

Although GTRS was intended to measure behavioral resilience, it does not directly measure

how this occurs. For example, if TPE is greatly reduced by a failure, but TPE on the subsequent target returns to a high level, then GTRS would be large, which would fit the concept of resilience as recovery (Woods, 2015). In this case, we should observe a negative correlation between larger GTRS and shorter relaxation times. On the other hand, if a team reorganizes so quickly (shorter relaxation time) that TPE on the failure target remains high, and TPE on the subsequent target also remains high, then GTRS would be small, and this would fit the concept of resilience as robustness (Woods, 2015). It is also possible for low-performing teams, who were poor on the failure target and the subsequent target (i.e., TPE small on both occasions), to obtain a small GTRS. In these latter cases, we should observe a positive correlation between smaller GTRS and shorter relaxation times.

## Dynamical Systems Resilience Metrics

*Layered Dynamics.* We analyzed four layers of RPAS coordination that represent HAT reorganization (Gorman et al., 2019). System layers included (1) communication layer – message sending and receiving among team members through the chat system (i.e., pilot → navigator; navigator → pilot and photographer; etc.); (2) vehicle layer – actions and states of the vehicle (i.e., changes in speed; altitude; fuel; heading; etc.); (3) controls layer – the controls used to interface with the vehicle and other teammates (i.e., changes in pilot's vehicle controls; photographer's camera controls; navigator's route planning controls; etc.); and (4) system layer – overall system state across all layers.

A vector of binary symbols represents the states of all system components, within each layer and the system overall, as a time series (1 Hz). The sensors within the layers—Communication = 9, Vehicle = 9, Controls = 21—comprise an overall RPAS state vector (Figure 1). This overall RPAS vector (i.e., the "system layer") is thus a 39-component vector. For continuous variables, states were determined by mapping the continuous dynamics of components onto a numeric alphabet for symbolic time series modeling (Nicolis & Prigogine, 1989) that preserves the dynamics (e.g., vehicle speed can be represented using four states/symbols: speeding up; slowing down; constant speed; alarm state; Gorman et al., 2019). The purpose of using symbolic dynamics is that by defining the symbols as mutually exclusive and collectively exhaustive symbol sets, we can sum across any collection or sensor states at 1 Hz (e.g., just the vehicle layer vs. the system overall) to efficiently obtain set intersections representing unique layer and system states. This method allows for the efficient computation of changing system and layer states on a second-by-second basis (Gorman et al., 2019).

Although the symbolic alphabet is numeric to allow for summation, we do not assume any ordinal relations (e.g., greater than) among the symbols. As illustrated in Figure 2, it does not matter what the symbols are except that the symbolic time series for each component sensor must be mutually exclusive with all other components, such that summing across component states yields a unique intersection (∩) for every unique system state. In Figure 2, the numeric symbols are binary numbers, with addition through horizontal concatenation. The purpose of using binary numbers is that it facilitates the scalable expansion of the mutually exclusive and exhaustive symbol sets if needed (e.g., if we needed to add another component to the system).

*Reorganization.* The following section describes the calculation of reorganization time series using layered dynamics. All data management and analytic procedures were the same across both experiments.

Entropy (Equation (2)) is a measure of the variety of system states within a window of time (Ashby, 1957), and moving window entropy is a continuous measure of the changing variety ("reorganization") of system states over time (Gorman et al., 2020; Stevens et al., 2016). In equation (2), $p_n$ is the relative frequency of any of the $n$ states in the window multiplied by $\log_2 p_n$. Entropy is used due to its computational efficiency relative to other measures such as recurrence-based determinism (Gorman et al., 2020). Based on prior work, we calculated

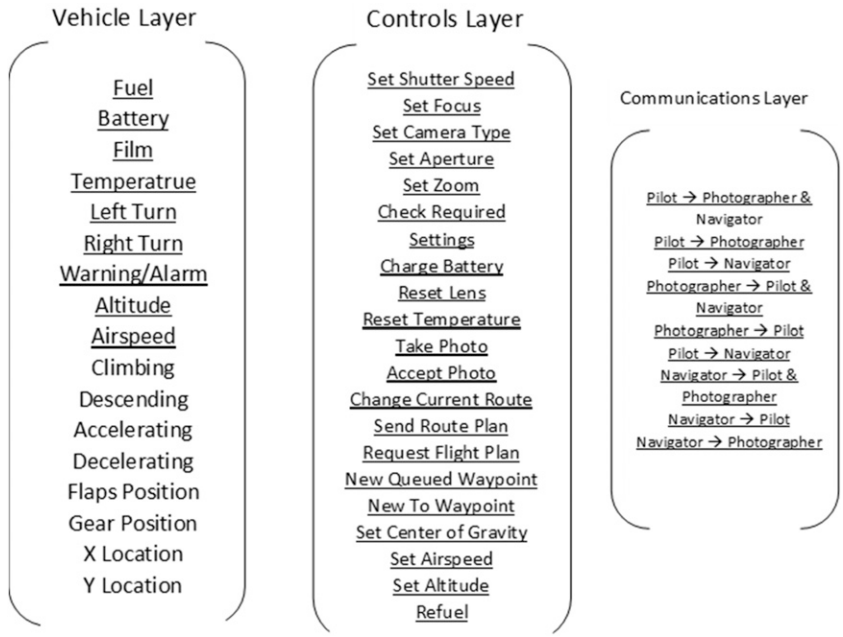



*Figure 1.* Input component signals for the Vehicle, Controls, and Communication layers. Non-underlined component in the Vehicle layer provide redundant information and were not used. Figure adapted from Gorman et al. (2019).

| System Component | Sample # (e.g., 1 Hz) | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 ($q_1$) | 000 | 000 | 100 | 100 | 000 | 000 |
| 2 ($q_2$) | 000 | 010 | 010 | 000 | 010 | 010 |
| 3 ($q_3$) | 110 | 110 | 110 | 000 | 000 | 000 |
| 4 ($q_4$) | 000 | 000 | 000 | 001 | 001 | 001 |
| Binary System State (Q') | 000000110000 | 000010110000 | 100010110000 | 100000000001 | 000010000001 | 000010000001 |
| Decimal System State (Q') | 48 | 176 | 2224 | 2049 | 129 | 129 |

*Figure 2.* Example illustrating symbolic time series using binary symbols for component states ($q_i$ = component i state; 000 = off state) and team state (Q'; component intersections) obtained by summing across (binary addition) component states at each time point (sample). For illustration, this example uses two on/off state for each component; however, the method is generalizable to higher order component states as in the current studies.

reorganization in the system layers and the system overall using a window size of 120 s with a 1 Hz window update rate (e.g., Gorman et al., 2019; Gorman et al., 2020). Using this approach, the more permutations of symbol intersections (i.e., unique states) a system goes through in a window of time, the greater the system variety and, hence, the greater the reorganization in that portion of the time series. Entropy spikes correspond to times of extreme reorganization, and dips in entropy correspond to times of low reorganization. We used Shannon entropy (Shannon & Weaver, 1949; Equation (2)) to calculate continuous system reorganization as the window was slid across the layered dynamics symbolic time series (e.g., Q′ in Figure 2).

$$Entropy = -\sum_{n=1}^{\#sym}(p_n \times \log_2 p_n) \qquad (2)$$

In accordance with the law of requisite variety (Ashby, 1957), we hypothesized (described later) that reorganization would be significantly

larger during failure perturbations compared to routine mission conditions that do not require as much reorganization. This corresponds also to the interpretation of increased entropy as critical variability during phase transitions (Heinzel et al., 2014; Wiltshire et al., 2018), as this increase can be indicative of a team or system transitioning from one state to another. Similar dynamics have been observed in symptom changes among patients with obsessive-compulsive disorder (Heinzel et al., 2014), denoting the generalizability of the approach. Figure 3 provides a visualization of the moving window entropy calculation on binary symbols, using team communication channels as an example.

The purpose of encoding sensor data using mutually exclusive states is that every unique intersection of sensor states (e.g., intersecting a Vehicle state with a Communication state) defines some new state. The possible combinations of sensor states for measuring system state (or layer state, if desired) can be enormous. A conservative estimate of the number of the possible system states in the current studies if each of the 38 sensors take on at most two states would be $2^{38} = 274{,}877{,}906{,}944$ unique system states. It is unlikely that all portions of this state space will ever be visited by the system: Some portions of the state space are likely to be visited more frequently than others (cf. attractors), whereas some portions of the state space may be inaccessible to the system (cf. repellors). Note, however, that the purpose of the present studies is not to enumerate specific states and attractors of the system; we leave that for future research, but to use layered dynamics models to develop generalizable real-time resilience metrics.

*Reorganization Novelty.* We used Kantz and Schreiber's (1997) nonlinear prediction algorithm to quantify reorganization novelty in terms of deviations (root mean square error; RMSE) of the observed reorganization time series from a predicted behavior reorganization time series. RMSE represents how different the current reorganization trajectory is from the

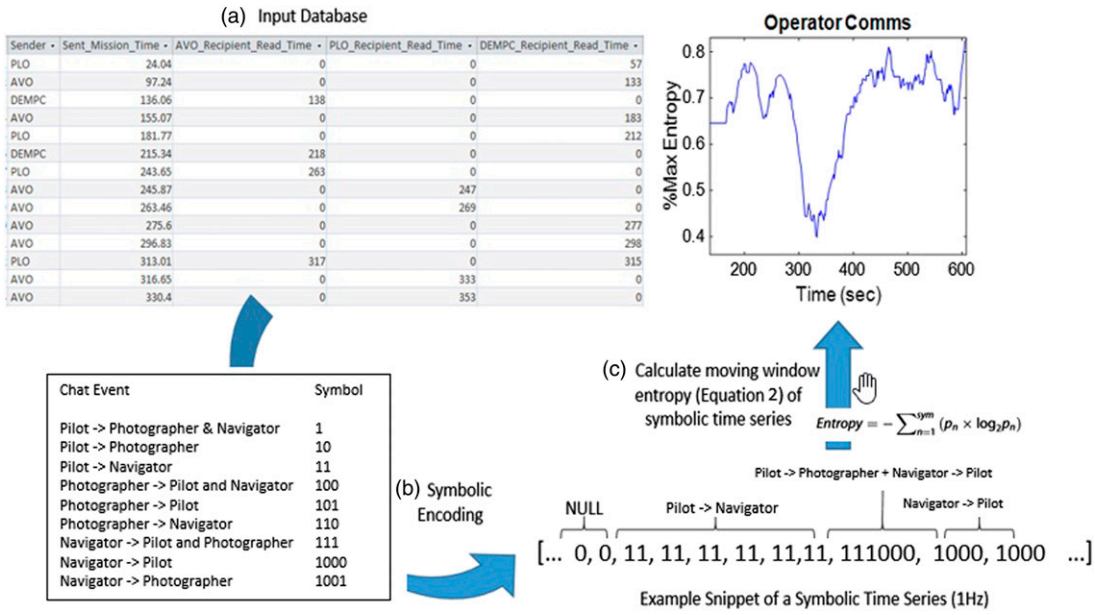



*Figure 3.* Illustration of moving window entropy calculation: (a) input database of text chat events generated by the task software (only time sent and read were used; content was not analyzed); (b) symbolic encoding represents each possible "From-To" chat event as a binary symbol; (c) moving window entropy calculated from (1 Hz) symbolic time series of chat events (higher entropy = more reorganization/variety). Figure adapted from Gorman et al. (2019).

predicted trajectory based on prior reorganization behavior (Figure 3).

