ARL-TR-9647 ● Feb 2023

DEVCOM
ARMY RESEARCH
LABORATORY

Potential Indicators of Team Trust Measured in a Next Generation Combat Vehicle Simulator

by Harrison Philip Crowell, Daniel Forster, Steven Thurman, Kimberly Pollard, David Chhan, Angelique Scharine, Andrea Krausman, Shan Lakhmani, and Justin R Brooks

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer’s or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-9647 ● FEB 2023

Potential Indicators of Team Trust Measured in a Next Generation Combat Vehicle Simulator

Harrison Philip Crowell, Daniel Forster, Steven Thurman, Kimberly Pollard, David Chhan, Angelique Scharine, Andrea Krausman, and Shan Lakhmani
DEVCOM Army Research Laboratory

Justin R Brooks
D Prime LLC

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)* February 2023	**2. REPORT TYPE** Technical Report	**3. DATES COVERED (From - To)** 30 September 2021–29 September 2022

4. TITLE AND SUBTITLE Potential Indicators of Team Trust Measured in a Next Generation Combat Vehicle Simulator	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Harrison Philip Crowell, Daniel Forster, Steven Thurman, Kimberly Pollard, David Chhan, Angelique Scharine, Andrea Krausman, Shan Lakhmani, and Justin R Brooks	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) DEVCOM Army Research Laboratory ATTN: FCDD-RLA-FB Aberdeen Proving Ground, MD 21005	**8. PERFORMING ORGANIZATION REPORT NUMBER** ARL-TR-9647
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES

ORCID IDs: Harrison Crowell, 0000-0001-8763-8733; David Chhan, 0000-0003-2470-8663; Daniel Forster, 0000-0001-8351-2009; Andrea Krausman, 0000-0003-1955-8867; Shan Lakhmani, 0000-0001-6052-439X; Kimberly Pollard, 0000-0002-5849-1987; Angelique Scharine, 0000-0001-8230-1824; Steven Thurman, 0000-0003-3962-8447

14. ABSTRACT

This report describes the analysis of subjective, physiological, and behavioral data as part of a project to assess team trust and identify potential indicators of team trust. The long-term goal of the project is to enhance human–autonomy teaming. The data came from a study of technology aids and crew size. In that study, Soldiers conducted various missions in a Next Generation Combat Vehicle simulator. For this report, approximately 300 variables were examined. The variables that correlated significantly with trust in human–human teams included heart rate, heart rate variability, respiration rate, pupil size, speech affect, verbosity, affiliative language, and linguistic style matching. For human–autonomy teams, the variables that correlated significantly with trust included heart rate, heart rate variability, respiration rate, frontal and posterior alpha modulated electroencephalogram signals, posterior theta band electroencephalogram signals, pupil size, eye fixation position, and speech descriptiveness. These results can guide the development of systems to monitor team trust and intervene if it becomes too low or too high.

15. SUBJECT TERMS

human–autonomy team, simulation, trust measurement, state assessment, team trust, physiological indicator, behavioral indicator, Humans in Complex Systems

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Harrison Philip Crowell
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	UU	68	**19b. TELEPHONE NUMBER (Include area code)** (410) 278-5986

Contents

List of Figures

List of Tables

Executive Summary

In response to the US Army's Next Generation Combat Vehicle (NGCV) Modernization Priority, the US Army Combat Capabilities Development Command Army Research Laboratory established the Human–Autonomy Teaming Essential Research Program (HAT ERP). The overarching goal of the HAT ERP is to address the integration and design challenges associated with "teaming" humans and autonomous systems within the future operational environment. The rationale behind human–autonomy teams is to create synergistic teams that function effectively and adapt to the complex and dynamic nature of combat. While integrating humans and autonomous systems has the potential to provide further team capabilities and a combat advantage, there are many challenges to be addressed. One specific challenge being addressed under Project 5 of the HAT ERP is to better understand how trust emerges within human–autonomy teams, which requires effective methods for measuring it, and to develop interventions to help humans and autonomous team members calibrate their trust in each other. Although trust research is decades old, the primary source of measurement has been self-report questionnaires that are effective for understanding trust at specific time points but are unable to track changes in trust in a more continuous fashion, thereby providing a more comprehensive picture of trust and the team dynamic. Specific goals of HAT Project 5 include identifying novel unobtrusive and real-time, or near real-time, methods for measuring the dynamic nature of trust over time and informing appropriate interventions to restore individual and team trust, if necessary.