For a reorganization (entropy) time series, select the current value, $x_N$, and define a neighborhood, $U_\zeta(x_N)$, of near neighbors, $x_n$, that are within $\varepsilon$ of $x_N$, where $\varepsilon$ is a noise factor. Next, generate predictions for the future evolution of $x_N$ over the next $\Delta n$ time steps (the "prediction horizon"), denoted by $x_{N+\Delta n}$, by taking the points, $x_{ni}$, in $U_\zeta(x_N)$, and following them $\Delta n$ time steps, to obtain a collection of predicted trajectories, $x_{ni+\Delta n}$. Rather than arbitrarily choosing any one predicted trajectory, calculate the ensemble average across the predictions, $\langle x_{ni+\Delta n}\rangle$. To calculate how much the current system trajectory, $x_{N+\Delta n}$, deviates from the ensemble average predicted trajectory, $\langle x_{ni+\Delta n}\rangle$, calculate RMSE = $\sqrt{(x_{N+\Delta n} - \langle x_{ni+\Delta n}\rangle)^2}$.

For the current studies, we set $\varepsilon = 3$ and $\Delta n =$ 20s, which have been shown to be effective for detecting novel system reorganization during perturbations in medical and submarine domains (Gorman et al., 2020; Grimm et al., 2017). RMSE time series were generated for each RPAS mission using the same moving window procedure described previously for entropy (Figure 4).

*Relaxation Time.* Dynamical systems approaches for studying team adaptation typically involve introducing perturbations to determine how the team responds through verbal communication reorganization (e.g., Gorman et al., 2020; Grimm et al., 2017). Using this approach, relaxation time is the time it takes for a team to adapt and recover by reorganizing following a perturbation. If this happens quickly, then the team's relaxation time is shorter. We define relaxation time as being made up of three components in line with the theoretical approach of Hoffman and Hancock (2017). Whereas relaxation time and resilience are often thought of as a singular time to rebound (Woods, 2015), we break it down into three functionally meaningful parts (Initial, Peak, End). In the current studies, we measure these relaxation time components across the sociotechnical system—across communication, vehicle, controls layers, and the system as a whole—in response to failure perturbations.

The first relaxation time component, *Initial*, is how long (in sec.) it takes a team's reorganization time series to exceed a 99% confidence interval (CI) following perturbation onset (described in detail later). The Initial measure operationalizes how quickly the team enacts a reorganization in response to a failure and represents the enaction component of resilience. The second component, *Peak*, is how long (in sec.) it takes reorganization to reach its most extreme value following perturbation onset. The Peak measure operationalizes how quickly the team reaches its maximum point of reorganization and represents the adaptation component of resilience. To parallel Hoffman and Hancock (2017), Initial measures the time to recognize the need for and enact a change,

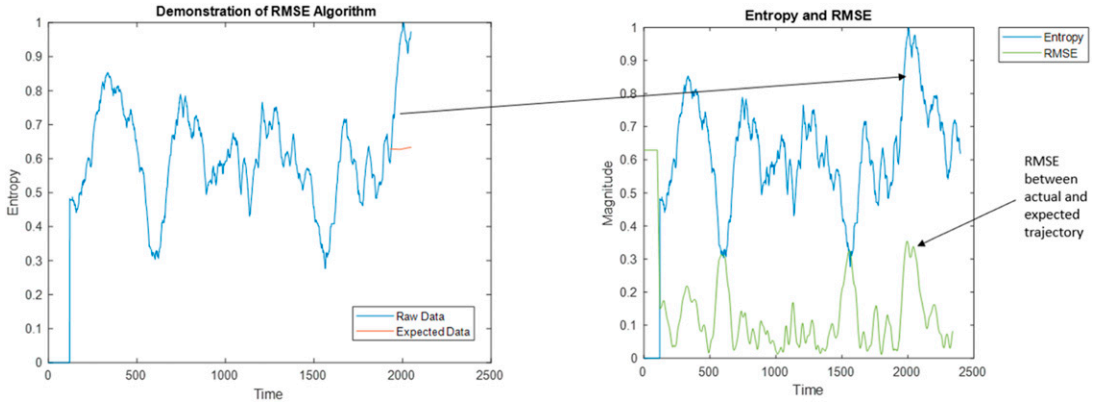
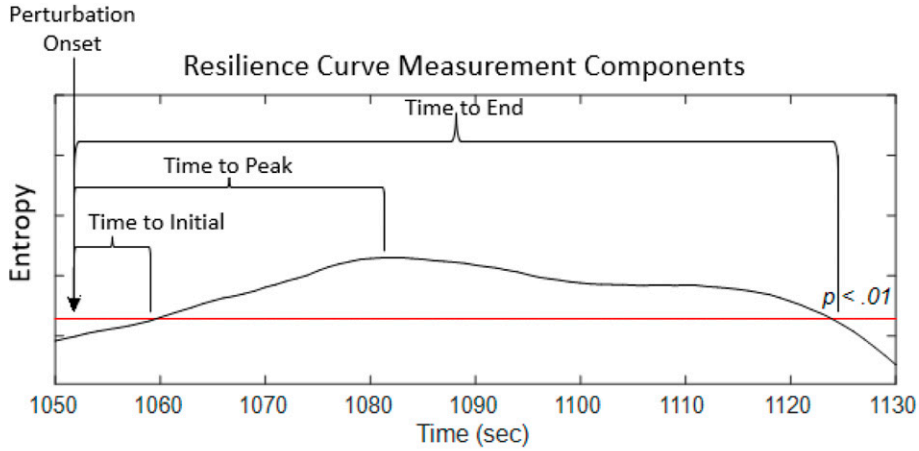



*Figure 4.* Depiction of the RMSE calculation. Larger deviations between observed ("Raw") and predicted ("Expected") yield larger RMSE values indicating greater reorganization novelty.

whereas Peak measures the time to implement the change (i.e., adaptation).

The third component, *End*, measures how long it takes for a team to return to a non-significant level of reorganization following enaction and adaptation. The End measure is defined as the last time point (in sec.) at which the reorganization time series is operating at statistically extreme levels (exceeds 99% CI) following a failure perturbation. This third metric closes the "resilience curve" comprising enaction, adaptation, and recovery (Figure 5), with recovery defined as a return to nominal levels of reorganization. As shown in Figure 5, all relaxation time component measures are calculated relative to a 99% CI computed over the reorganization time series from failure perturbation onset to perturbation offset. The purpose of using the distribution of observations within a failure's duration was to ensure that each of the three relaxation time components (Initial, Peak, End) could be measured for every failure perturbation.

Because the sampling rate was 1 Hz, the number of possible reorganization observations during a perturbation simply corresponds to the duration (in sec.) of the perturbation. Perturbation failure length ranged from 300–420 seconds in Experiment 1 (Cooke et al., 2020) and 300–600 seconds in Experiment 2 (Johnson et al., 2020). Relaxation times were always measured relative to the onset of perturbation, such that more rapidly closing the resilience curve (Figure 5) would result in relaxation times shorter than the full perturbation duration.




| Metric | Operational Definition | Characteristic Measured |
|---|---|---|
| **Time to initial (99% CI red line) entropy** | Time (in sec.) to reach the initial extreme level of reorganization following perturbation onset | **Enaction**; time taken for a team to begin its reorganization behavior |
| **Time to peak (99% CI red line) entropy spike (adaptation time)** | Time (in sec.) taken to reach the maximum amount of reorganization | **Adaptation**; time taken to reach the greatest amount of reorganization behavior |
| **Time to end (99% CI red line) entropy spike (relaxation time)** | Time (in sec.) to return to the normal (non-extreme) level of reorganization | **Recovery;** time taken to return to the routine level of reorganization behavior |

*Figure 5.* Resilience measurement components that complete a "resilience curve." The black trace represents moving window entropy over time, and the red line represents the 99% confidence interval (CI) used to measure Initial, Peak, and End relaxation time components. This figure illustrates resilience metrics for an entropy reorganization time series; however, the process is identical for measuring resilience for an RMSE reorganization novelty time series.

### Study Hypotheses

Our first hypothesis examined how relaxation time metrics relate to maintaining team effectiveness. We hypothesized that shorter relaxation times, which indicate faster enaction, adaptation, and recovery, should be associated with higher team performance.

- Hypothesis 1: Shorter relaxation times (greater adaptive ability/recovery) will be correlated with greater team effectiveness (higher performance scores) across both experiments.

Our second hypothesis was that RPAS HATs should exhibit significantly greater system reorganization during failure perturbations compared to routine mission conditions containing no failures. This hypothesis is akin to the *law of requisite variety* (Ashby, 1957), which states that for a system to maintain effectiveness, the controller ("team") must be able to produce sufficient coordination variety (variety = number of states) to match or exceed the variety demanded by the environment. We further hypothesized that this increase in reorganization behavior would be larger for more effective teams.

- Hypothesis 2a: Teams will exhibit greater reorganization behavior during failure perturbations compared to routine mission segments.
- Hypothesis 2b: This effect will be larger for higher-performing teams.

We examined criterion validity by correlating our resilience metrics with a ground truth resilience score, which measured the impact and subsequent recovery of performance following a failure perturbation. As described earlier, whether the correlation between our resilience metrics and ground truth resilience was negative or positive indicates either the classic form of resilience as recovery or resilience as robustness to perturbation (Woods, 2015). Therefore, our hypothesis with respect to ground truth resilience was non-directional.

- Hypothesis 3: Relaxation times will be correlated with ground truth resilience, with the direction of correlation indicating the nature of resilience (i.e., recovery vs. robustness).

In Experiment 2, teams received different types of training designed to help them overcome either automation failures ("Coordination Coaching") or autonomy failures ("Trust Calibration"), with a third group receiving no special training ("Control"). Because coordination coaching was intended to help teams overcome automation failures, and Trust Calibration was intended to help teams overcome autonomy failures (Johnson et al., 2020), we hypothesized that our resilience metrics would reflect this difference. Specifically, we predicted that relaxation time-performance/GTRS correlations would be stronger for automation failures for Coordination Coaching teams, whereas these correlations would be stronger for autonomy failures for Trust Calibration teams. These two training conditions, Trust Calibration and Coordination Coaching, are described in detail in the Experiment 2 Methods section.

- Hypothesis 4: Teams receiving coordination coaching will display greater resilience in the form of stronger resilience correlations for automation failures, whereas teams receiving trust calibration training will display stronger resilience correlations for autonomy failures.

Hypotheses 1–3 (but not 4) were tested in both experiments. Therefore, in the following results sections we refer to each hypothesis according to its experiment and hypothesis number. For example, Experient 1, Hypothesis 1 is labeled E1.H1, Experient 2, Hypothesis 1 is labeled E2.H1, etc.

## Experiment 1

### Participants

Forty-four participants (22 teams) between 18 to 36 years of age ($M = 23.0$, $SD = 3.90$) were recruited from Arizona State University and surrounding areas. The gender distribution was 21 males and 23 females. Participants were required to have normal or corrected-to-normal vision and fluency in English. All participants were compensated $10 per hour. The experiment was approved by the Cognitive Engineering Research Institute Institutional Review Board.

### Procedure

Experiment 1 took place across two sessions, with a one- to two-week interval between sessions. A trained experimenter was placed in the

pilot role and performed as the autonomous agent in a WoZ paradigm (Kelley, 1983) using a script to mimic actions and communications consistent with the synthetic teammate. Participants were randomly assigned to either navigator or photographer and were instructed that they were working with a synthetic teammate. The experimenter in the synthetic teammate role was in a separate room, and the participants were located together in another room and were separated by a partition. Each participant individually received 30 min of PowerPoint training on the task and their roles. Subsequently, they performed a 30 min hands-on training mission as a team, during which other experimenters used a checklist to ensure that the navigator and photographer were sufficiently trained in their roles.

The first 40-min mission was a baseline mission with no failures. From Missions 2 to 9, there were two failures (one automation and one autonomy) per mission. These failures were introduced to measure team resilience to low-level system automation failures versus perturbations stemming from glitches in the autonomous agent teammate and allowed us to directly compare team behavior during failures to routine mission segments. A malicious cyberattack was introduced during the final 10 minutes of the last mission. Table 2 summarizes the experimental procedure, including the schedule for introducing different types of failures. As described next, automation and autonomy failures each had three types.

### Failure Types

This section describes the three types of automation and autonomy failures. Because the malicious cyberattack occurred only once, it was not included in the inferential statistical analysis of Experiment 1. However, the malicious attack failure was examined in

**TABLE 2:** Procedure for Experiment 1

| | Application of failures during specific targets | | |
|---|---|---|---|
| | Target/ Automation | Target/ Autonomy | Target/Malicious attack |
| Session I (total session with breaks ~6 hours) | | | |
| Consent (15 min) | | | |
| Training - PowerPoint + hands on | No failure | No failure | No failure |
| **Mission I (40 min)** | No failure | No failure | No failure |
| NASA TLX (15 min) | | | |
| **Mission 2 (40 min)** | 2nd/Type I | 4th/Type I | No failure |
| **Mission 3 (40 min)** | 4th/Type II | 2nd/Type II | No failure |
| **Mission 4 (40 min)** | 1st/Type III | 3rd/Type III | No failure |
| NASA TLX-II, trust and anthropomorphisms, and demographics (30 min) | | | |
| Session II (total session with breaks ~7 hours) | | | |
| **Mission 5 (40 min)** | 2nd/Type III | 4th/Type II | No failure |
| NASA TLX I (15 min) | | | |
| **Mission 6 (40 min)** | 4th/Type I | 2nd/Type I | No failure |
| **Mission 7 (40 min)** | 1st/Type II | 3rd/Type II | No failure |
| **Mission 8 (40 min)** | 3rd/Type III | 1st/Type III | No failure |
| **Mission 9 (40 min)** | 3rd/Type II | 1st/Type III | No failure |
| **Mission 10 (40 min)** | 2nd/Type III | 4th/Type III | Last 10 min |
| NASA TLX-II, trust, anthropomorphism, demographics, and debriefing (30 min) | | | |
| Post-check procedure (15 min) | | | |

*Note.* Automation and autonomy failures of different types were implemented in a specified order. Failure types are described in the text.

Experiment 2, due to its importance for examining failure complexity.

*Automation Failures.* The Type I Automation Failure affected the photographer for a total duration of 300 sec. This failure prevented the photographer from viewing current and next target waypoint information, remaining time, distance to the current target, bearing, and course deviation to target, such that the photographer had to obtain that information by communicating with other team members. The Type II Automation Failure affected the pilot for a total duration of 420 sec. This failure prevented the pilot from viewing current altitude and airspeed settings and from entering new altitude and airspeed information, such that the pilot had to obtain that information by communicating with other team members. The Type III Automation failure also affected the pilot for a duration of 420 sec. This failure was more intense than the Type II automation failure. In addition, the pilot was unable to see the remaining time, distance, and bearing to the current target waypoint, such that the pilot had to communicate with other team members to obtain accurate target information. Figure 6 displays an example of a Type II automation failure.

*Autonomy Failures.* The experiment included three types of autonomy failures in which the synthetic teammate pilot failed, each lasting 420 seconds. The Type I Autonomy Failure was a comprehension failure in which a human team member provided information to the synthetic agent, but the agent repeatedly requested the same information due to its inability to comprehend. To overcome this failure, the human team member had to notice the synthetic pilot's incorrect behavior and re-send the correct target waypoint information (i.e., required altitude and airspeed; Cooke et al., 2020). The Type II Autonomy Failure was an anticipation failure in which the synthetic agent did not give the photographer sufficient time to take a good photo and prematurely changed course to the next target. To overcome this failure, the photographer or navigator must notice this failure and instruct the pilot to go back to the target waypoint. The Type III Autonomy Failure was also a comprehension failure in which the synthetic agent failed to understand a message due to its limited communication abilities and misinterpreted target information (altitude, airspeed) from the navigator and photographer. To overcome this failure, the photographer had to repeat the correct information until the pilot correctly adjusted the necessary settings.

*Data Analysis Overview.* To classify reorganization or novelty values as exceeding the critical threshold, we focused on the distribution of observations within the timespan of a failure,


*Figure 6.* Example of a Type II automation failure. The left image displays the pilot's screen during normal (routine) conditions. The right image displays the failures that occurred during the Type II automation failure: The pilot cannot see Altitude and Airspeed and must obtain this information from other team members.

and identified reorganization observations that exceeded the 99% CI of the observations within the timespan of that failure, corresponding to a .01 alpha level (Cohen et al., 2013). From these extreme values, we calculated the relaxation time component metrics (i.e., Initial, Peak, and End) for each failure perturbation. To test H1 (shorter relaxation times are correlated with greater performance), H3 (relaxation times are correlated with GTRS), and H4 (training effects will be present), we correlated each relaxation time component for each system layer (Vehicle, Communications, Controls, System Overall) with all performance scores and GTRS. This was done separately for the reorganization (entropy) and novelty (RMSE) metrics. To test H2a (greater reorganization during failures), we conducted ANOVAs to test for main effects between failure and routine mission segments; to test H2b (that the effect would be larger for higher performing teams), we examined mission segment × performance cluster (low, medium, high) interactions from these ANOVAs.

All correlations and ANOVAs for both Experiment 1 and 2 were conducted using IBM SPSS Statistics (Version 28.0.1.0). To calculate the resilience metrics (relaxation time components, reorganization, RMSE, GTRS, and moving window measures) we used MatLab (Versions 2019–2021a) for both Experiment 1 and 2. All MatLab scripts were written by the authors, except for the entropy function, which was downloaded from the MatLab File Exchange (Dwinnell, 2023).

## Results and Discussion

### Hypothesis E1.H1

This hypothesis predicted that faster relaxation times would correlate with greater team performance. Significant relaxation time—performance metric correlations are presented in Table 3. This table includes system layers, reorganization and reorganization novelty measures (entropy, RMSE), and relaxation time components (Initial, Peak, End). Table 3 reports all significant correlations at the $\alpha = .05$ level. However, given that there were 192 correlations (2 failure types × 2 dynamical system measures × 4 performance measures × 4 layers × 3 relaxation time components), we focus on medium to large effect sizes ($|r| > .3$; Cohen, 1988) to reduce the risk of Type I errors and to assess the correlations in terms of their practical significance (Cumming, 2012, 2014). Prior work by Cumming describes how relying on *p*-values may lead to poor replication due to high variability and a wide range of possible *p*-values, whereas effect sizes perform better in replications under simulated conditions (Cumming, 2008; Cumming, 2014; Cumming & Maillardet, 2006). Additionally, relying on a Bonferroni-corrected $\alpha = \frac{.05}{192} = .00026$ would have increased the probability of Type II error.

Using this criterion, all three relaxation time components were negatively correlated with TPE in the vehicle layer during autonomy failures, with the system layer falling just below our practical significance criterion (all $|r| > .24$). Thus, the vehicle and to some degree the system layer produced consistent correlations in the hypothesized direction for autonomy failures, whereas the results across all other system layers were less consistent. These results provide some support for E1.H1, that faster relaxation times would be correlated with greater team performance, across all three relaxation time components (Initial, Peak, End). The positive vehicle and overall system correlations were also sizable with respect to GTRS for autonomy failures.

### Hypothesis E1.H2

Hypothesis 2a was that teams would display greater reorganization during failure perturbations compared to routine mission segments, and Hypothesis 2b was that this effect would be larger for more effective teams. To test this hypothesis, we calculated average system layer entropy separately for routine, automation failure, and autonomy failure segments of each mission. We obtained $n = 788$ average entropy values (9 missions × 22 teams × 4 layers; four observations were missing due to a file that failed to save) for each level of failure status (routine, automation, autonomy). To test the team effectiveness hypothesis (H2b), we clustered (*k*-means) teams on TPE, Team Performance, and Overcome, to classify low, medium, and high-performing teams across the three performance scores. We then analyzed mean

**TABLE 3:** Experiment 1 Results: Significant Relaxation Time Correlations

| Failure type | Dynamical reorganization measure | Performance measure | Layer | Relaxation time component |
|---|---|---|---|---|
| Automation failure | Entropy | Ground truth resilience score (GTRS) | Vehicle | Initial ($r$ = .153, $p$ = .041) |
| | | Team performance (mission level) | Vehicle | Initial ($r$ = .205, $p$ = .004)<br>Peak ($r$ = .164, $p$ = .023)<br>End (r = .153, $p$ = .034) |
| | | Overcome | Vehicle | Initial ($r$ = .153, $p$ = .041) |
| | | | Control | Initial ($r$ = .166, $p$ = .022)<br>Peak ($r$ = .166, $p$ = .021)<br>End ($r$ = .176, $p$ = .015) |
| | | | System | Initial ($r$ = .161, $p$ = .025)<br>Peak ($r$ = .154, $p$ = .033)<br>End ($r$ = .146, $p$ = .043) |
| | RMSE | Team performance (mission level) | Vehicle | Peak ($r$ = −.160, $p$ = .029)<br>End ($r$ = −.162, $p$ = .027) |
| | | | Control | Initial ($r$ = −.158, $p$ = .031)<br>Peak ($r$ = −.159, $p$ = .031)<br>End ($r$ = −.151, $p$ = .040) |
| | | Overcome | System | Initial ($r$ = .172, $p$ = .019)<br>Peak ($r$ = .173, $p$ = .018)<br>End ($r$ = .172, $p$ = .019) |
| Autonomy failure | Entropy | Target processing efficiency (TPE) | Vehicle | **Initial ($r$ = −.317, $p$ < .001)****<br>**Peak ($r$ = −.333, $p$ < .001)****<br>**End ($r$ = −.332, $p$ < .001)**** |
| | | | System | Initial ($r$ = −.270, $p$ < .001)**<br>Peak ($r$ = −.244, $p$ = .002)*<br>End ($r$ = −.260, $p$ < .001)** |
| | | Ground truth resilience score (GTRS) | Vehicle | Initial ($r$ = .298, $p$ < .001)**<br>Peak ($r$ = .202, $p$ = .016)<br>End ($r$ = .201, $p$ = .016) |
| | | | System | Initial ($r$ = .260, $p$ = .002)*<br>Peak ($r$ = .246, $p$ = .003)*<br>End ($r$ = .266, $p$ = .001)* |
| | RMSE | Overcome | Communication | End ($r$ = −.159, $p$ = .040) |

*Note.* Significant correlations of relaxation time metrics with outcome measures. Medium to large correlations are in bold, with asterisks denoting the following: *$p$ < .01, **$p$ < .001.

entropy using a 3 (Performance Cluster [Low, Medium, High]) × 3 (Failure Status [Routine, Automation, Autonomy]) mixed Analysis of Variance (ANOVA), with Performance Cluster as a between-subjects factor and Failure Status as a within-subjects factor.

The main effect of Failure Status was significant, $F$ (1.78, 1237.56) = 49.45, $p$ < .001, $\eta_p^2$ = .066 (Greenhouse-Geisser correction used). Post-hoc Least Significant Difference (LSD) comparisons revealed that teams exhibited significantly greater reorganization during autonomy failures compared to automation failures ($p$ = .003) and routine mission segments ($p$ < .001) and greater reorganization during automation failures compared to routine mission segments ($p$ = .008). These results support E1.H2a (Figure 7). This result suggests that teams display greater reorganization in response to increasing demands for system variety caused

by failure perturbations. The interaction between Failure Status and Performance Cluster was not significant, $F$ (3.551, 1237.56) = .209, $p$ = .917, $\eta_p^2$ = .001, indicating that low, medium, and high-performing teams exhibited similar amounts of increased entropy (reorganization) in responding to failures compared to routine mission segments. These results do not support E1.H2b, that the effect of greater reorganization during failure perturbations would be larger for higher performing teams. The Performance Cluster main effect was not significant, $F$ (2, 697) = .363, $p$ = .695, $\eta_p^2$ = .001, such that low, medium, and high-performing clusters displayed similar entropy levels overall.

### Hypothesis E1.H3

This hypothesis predicted that relaxation times would be correlated with GTRS, with the direction of correlation suggesting the nature of resilience. Table 3 shows significant correlations between vehicle and overall system entropy and GTRS (difference between failure target TPE and subsequent target TPE), all in the positive direction. Close inspection of the autonomy failure data revealed teams that performed well on both the failure target and subsequent target and, therefore, had a small GTRS. Conversely, there were teams that performed poorly on the failure target but recovered and performed well on the subsequent target and, therefore, had a large GTRS (see Table 4). Based on this pattern of findings, teams that are relatively unaffected by failures (high TPE on failure target) followed by high TPE on the subsequent target display robustness. Conversely, teams that are negatively impacted by failures (low TPE on failure target) but subsequently score high on the follow-up target display recovery. This interpretation of the data was corroborated by the large negative correlation between failure target TPE and GTRS, $r$ (140) = −.634, $p$ < .001.

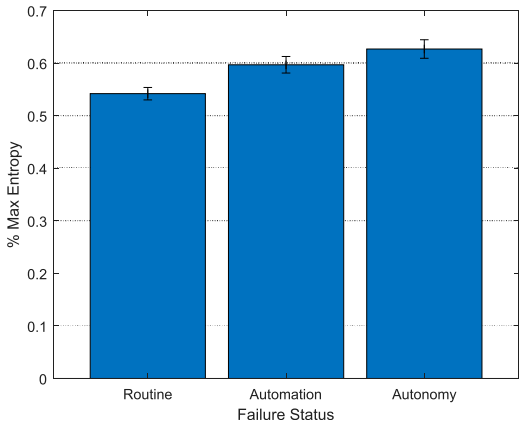



*Figure 7.* Main effect of Failure Status on reorganization behavior (% max entropy) in Experiment 1. Error bars represent standard errors.

Considering that shorter relaxation times were generally correlated with higher TPE (Table 3), Figure 8 illustrates the empirical relationships underlying the positive correlation between relaxation time and GTRS. This positive correlation undergirds two interpretations of resilience in the current study, resilience as robustness and resilience as recovery (Woods, 2015).

## Experiment 2

Experiment 1 revealed that the dynamical systems resilience metrics were more sensitive to autonomy failures versus automation failures, in terms of the resilience-performance correlations. Experiment 1 also indicated separate interpretations of resilience using the metrics: resilience as robustness versus resilience as recovery and that the reorganization profiles suggested that autonomy failures required greater reorganization than automation failures,

**TABLE 4:** Sample Data Points for Target Processing Efficiency (TPE) Scores on Autonomy Failures to Illustrate the Generation of Large and Small Ground Truth Resilience Scores (GTRS)

| | TPE on autonomy failure | TPE on target following failure | GTRS |
|---|---|---|---|
| High performance on autonomy failure | 938.49 | 974.24 | 35.75 |
| Low performance on autonomy failure | 622.4 | 935.21 | 312.81 |

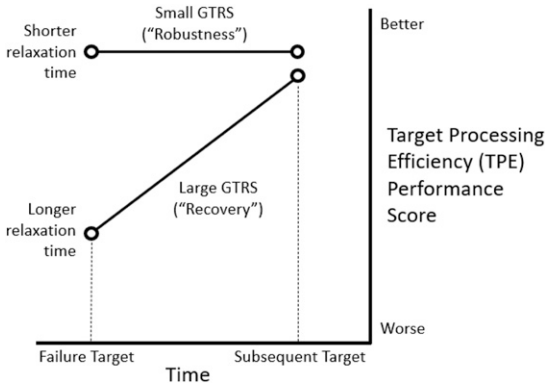



*Figure 8.* Graph of the relationships between relaxation time, target processing efficiency (TPE), and ground truth resilience score (GTRS) underlying the interpretation of resilience as robustness versus recovery.

which in turn required greater reorganization than routine mission segments. Taken together, these results suggest that although they were of similar time durations (i.e., all failures were 420 sec except for Automation Type I, which was 300 sec), autonomy failures may have been more complex than automation failures, in that they required greater amounts of system reorganization by the team. Experiment 2 further investigates this effect by introducing even more complex failures in the form of hybrid automation-autonomy failures, system powerdowns, and malicious cyberattacks. Experiment 2 was also designed to test separate training strategies for increasing resilience to automation and autonomy failures, providing the opportunity to examine the sensitivity of the dynamical systems resilience metrics to differences in team training.

## Participants

Sixty participants (30 teams) between 18 to 33 years of age ($M$ = 22.6, $SD$ = 3.61) were recruited from Arizona State University and surrounding areas. The sample had a gender distribution of 52 males and 7 females, with one participant not responding. Ten teams were randomly assigned to each of the training conditions. Participants were required to have normal or corrected-to-normal vision and fluency in English. All participants were compensated $10 per hour.

## Procedure

Experiment 2 took place over one session. As in Experiment 1, it used the WoZ paradigm in the CERTT-RPAS-STE with a trained experimenter in the synthetic pilot role and participants randomly assigned to either navigator or photographer. Participants were told that they were working with a synthetic teammate. Like Experiment 1, the experimenter was in a separate room, with the participants located in another room and separated by a partition.

Experiment 2 included a between-subjects training manipulation (Control, Coordination Coaching, Trust Calibration) and three additional types of failures (hybrid, system powerdown, communication cut; described later), with the new types of failures intended to be more complex than the failures in Experiment 1. The Control condition was the standard training used in Experiment 1. In Coordination Coaching, participants were trained to push and pull information with the synthetic pilot in a timely manner. The hands-on training mission included a synthetic "super-pilot" coach that directed information pushing and pulling coordination patterns by instructing participants to send relevant information if it was not sent in a timely manner. The goal of the Trust Calibration condition was to appropriately calibrate participants' trust in the synthetic agent. Participants were informed that the synthetic teammate was "imperfect" and "still under development." There were simulated agent coordination delays during hands-on training, and experimenters reinforced that participants should be persistent in coordinating with the agent. The Coordination Coaching condition was intended to increase effectiveness in responding to automation failures, whereas the Trust Calibration condition was intended to increase effectiveness in responding to autonomy failures (Johnson et al., in press). Table 5 describes the training condition manipulations used in this experiment.

The current study examines resilience-performance correlations as a function of training condition; Johnson et al. (in press)

presents the details of the training hypotheses and their direct impact on team process and the team performance metrics. Table 6 shows the Experiment 2 procedure.

### Failure Types

Experiment 2 included Automation Type III and Autonomy Type III failures as previously described. Other failures included malicious cyberattacks, hybrid failures, system power down failures, and communication cuts.

*Malicious Cyberattacks.* Malicious cyberattacks, introduced during the final 10 minutes of the final mission, simulated the synthetic agent being hijacked through cyberattack resulting in the agent providing false information detrimental to mission completion. In addition, the synthetic agent pilot attempted to fly the RPA to an enemy-designated waypoint. To overcome this failure, either the navigator or photographer had to explicitly inform Intelligence (an experimenter) that the RPA was off-route and was flying toward an enemy-designated area via chat message.

*Hybrid Failure.* The hybrid failure was a combination of the Type II automation failure and Type II autonomy failure from Experiment 1. The Type II automation failure affected the pilot, wherein the pilot was not able to view the altitude and airspeed for the next target and had to communicate with the navigator and photographer to achieve proper airspeed and altitude. The Type II autonomy failure portion was an anticipation failure, wherein the pilot began flying to the next waypoint before the photographer could take a photo of the target waypoint. To overcome this failure, the team needed to negotiate the proper settings with the synthetic pilot (Johnson et al., 2020). Since this is a combination of an automation and autonomy failure, the solution would naturally be a combination of solutions to these specific failures as described in Experiment 1.

*System Power Down Failure.* This failure simulated a system power down and rebooting during the mission. During this failure, there was a gradual power down of all screens over the course of 330 sec. The screens were powered down in order from pilot→navigator→photographer. After each team member lost their common information screen, the sequence repeated with each team member losing their role-specific screen. The screens then rebooted in reverse order. The photographer was still able to

**TABLE 5:** Training Condition Manipulations

| Training condition | Manipulations to training | Goal of training | Targeted failure type |
|---|---|---|---|
| Coordination coaching | Participants were informed of teammates' informational needs and encouraged to expediently push and pull information during interactive slideshow training; the synthetic pilot pushed and pulled information across the team members in a timely manner to "coach" coordination during hands-on training. | Improve the speed at which the team coordinates and sends information to one another | Automation |
| Trust calibration | Participants were informed that the synthetic pilot is "imperfect" and "under development" during interactive slideshow training; during the training mission, the synthetic pilot experiences delays in responding, and participants were encouraged to be persistent in communicating with the agent. | Calibrate participants' expectations of the synthetic teammate's abilities and limitations | Autonomy |

**TABLE 6:** Procedure for Experiment 2

| | Condition 1: Control | Condition 2: Coordination coaching | Condition 3: Trust calibration |
|---|---|---|---|
| Consent (15 min) | | | |
| **Training- PowerPoint (40 min)** | Control: Filler | Automation: + push/pull | Autonomy: Calibration of expectations |
| **Training – Hands-on (40 min)** | Standard | Super-AVO/pilot + push/pull coach | Faulty-AVO/pilot + persistence coach |
| Mission I (40 min) | No failure | No failure | No failure |
| Mission 2 (40 min) | 2nd/Automation (type I) | 2nd/Automation (type I) | 2nd/Automation (type I) |
| | 4th/Autonomy (type I) | 4th/Autonomy (type I) | 4th/Autonomy (type I) |
| Mission 3 (40 min) | 3rd/Automation (type III) | 3rd/Automation (type III) | 3rd/Automation (type III) |
| | 1st/Autonomy (type III) | 1st/Autonomy (type III) | 1st/Autonomy (type III) |
| Mission 4 (40 min) | 2nd/Hybrid (automation II and autonomy II) | 2nd/Hybrid (automation II and autonomy II) | 2nd/Hybrid (automation II and autonomy II) |
| | 4th/Communication | 4th/Communication | 4th/Communication |
| Mission 5 (40 min) | 2nd/System power down | 2nd/System power down | 2nd/System power down |
| | 4th/Malicious attack | 4th/Malicious attack | 4th/Malicious attack |
| **Debrief, trust and anthropomorphism questionnaires** | | | |

take a successful photo (until the last screen lost power) if the team adapted in a timely manner to ensure that all necessary information was provided to the affected team member before losing power.

*Communication Cut Failure.* In the communication cut failure, communication from the photographer to the pilot was cut; however, the pilot to photographer link remained active. Because the pilot's communication links to the photographer and navigator remained intact, the pilot was unaware of the communication cut. To overcome this failure, the photographer had to communicate through the navigator to relay information to the pilot.

## Results and Discussion

### Hypothesis E2.H1

Significant relaxation time—performance correlations across failure types, measures, system layers, and relaxation time components are shown in Table 7. Medium to large correlations are shown in bold.

As in Experiment 1, although significant, the automation failure correlations did not meet the medium-to-large effect size criterion. In addition, the communication cut correlations were not significant. All performance correlations were in the hypothesized direction except for the autonomy failure, for which we found positive correlations between relaxation time in the vehicle layer for team performance. Unlike the performance measures that correlated in the hypothesized direction (i.e., Overcome and TPE), the team performance outcome measure was taken at the mission-level rather than the failure target level, perhaps contributing to this unexpected result. As indicated by the changing patterns of medium to large correlations across system layers and failure types, these results suggest that specific patterns of relaxation times and system reorganization across system layers depend on failure type. The positive correlations between relaxation times and GTRS replicated the finding from Experiment 1 and are discussed later, under Hypothesis E2.H3.

**TABLE 7:** Experiment 2 Results

| Failure type | Dynamical system measure | Outcome measure | Layer | Time Point(s) |
|---|---|---|---|---|
| Automation failure | Entropy | Target processing efficiency (TPE) | Vehicle | Initial ($r = -.282$, $p = .026$)<br>Peak ($r = -.281$, $p = .027$)<br>end ($r = -.285$, $p = .025$) |
| | RMSE | No significant findings | | |
| Autonomy failure | Entropy | Team performance (mission level) | System | Initial ($r = .276$, $p = .034$)<br>Peak ($r = .260$, $p = .047$) |
| | RMSE | Team performance (mission level) | Vehicle | **Initial ($r = .328$, $p = .014$)**<br>**Peak ($r = .330$, $p = .013$)**<br>End ($r = .287$, $p = .032$) |
| Hybrid failure | Entropy | Team performance (mission level) | Vehicle | **Peak ($r = -.374$, $p = .038$)**<br>**End ($r = -.357$, $p = .048$)** |
| | | | System | **Initial ($r = -.371$, $p = .040$)** |
| | | Overcome | Communication | **Initial ($r = -.368$, $p = .039$)**<br>**Peak ($r = -.377$, $p = .033$)**<br>**End ($r = -.427$, $p = .015$)** |
| | RMSE | Target processing efficiency (TPE) | Communication | **Initial ($r = -.522$, $p = .011$)**<br>**Peak ($r = -.522$, $p = .011$)**<br>**End ($r = -.524$, $p = .010$)** |
| | | Ground truth resilience score (GTRS) | Communication | **Initial ($r = .512$, $p = .012$)**<br>**Peak ($r = .513$, $p = .012$)**<br>**End ($r = .513$, $p = .012$)** |
| Communication cut | No significant findings | | | |
| System failure (power down) | Entropy | Target processing efficiency (TPE) | System | **Initial ($r = -.395$, $p = .028$)**<br>**Peak ($r = -.400$, $p = .026$)**<br>**End ($r = -.394$, $p = .028$)** |
| | | Ground truth resilience score (GTRS) | Control | **Initial ($r = -.389$, $p = .031$)** |
| | RMSE | No significant findings | | |

*(Continued)*

**TABLE 7:** (Continued)

| Failure type | Dynamical system measure | Outcome measure | Layer | Time Point(s) |
|---|---|---|---|---|
| Malicious attack | Entropy | Team performance (mission level) | Vehicle | **Initial ($r = -.521$, $p = .003$)*** |
| | | | | **Peak ($r = -.532$, $p = .002$)*** |
| | | | | **End ($r = -.437$, $p = .016$)** |
| | | Overcome | System | **Initial ($r = -.381$, $p = .035$)** |
| | | | | **Peak ($r = -.400$, $p = .026$)** |
| | | | Vehicle | **Initial ($r = -.356$, $p = .049$)** |
| | | | | **Peak($r = -.377$, $p = .037$)** |
| | RMSE | Target processing efficiency (TPE) | Vehicle | **Initial ($r = -.410$, $p = .034$)** |
| | | | | **Peak ($r = -.408$, $p = .035$)** |
| | | | | **End ($r = -.430$, $p = .025$)** |
| | | Ground truth resilience score (GTRS) | Communication | **Initial ($r = .521$, $p = .013$)** |
| | | | | **Peak ($r = .520$, $p = .013$)** |
| | | | | **End ($r = .509$, $p = .016$)** |
| | | Overcome | Vehicle | **End ($r = -.431$, $p = .017$)** |
| | | | System | **Initial ($r = -.464$, $p = .010$)** |
| | | | | **Peak ($r = -.466$, $p = .009$)*** |
| | | | | **End ($r = -.468$, $p = .009$)*** |

*Note*. Significant correlations of relaxation time metrics with outcome measures across all failure types (Automation, Autonomy, Communication Cut, System Power Down, Malicious Attack). Medium to large correlations are in bold, with asterisks denoting the following: $^{*}p < .01$, $^{**}p < .001$.

### Hypothesis E2.H2

Hypothesis 2a was that teams would display greater reorganization during failure perturbations compared to routine mission segments, and Hypothesis 2b was that this effect would be larger for higher-performing teams. We carried out the same analysis as for Experiment 1 by conducting a *k*-means cluster analysis to identify low, medium, and high-performing teams, calculating average entropy according to failure status, and running a mixed ANOVA on the average entropy values. We obtained $n = 348$ average entropy values (29 teams × 4 layers × 3 failure complexity [described below]; one team was missing due to a file that failed to save). The Experiment 2 analysis also included Training Condition as a between-subjects factor. Because each mission had two failures (Table 6), we classified the various failure types as Failure One and Failure Two, constituting a within-subject variable, Failure Status. We analyzed the average entropy scores using a 3 (Performance Cluster [Low, Medium, High]) × 3 (Failure Status [Routine, Failure One, Failure Two]) × 3 (Training Condition [Control, Coordination Coaching, Trust Calibration]) mixed ANOVA. Due to the different failure types included in the Failure One and Failure Two factor, we included a covariate, Failure Complexity, that indexed the different possible failure combinations.

There was a main effect of Failure Type, $F(1.702, 549.762) = 17.373, p < .001, \eta_p^2 = .051$. Pairwise comparisons (Bonferroni) showed that Failure Two resulted in significantly greater reorganization than Failure One ($p = .048$) and Routine ($p < .001$), and Failure One displayed significantly more reorganization than Routine ($p = .002$). Figure 9 displays this effect. The Performance Cluster effect, Training Condition effect, and all interactions with these factors were non-significant. As in Experiment 1, these results support the increased variety in response to failure perturbation (main effect, E2.H2a) portion of Hypothesis 2. Also, as in Experiment 1, the portion of Hypothesis 2 that predicted that this effect would be larger for higher performing teams (interaction, E2.H2b) was not supported.

The Failure Status × Failure Complexity (covariate) interaction was significant, $F(1.690, 544.262) = 18.587, p < .001, \eta_p^2 = .055$ (Greenhouse-Geisser correction used). We conducted a simple effects analysis, examining the effect of Failure Type at each level of Failure Complexity. At Complexity = 1 (Missions 2 and 3), we found that the autonomy failure resulted in greater reorganization than the automation failure ($p < .001$) and routine ($p < .001$) mission segments, which replicates the Experiment 1 finding. However, at Complexity = 2 (Mission 4) and Complexity = 3 (Mission 5), there were no significant differences between Failure One and Failure Two. As shown in Figure 10, although high complexity failure types resulted in higher than routine entropy, autonomy failures resulted in the greatest amount of reorganization.

### Hypothesis E2.H3

The three relaxation time metrics resulted in medium-to-large correlations with GTRS (Table 7), supporting Hypothesis 3. The Hybrid and Malicious Attack failures resulted in large correlations in the Communication layer and medium correlations in the Controls layer for the system power down. As in Experiment 1, these

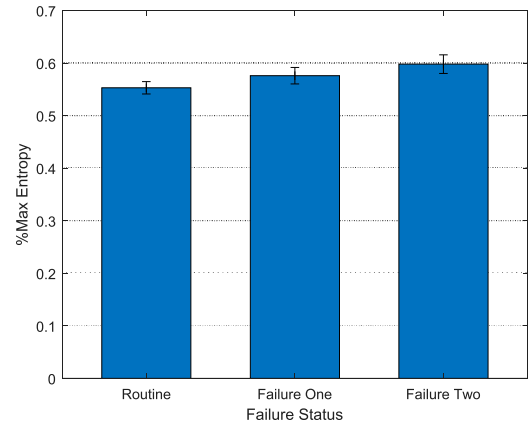



*Figure 9.* The main effect of Failure Type in Experiment 2 shows that failure perturbations resulted in greater entropy (reorganization) compared to routine mission segments. Failure 2 entropy was also significantly greater than Failure 1 entropy. Error bars represent standard errors.

correlations were all positive. We further inspected the relationship between TPE and GTRS by comparing TPE scores on the failure target with TPE scores on the follow-up target. Table 8 shows example scores for the Hybrid and Malicious Attack failures, demonstrating the same pattern observed in Experiment 1.

The pattern of TPE scores in Table 8 predicts a strong, negative correlation between failure target TPE and the resulting GTRS, due to the mapping of high TPE on the failure target onto small GTRS ("robustness") and low TPE on the failure target onto large GTRS ("recovery"). Another possibility that must be accounted for is that low TPE on both the failure target and the follow-up target ("non-resilient") can result in small GTRS. In that case, we would expect a smaller, insignificant correlation between TPE on the failure target and GTRS, due to the inconsistent mapping between TPE failure target score (equal mix of high and low TPE) onto GTRS. Indeed, the correlations between TPE on the failure target and GTRS were strong and negative for both the Hybrid Failure, $r\,(21) = -.669$, $p < .001$, and Malicious Attack, $r\,(20) = -.755$, $p < .001$, indicating that our results were primarily due to a mix of the robustness and recovery forms of resilience. This same pattern of correlations between TPE and GTRS was observed across experiments, despite

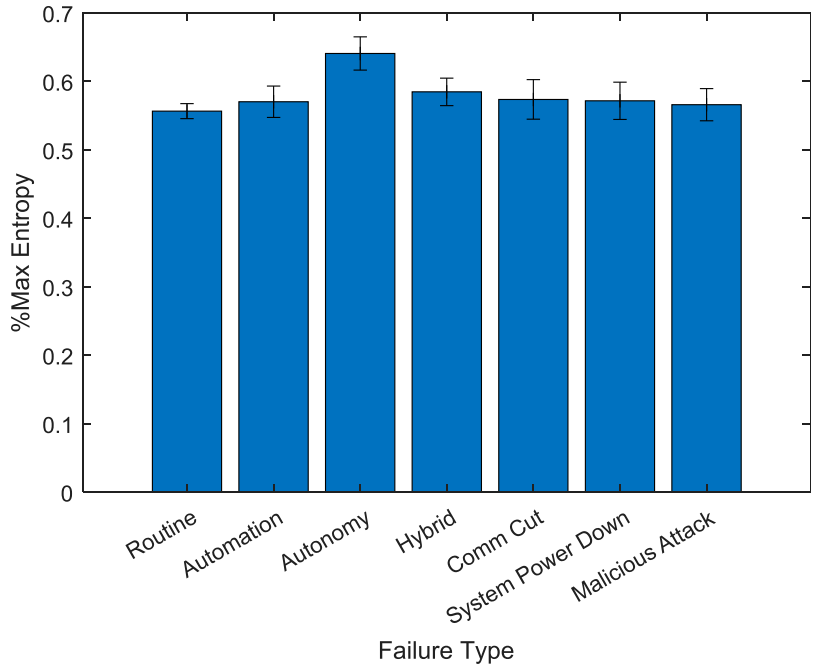



*Figure 10.* The amount of reorganization (entropy) differed as a function of failure type, with autonomy failures producing the greatest amount of entropy. Error bars represent standard errors.

**TABLE 8:** Sample Data Points for Target Processing Efficiency (TPE) Scores on Hybrid Failure and Malicious Attack to Illustrate the Generation of Large and Small Ground Truth Resilience Scores (GTRS)

| | TPE on failure target | TPE on target following failure | GTRS |
|---|---|---|---|
| High performance on hybrid failure | 948 | 927.65 | −20.35 |
| Low performance on hybrid failure | 725.15 | 880.08 | 154.93 |
| High performance on malicious attack | 974.15 | 979.79 | 5.64 |
| Low performance on malicious attack | 458.08 | 852.96 | 394.88 |

the different experimental manipulations across the two experiments. As explained later (Table 11), low GTRS scores resulting from the "robustness" pattern of TPE scores were approximately 47% more frequent than low GTRS resulting from the "non-resilient" pattern of TPE scores across both experiments.

### Hypothesis E2.H4

We hypothesized that teams trained in the Coordination Coaching condition would display stronger, negative correlations between relaxation time and performance when overcoming automation failures and that teams trained in the Trust Calibration condition would display stronger, negative correlations between relaxation time and performance when overcoming autonomy failures. Thus, we examined how these differently trained teams' relaxation times correlated with all performance metrics and GTRS (Table 9).

Counter to our hypothesis, Trust Calibration teams exhibited strong correlations for automation failures rather than autonomy failures, although they were in the predicted direction. As hypothesized, Coordination Coaching teams exhibited strong performance correlations for automation failures, although not in the predicted direction. Specifically, as in Experiment 1 we observed negative correlations for the target-level performance variables (TPE, Overcome) but some correlations were positive for the mission-level performance variable (team performance). In addition to the strong automation failure correlations, Coordination Coaching teams unexpectedly exhibited strong negative correlations for autonomy failures. That these correlations involving GTRS were negative suggests that this effect may be interpreted as resilience as recovery. In sum, these results only partially support Hypothesis 4, primarily due to the unexpected direction of effects and locus of resilience (i.e., automation vs. autonomy) across these training conditions. These results indicate that the resilience metrics are sensitive to differences in training conditions, although the exact nature of this relationship remains to be determined. Table 10 provides a summary of findings across both studies.

## General Discussion

Although they may occur infrequently compared to routine conditions, failure perturbations are to be expected in dynamic HAT environments. These are often high-stakes, one-of-a-kind events that require enaction, adaptation, and recovery by HATs. The current studies developed and tested a novel approach for measuring HAT resilience in a controlled laboratory setting. We envision, however, that versions of these metrics will be deployed in operational environments, wherein team adaptation and resilience are critical for safe and effective operations. Based on the current results, these metrics are promising as real-time indicators of when and how HATs resolve failure perturbations.

Consistent with Hoffman and Hancock's (2017) theoretical resilience measurement model, we operationally defined three components of a "resilience curve" (Figure 5), comprising initial ("enaction"), peak ("adaptation"), and final ("recovery") relaxation time components. Considering only medium-to-large relaxation time-performance correlations, we found support for the hypothesis that relaxation times were negatively correlated with better team performance (Hypothesis 1) across both experiments. These results suggest that faster enaction, adaptation, and recovery during failures predict greater team effectiveness. Although the preponderance of correlations supports this hypothesis, mission-level team performance and GTRS (discussed below) indicated positive correlations. Regarding mission-level team performance, this performance measurement was taken at the mission level, whereas the other measures were taken during and just after a failure at the target-level. The mission-level performance score also included the amount of time spent in warning or alarm states. It is possible that faster relaxation time coupled with less time spent in warnings or alarms over the whole mission could account for the positive correlation with mission-level performance.

**TABLE 9:** Experiment 2 Training Effect Correlations

| Training condition | Failure type | Dynamical system measure | Outcome measure | Layer | Time point |
|---|---|---|---|---|---|
| Coordination coaching (hypothesized stronger automation correlations) | Automation failure | Entropy | Team performance (mission level) | Communication | **Initial ($r$ = .609, $p$ = .003)**<br>**Peak ($r$ = .561, $p$ = .007)**<br>**End (r = .640, p = .001)** |
| | | RMSE | No significant findings | | |
| | Autonomy failure | Entropy | No significant findings | | |
| | | RMSE | Ground truth resilience score | Controls | **Initial ($r$ = −.509, $p$ = .031)**<br>**Peak ($r$ = −.507, $p$ = .032)**<br>**End ($r$ = −.514, $p$ = .029)** |
| Trust calibration (hypothesized stronger autonomy correlations) | Automation failure | Entropy | Target processing efficiency (TPE) | Communication | **Initial ($r$ = −.526, $p$ = .037)**<br>**Peak ($r$ = −.522, $p$ = .038)**<br>**End ($r$ = −.535, $p$ = .033)** |
| | | RMSE | Overcome | Communication | **Initial ($r$ = −.672, $p$ = .004)**<br>**Peak ($r$ = −.673, $p$ = .004)**<br>**End ($r$ = −.667, $p$ = .005)** |
| | Autonomy failure | Entropy | No significant findings | | |
| | | RMSE | No significant findings | | |

*Note*. Correlations split across training condition type. Medium to large correlations are in bold.

**TABLE 10:** Summary of Findings Across Experiments 1 and 2

| *Hypothesis* | *Support* |
|---|---|
| 1 Negative correlation between performance metrics and relaxation times. | **Supported:** This relationship was particularly salient in the vehicle and system layers but also observed in the controls and communications layers. |
| 2 Greater reorganization during failure perturbations compared to routine mission segments, such that this relationship would be more pronounced for more effective teams. | **Partially supported:** There was greater reorganization during failures compared to routine mission segments, but this effect was similar across team effectiveness levels. |
| 3 Relaxation times will be correlated with ground truth resilience scores (GTRS). | **Supported:** Relaxation times can reflect resilience as either recovery (negative correlation with GTRS), robustness (strong positive correlation with GTRS), or non-resiliency (weak positive correlation with GTRS). |
| 4 There will be predictable training effects on resilience in terms of the locus of resilience (i.e., automation failure vs. autonomy failure resilience). | **Partially supported**, training condition appears to moderate the resilience – performance relationship; however, the directionality and locus of these effects are hard to predict. |

One potential difficulty with relaxation time metrics is that as failures become increasingly complex (e.g., moving from Experiment 1 to Experiment 2), the locus of resilience (e.g., system layer) may be difficult to predict. In Experiment 1, correlations were found primarily in the vehicle and system layers, whereas in Experiment 2, correlations were found in all system layers. This may reflect the bespoke nature of adaptation and resilience to increasingly complex system failures. Although currently our resilience metrics have the advantage of identifying where and when a system reorganizes across system layers, there is no simple law relating system layers to failure complexity. This may still be a potential benefit to resilience engineering, however, which views teams as large systems containing numerous components with complex interactions that are difficult to predict in perturbed operational settings (Hollnagel et al., 2007). We argue, therefore, that this approach can help identify, if not predict *a priori*, which system layers are key to resilience in dynamic environments.

There is, however, a straightforward law that relates the amount of reorganization to the variety demanded by the environment in terms of maintaining stable system performance, the law of requisite variety (Ashby, 1957). In both experiments, we found support for the hypothesis (Hypothesis 2a) that teams would display significantly greater reorganization behavior during failures compared to routine (nominal) mission segments, regardless of the source (system layers) of reorganization. We additionally hypothesized that this effect would be greater for more effective teams (Hypothesis 2b), such that teams with higher performance scores would exhibit larger differences between failure and routine reorganization. We did not find support for the latter hypothesis. However, the point of this law is that a controller must be able to match the variety required by the system it controls. The question is whether it is effective in doing so. For instance, a poorly designed traffic system takes longer to match the same amount of requisite variety compared to an effectively designed traffic system. In the current study, all HATs were exposed to the same amounts of requisite variety. However, as demonstrated by the relaxation time metrics and support for Hypothesis 1, timeliness of response was key, over and above the amount of variety the system can match. Although not observed in the current studies, it should be noted that Ashby's Law is a key contribution to the inverse-U relationship between controller complexity and performance. This means that if there is *too much* complexity in the controller, then there is wasted effort, which can be a detriment to a team's

efficiency (Boisot & McKelvey, 2011; Friston, 2010; Guastello, 2015; Hong, 2010).

In support of the hypothesis that relaxation times would predict ground truth resilience (GTRS; Hypothesis 3), we found medium to strong correlations between relaxation times and GTRS, largely in the positive direction, in both experiments. Moreover, shorter relaxation times (i.e., faster enaction, adaptation, and recovery) at the failure target were correlated with higher TPE at the failure target, indicating that resilience in the current studies can be interpreted as either robustness or recovery. This suggests that although resilience as recovery may contribute to resilience, it can also be associated with robustness, or the ability to handle increasing complexity at the point of the system failure (Woods, 2015).

Table 11 acknowledges four possible outcomes to explain the relationship between GTRS and the dynamic resilience metrics, including the number of teams across both experiments whose mode GTRS falls into each outcome. First, a team could display high performance on the failure target as well as the subsequent target. This is a high-performing team and one that is both robust and resilient because it effectively handled the complexity of the failure while maintaining high performance on the follow-up target. However, this team would have a low GTRS or possibly negative, which partially explains the positive correlation between relaxation time and GTRS. This was the most frequently occurring pattern in the current studies ($n$ = 29).

A second possibility is that a team performed well on the failure target but performed poorly on the subsequent target. This team could be considered robust initially but possibly lacking in recovery after doing so, thereby failing to perform well on the follow-up target. It is possible, for example, that such teams may expend their energy on the failure target, making it difficult to function effectively afterwards. However, this pattern was not observed in the current studies ($n$ = 0). A third possibility is that a team may perform poorly at the point of failure but performs well on the subsequent target. These teams would have a high GTRS and would be considered resilient in terms of recovery but not robustness. This pattern, which would account for the remainder of the negative correlation between relaxation time and GTRS, was not as frequently observed in the current studies ($n$ = 9).

A final possibility is that teams perform poorly on both the failure target and the subsequent target. Teams in this category would be non-resilient from both a robustness and recovery perspective (Woods, 2015). This pattern, however, does not map onto the strong negative correlations between failure TPE and GTRS. Nevertheless, this pattern was observed in the current studies ($n$ = 19), although 47% less often than the robust and resilient pattern, but it could contribute to positive correlations between relaxation time and GTRS. Thus, relaxation time metrics should be supported by classification metrics to disentangle the formal nature of team

**TABLE 11:** Mode Resilience Profile Classifications Using Ground Truth Resilience Scores (GTRS) for Teams Across Experiments 1 and 2

| Failure target performance | Subsequent target performance | Change between scores (GTRS) | Classification | Number of teams (across experiments) |
|---|---|---|---|---|
| High | High | Low (on average) | Robust and resilient | 29 |
| High | Low | Negative | Initially robust but non-resilient | 0 |
| Low | High | Positive | Recovery after failure | 9 |
| Low | Low | Low (on average) | Non-resilient | 19 |

*Note.* A description of four possible outcomes and corresponding performance classifications when measuring GTRS. We calculated the failure target performance classification per failure during each mission. To aggregate this value to the team level, we took the most frequently occurring target performance category (i.e., the mode) per team.

resilience. The approach used in the current studies could be implemented in real time using real-time TPE scoring. At a minimum, to create these profile classifications, one needs to examine not just the change scores, but the different performance classifications at both the point of failure and subsequent targets (e.g., Table 11).

These findings align with previous research in which teams competed against a sentient attacker (Guastello, 2010; Guastello et al., 2017). Like our studies, that research focused on dynamical systems metrics, which quantified adaptability and resilience using the largest Lyapunov exponent (an index of stability and chaotic behavior). When enemy attacks were making progress, team performance dropped both during the current performance opportunity as well as the subsequent performance opportunity, during which decision making was hampered. However, teams exhibited higher levels of adaptation compared to attackers as measured by larger values of their largest Lyapunov exponents.

With respect to our training hypothesis (Experiment 2, Hypothesis 4), results were mixed. Teams trained in the Coordination Coaching condition demonstrated strong relaxation time correlations when overcoming automation failures as hypothesized, although the direction of correlation (positive) was not in the predicted direction. These teams also demonstrated the hypothesized strong negative relaxation time correlation in response to autonomy failures, which was not hypothesized. Teams trained in the Trust Calibration condition did not demonstrate the hypothesized strong relaxation time correlations when overcoming autonomy failures; however, they did demonstrate strong negative relaxation time correlations when overcoming automation failures, which was not hypothesized. Interestingly, these training effects were most apparent in the communication layer and to a lesser extent in the controls layer. Consistent with other analyses of these training effects demonstrating that coordination training primarily impacts communication, this implies that communication reorganization may be critical when overcoming failures of increasing complexity, and that the relationship between communication reorganization and performance differs depending on training condition (Johnson et al., in press). Overall, Coordination Coaching teams appear to be more resilient in terms of their relaxation time—performance relationships because these relationships were observed for both automation and autonomy failures, whereas Trust Calibration teams only demonstrated these relationships for automation failures.

Given the many definitions of resilience in the literature (Hollnagel et al., 2007; Woods, 2015), as well as recommendations to practitioners on how to apply concepts of resilience engineering in practice (Hollnagel, 2013), there is currently a lack of methods for objectively measuring team resilience in dynamic sociotechnical environments. However, layered dynamics and relaxation time metrics demonstrate the potential for objective, real-time methods that can be used to quantify resilience across system layers that are predictive of various aspects of team performance. The current approach builds on the work of Hoffman and Hancock (2017), who proposed a theoretical approach for measuring resilience. We submit that their proposal, which comprises recognition, design, and implementation of system changes in response to failures, aligns with the operational definitions of enaction (Initial), adaptation (Peak), and recovery (Final) relaxation time metrics that comprise a "resilience curve" in the current studies. Although we do not claim a strict, one-to-one match with their proposed theoretical framework, the concept of measuring resilience as the capacity to overcome a failure and rapidly and efficiently recover to a stable state, we claim captures the essence of their theoretical approach.

The current results may further be tied to theoretical concepts in resilience engineering. Woods (2015) defines four concepts of resilience, including resilience as robustness, resilience as rebound from degraded conditions, resilience as graceful extensibility, and resilience as sustained adaptability. The current results most directly apply to the first three. In relation to ground truth resilience (e.g., GTRS), we argue that relaxation time can be indicative of robustness (performance on both targets is high)

or low-performing, non-resilience (performance on both targets is low) when the correlation is positive and rebound or recovery when negative. Thus, relaxation times are indicative of different theoretical conceptualizations of resilience depending on the robustness of the system and its ability to recover from a failure. The third concept, graceful extensibility, is the ability of a system to extend its capacity in response to novel disturbances. Although we did not aim to directly measure this property in the current studies, the "extensibility" component of this property is arguably represented by the consistently greater system reorganization values observed under failure conditions compared to routine conditions. The "graceful" component may be embedded within the relaxation time—performance correlations, which should be explored in future research. Overall, the aims of the current research involved generating dynamic systems-based resilience metrics, wherein rapid reorganization tends to be associated with better performance, and sublayers can be used to identify the sources of rapid responses and resilient behavior within a system. On those counts, the current results are promising for future applications of these metrics in training and operational settings.

### Limitations and Future Directions

Although we can generally argue that relaxation time metrics measure robustness, recovery, and extensibility, due to the number of system layers, types of failures, relaxation time metrics, and performance measures, there were a very large number of correlations, making a clean interpretation of the results difficult. In the future, it may be beneficial to use a less complicated experimental apparatus with fewer layers and types of failures to parse out the causal relations between perturbations and relaxation times as they relate to resilience. Although it would appear to bely the bespoke nature of unique events that require resilience, a more highly controlled and simpler experiment might help unpack the complex dynamics observed in the current studies.

Additionally, more conceptual work is needed on the two dynamical system measures used, entropy (reorganization) and RMSE (novelty), and how they meaningfully relate to system response. Entropy describes system response in terms of the number of unique states occupied by the system during a span of time. Currently, we think this is analogous to variety in Ashby's (1957) law of requisite variety. On the surface, RMSE captures the novelty of system response in terms of the deviation from a predicted trajectory of the reorganization time series. However, RMSE is agnostic as to the source of a novel trajectory of the entropy time series. Moreover, because it is a square root, RMSE, in terms of novelty, could correspond to extremes of either increasing or decreasing variety. Future work should focus on more exactly tying novel team states to requisite variety that maintain system effectiveness under degraded conditions.

Along these lines, future directions in resilience measurement could disentangle these two metrics based on their respective explanations of a system response. Thus, another future direction involves selectively filtering out sources of variation, such as team members or sublayers, to identify which are critical for reorganization and novelty in a team response (e.g., by using the filtering method described in Gorman et al., 2020). For example, one could filter the control layer from the overall system layer separately for the reorganization and novelty metrics to determine if significant correlations between relaxation time and performance persist across other layers.

Finally, it is worth noting that the relaxation times do not necessarily indicate whether a system is revisiting a previous state, or if it is moving into a new state. Our metrics quantify reorganization and resilience by computing unique states of the system, but they do not currently capture qualitative differences among those states. In this light, there are other metrics available for quantifying resilience based in dynamical systems theory that are relevant to this line of research (Guastello & Gregson, 2011).

## Conclusion

The methodological approaches described in the current paper have the potential to impact the training and assessment of HATs as well as teams in other sociotechnical contexts. These environments can be highly dynamic and are susceptible to system failures, errors, and crises. This work is beneficial in many situations in which team flexibility, preventive behavior, and resilience are critical, and real-time metrics are needed. The metrics developed in this work have strong potential for real-time implementation that would benefit the training of more resilient teams and the design of more resilient systems by providing real-time feedback and guidance during training and simulation, as well as understanding how systems reorganize to maintain high levels of effectiveness during one-of-a-kind, anomalous events (e.g., Gorman et al., 2020; Grimm et al., 2017). Real-time analysis of reorganization and resilience may also enable analysts and operators in operational environments to detect early onset of maladaptive and possibly dangerous team actions. Taken together, we propose that these types of measures will inform and generate new approaches to teamwork measurement, monitoring, and assessment strategies in sociotechnical work domains in which timely and resilient responses are critical.

## Acknowledgments

We acknowledge Paul Jorgeson and Steve Shope for their assistance modifying the CERTT-RPAS-STE; John Flach for contributing to our understanding of requisite variety; and Cody Radigan, Craig Johnson, Sophie He, Matthew Lin, Tanvi Tandolkar, Garrett Zabala, Alexandra Wolff, for data collection efforts. Lastly, we acknowledge Dr. Ronald H. Stevens for inspiring the methodological approaches described in this paper.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by ONR Award N000141712382 (Program Managers: Marc Steinberg; Micah Clark).

## ORCID iDs

Mustafa Demir https://orcid.org/0000-0002-5667-3701

Nathan J. McNeese https://orcid.org/0000-0002-9143-2460

## References

Abraham, R., & Shaw, C. D. (1992). *Dynamics: The geometry of behavior.* Addison-Wesley.

Alliger, G. M., Cerasoli, C. P., Tannenbaum, S. I., & Vessey, W. B. (2015). Team resilience: How teams flourish under pressure. *Organizational Dynamics*, *44*(3), 176–184. https://doi.org/10.1016/j.orgdyn.2015.05.003

Amazeen, P. G., & Amazeen, E. L. (2017). A systems approach to perception and action. *Ecological Psychology*, *29*(3), 213–220. https://doi.org/10.1080/10407413.2017.1330119

Anderson, J. R., Matessa, M., & Lebiere, C. (1997). ACT-R: A theory of higher level cognition and its relation to visual attention. *Human-Computer Interaction*, *12*(4), 439–462. https://doi.org/10.1207/s15327051hci1204_5

Ashby, W. R. (1957). Requisite variety. *An introduction to cybernetics* (pp. 202–218). Chapman and Hall Ltd.

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

Boin, A., & Bynander, F. (2015). Explaining success and failure in crisis coordination. *Geografiska Annaler - Series A: Physical Geography*, *97*(1), 123–135. https://doi.org/10.1111/geoa.12072

Boisot, M., & McKelvey, B. (2011). Complexity in organization-environment relations: Revisiting Ashby's law of requisite variety. In P. Allen, S. Maguire, & B. McKelvey (Eds.), *The Sage handbook of complexity and management* (pp. 279–298). Sage.

Campbell, M., Egerstedt, M., How, J. P., & Murray, R. M. (2010). Autonomous driving in urban environments: Approaches, lessons and challenges. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, *368*(1928), 4649–4672, https://doi.org/10.1098/rsta.2010.0110

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

Colten, C. E., Kates, R. W., & Laska, S. B. (2008). Three years after Katrina: Lessons for community resilience. *Environment: Science and Policy for Sustainable Development*, *50*(5), 36–47. https://doi.org/10.3200/envt.50.5.36-47

Cooke, N., Demir, M., McNeese, N., Gorman, J., & Myers, C. (2020). *Human–autonomy teaming in remotely piloted aircraft systems operations under degraded conditions*. Office of Naval Research Technical Report for Grant No. N00014-17-1-2382.

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*(4), 286–300.

Cumming, G. (2012). *Understanding the new statistics effect sizes, confidence intervals, and meta-analysis (Multivariate applications book series)*. Routledge.
Cumming, G. (2014). The new statistics: Why and how. *Psychological science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966
Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: Where will the next mean fall? *Psychological methods*, *11*(3), 217.
Dwinnell, W. (2023). Entropy. *MATLAB Central File Exchange*. Retrieved.https://www.mathworks.com/matlabcentral/fileexchange/28692-entropy (12 May 2023).
Elliott, E., & Kiel, L. D. (Eds.), (2022). *Complex systems in the social and behavioral sciences: Theory, method and application*. University of Michigan Press.
Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127. https://doi.org/10.1038/nrn2787
Gorman, J. C., Cooke, N. J., & Amazeen, P. G. (2010). Training adaptive teams. *Human Factors*, *52*(2), 295–307. https://doi.org/10.1177/0018720810371689
Gorman, J. C., Demir, M., Cooke, N. J., & Grimm, D. A. (2019). Evaluating sociotechnical dynamics in a simulated remotely-piloted aircraft system: A layered dynamics approach. *Ergonomics*, *65*(5), 629–643. https://doi.org/10.1080/00140139.2018.1557750
Gorman, J. C., Grimm, D. A., Stevens, R. H., Galloway, T., Willemsen-Dunlap, A. M., & Halpin, D. J. (2020). *Measuring real-time team cognition during team training*. Human Factors.
Grimm, D., Demir, M., Gorman, J. C., & Cooke, N. J. (2018). The complex dynamics of team situation awareness in human-autonomy teaming. In 2018 IEEE conference on cognitive and computational aspects of situation management (CogSIMA), Boston, MA, June 2018 (pp. 103-109). IEEE.
Grimm, D., Gorman, J. C., Stevens, R. H., Galloway, T., Willemsen-Dunlap, A. M., & Halpin, D. J. (2017). Demonstration of a method for real-time detection of anomalies in team communication. In Proceedings of the human Factors and ergonomics society 59th annual meeting, Austin, TX, October 2017 (pp. 282–286). Human Factors and Ergonomics Society.
Guastello, S. J. (2010). Nonlinear dynamics of team performance and adaptability in emergency response. *Human Factors*, *52*(2), 162–172. https://doi.org/10.1177/0018720809359003
Guastello, S. J. (2015). The complexity of the psychological self and the principle of optimum variability. *Nonlinear Dynamics, Psychology, and Life Sciences*, *19*(4), 511–527.
Guastello, S. J., & Gregson, R. A. M. (Eds.), (2011). *Nonlinear dynamical systems analysis for the behavioral sciences using real data*. C R C Press/Taylor and Francis.
Guastello, S. J., Marra, D. E., Castro, J., Gomez, M., & Perna, C. (2017). Performance and participation dynamics in an emergency response simulation. *Nonlinear Dynamics, Psychology, and Life Sciences*, *21*(2), 217–250.
Heinzel, S., Tominschek, I., & Schiepek, G. (2014). Dynamics patterns in psychotherapy: Discontinuous changes and critical instabilities during the treatment of obsessive compulsive disorder. *Nonlinear Dynamics, Psychology, and Life Sciences*, *18*(2), 155–176.
Hoffman, R. R., & Hancock, P. A. (2017). Measuring resilience. *Human factors*, *59*(4), 564–581. https://doi.org/10.1177/0018720816686248
Hollnagel, E. (Ed.), (2013). *Resilience engineering in practice: A guidebook*. Ashgate Publishing, Ltd.
Hollnagel, E., Woods, D. D., & Leveson, N. (2007). *Resilience engineering: Concepts and precepts*. Ashgate Publishing, Ltd.
Hong, S. L. (2010). The entropy conservation principle: Applications in ergonomics and human factors. *Nonlinear Dynamics, Psychology, and Life Sciences*, *14*(3), 291–315.
Johnson, C. J., Demir, M., McNeese, N. J., Gorman, J. C., Wolff, A. T., & Cooke, N. J. (in press). *The impact of training on human-autonomy team communications and trust calibration*. Human Factors.
Johnson, C. J., Demir, M., Zabala, G., He, H., Grimm, D., Radigan, C., Wolff, A., Cooke, N., McNeese, N. J., & Gorman, J. (2020). Training and verbal communications in human-autonomy teaming under degraded conditions. In 2020 IEEE conference on cognitive and computational aspects of situation management (CogSIMA), Virtual due to COVID-19, August 2020. IEEE.
Kantz, H., & Schreiber, T. (1997). Determinism and predictability. *Nonlinear time series analysis* (pp. 42–57). Cambridge University Press.
Kayes, D. C. (2004). The 1996 Mount Everest climbing disaster: The breakdown of learning in teams. *Human Relations*, *57*(10), 1263–1284. https://doi.org/10.1177/0018726704048355
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 systems* (pp. 193–196), ACM.
Kruijff, G. J. M., Janíček, M., Keshavdas, S., Larochelle, B., Zender, H., Smets, N. J., & Liu, M. (2014). Experience in system design for human-robot teaming in urban search and rescue. In *Field and service robotics* (pp. 111–125). Springer.
Leonard, H. B., & Howitt, A. M. (2006). Katrina as prelude: Preparing for and responding to Katrina-class disturbances in the United States—testimony to U.S. Senate committee, march 8, 2006. *Journal of Homeland Security and Emergency Management*, *3*(2), 1–20. https://doi.org/10.2202/1547-7355.1246
McGrath, J. E., Arrow, H., & Berdahl, J. L. (2000). The study of groups: Past, present, and future. *Personality and Social Psychology Review*, *4*(1), 95–105. https://doi.org/10.1207/s15327957pspr0401_8
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
Mermin, N. D. (1970). Lindhard dielectric function in the relaxation-time approximation. *Physical Review B*, *1*(5), 2362–2363. https://doi.org/10.1103/physrevb.1.2362
Morgan, P. B., Fletcher, D., & Sarkar, M. (2017). Recent developments in team resilience research in elite sport. *Current opinion in psychology*, *16*, 159–164. https://doi.org/10.1016/j.copsyc.2017.05.013
Nicolis, G., & Prigogine, I. (1989). Randomness and complexity. *Exploring complexity: An introduction* (pp. 147–192). W. H. Freeman and Company.
Salas, E., DiazGranados, D., Klein, C., Burke, C. S., Stagl, K. C., Goodwin, G. F., & Halpin, S. M. (2008). Does team training improve team performance? A meta-analysis. *Human Factors*, *50*(6), 903–933. https://doi.org/10.1518/001872008X375009
Scalia, M. J., Zhou, S., Grimm, D. A. P., Harrison, J. L., & Gorman, J. C. (2022). The role of timing of information front-loading and planning ahead in all-human vs. Human-autonomy team performance. *Proceedings of the Human Factors and Ergonomics Society - Annual Meeting*, *66*(1), 530–534. https://doi.org/10.1177/1071181322661251
Shannon, C., & Weaver, W. (1949). *The mathematical theory of communication*. University of Illinois Press.
Shively, R. J., Lachter, J., Brandt, S. L., Matessa, M., Battiste, V., & Johnson, W. W. (2017). Why human-autonomy teaming? In *International conference on applied human factors and ergonomics* (pp. 3–11). Springer.
Stevens, R., Galloway, T., Halpin, D., & Willemsen-Dunlap, A. (2016). Healthcare teams neurodynamically reorganize when resolving uncertainty. *Entropy*, *18*(12), 427. https://doi.org/10.3390/e18120427
Tambe, M., Shen, W. M., Mataric, M., Pynadath, D. V., Goldberg, D., Modi, P. J., & Salemi, B. (1999). Teamwork in cyberspace: Using TEAMCORE to make agents team-ready. In Proceedings of the AAAI spring symposium on agents in cyberspace, Palo Alto, CA, March 1999 (pp. 136–141).
Thorén, H. (2014). Resilience as a unifying concept. *International Studies in the Philosophy of Science*, *28*(3), 303–324. https://doi.org/10.1080/02698595.2014.953343
Trotzky, S., Chen, Y. A., Flesch, A., McCulloch, I. P., Schollwöck, U., Eisert, J., & Bloch, I. (2012). Probing the relaxation towards equilibrium in an isolated strongly correlated one-dimensional Bose gas. *Nature Physics*, *8*(4), 325–330. https://doi.org/10.1038/nphys2232
Wiltshire, T. J., Butner, J. E., & Fiore, S. M. (2018). Problem-solving phase transitions during team collaboration. *Cognitive Science*, *42*(1), 129–167. https://doi.org/10.1111/cogs.12482

Woods, D. D. (2015). Four concepts for resilience and the implications for the future of resilience engineering. *Reliability Engineering and System Safety*, *141*, 5–9. https://doi.org/10.1016/j.ress.2015.03.018

Woods, D. D., Roth, E. M., & Pople, H., Jr. (1988). Modeling human intention formation for human reliability assessment. *Reliability Engineering and System Safety*, *22*(1–4), 169–200. https://doi.org/10.1016/0951-8320(88)90073-7

Yin, X., Clark, J., Johnson, C., Grimm, D., Zhou, S., Wong, M., Cauffman, S., Demir, M., Cooke, N. J., & Gorman, J. C. (2022). Development of a distributed teaming scenario for future space operations. *Proceedings of the Human Factors and Ergonomics Society 66th Annual Meeting*, *66*(1), 788–792. https://doi.org/10.1177/1071181322661405

David A.P. Grimm is a Ph.D. student in Engineering Psychology at the Georgia Institute of Technology. He received his M.S. in Psychology from the Georgia Institute of Technology in 2020.

Jamie C. Gorman is a professor of Human Systems Engineering at Arizona State University, Deputy Director of ASU's Center for Human, Artificial Intelligence, and Robot Teaming, and Senior Personnel in the NSF Institute for Student-AI Teaming at the University of Colorado—Boulder. He received his Ph.D. in Psychology from New Mexico State University in 2006.

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.

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 Ph.D. 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.