Using a comprehensive process of literature reviews and data gathered during laboratory and field research, the Project 5 team identified several potential indicators of trust and created a multimodal conceptual toolkit, which consists of subjective measures (i.e., self-reporting), behavioral indicators such as communication content and flow, and physiological indicators like heart rate and pupil size. The benefit of these indicators is that they provide a rather unobtrusive, real-time, continuous measurement capability so researchers are able to capture changes in trust over the course of a mission and arrive at a more robust understanding of when and why trust changes.[1] Here, we sought to understand the relationships between these indicators and to arrive at a better understanding of what they can tell us about team trust, team dynamics, and how these measures can support the development of trust metrics. The specific focus of this report is to

[1] Krausman A, Neubauer C, Forster D, Lakhmani S, Baker AL, Fitzhugh SM, Gremillion G, Wright JL, Metcalfe JS, Schaefer KE. Trust measurement in human-autonomy teams: development of a conceptual toolkit. Trans Hum Robot Interact. 2022;11(3). https://doi.org/10.1145/3530874.

document initial validation work using data collected during a human–autonomy simulation experiment.

In this report, we present our analyses of data collected during the technology aids and crew size study conducted by Cox et al.[2] In that study, Soldiers representing the members of an NGCV section performed a variety of typical armor missions in a simulator. The goal of our exploratory analysis was to identify physiological and behavioral measures related to team trust. The long-term goal for this work is to enhance human–autonomy teaming. In our analyses, we considered two kinds of teams (i.e., human–autonomy teams and human–human teams). We examined approximately 300 different physiological and behavioral variables for their relationship to team trust. We assessed trust with the responses from three different questionnaires given after each data collection session. We collected physiological data with instruments that collected electrical signals from the heart and brain. We also tracked eye movements. The behavioral data we used came from recordings of communications between section members and records of events and vehicle locations during the missions that were captured as part of the simulation.

In our analyses, we found the following variables to be correlated to human–human trust:

- Heart rate
- Heart rate variability
- Respiration rate
- Pupil size for pairs of subjects controlling the same vehicle
- Speech communication content
    - Affect
    - Verbosity
    - Affiliative language
- Linguistic style matching

We found the following variables to be correlated to human–autonomy trust:

- Heart rate

[2] Cox KR, Rexwinkle JT, Gremillion MG, Perelman BS, Brooks JR, Dyer P, Pollard KA, Forster DE, Chhan D, Carter EC, et al. Effects of technology aids and crew size on a Next Generation Combat Vehicle platoon section. DEVCOM Army Research Laboratory (US); 2022. Report No.: ARL-TR-9403.

- Heart rate variability
- Respiration rate
- Frontal and posterior alpha modulated electroencephalogram (EEG) signals
- Posterior theta band EEG signals
- Standard deviation of pupil size
- Eye fixation position
- Speech communication content
    - Descriptiveness

Following up on our analyses will involve analyzing data from similar studies, conducting studies specifically focused on assessing trust, and examining additional physiological and behavioral measures. Two recently completed studies will provide larger data sets that we can analyze to further our understanding of relationships among the different measures. In future studies we will manipulate levels of trust to verify that the most relevant physiological and behavioral variables are being measured. All of this will be important for the development of systems to monitor the level of team trust and implement interventions if trust levels become too high or too low.

1. Introduction

Trust is understood to be critical for team functioning, as trust facilitates various team processes such as information sharing, collaborative decision making, positive team interactions, and overall team effectiveness (Salas et al. 2005; Grossman and Feitosa 2018). While research investigating trust in human teams spans decades, many questions remain with respect to trust in human–autonomy teams. Human–autonomy teams are composed of one or more human teammates working interdependently with one or more autonomous systems or intelligent agents, to accomplish a mutual goal (Demir et al. 2019). In both human and human–autonomy teams, inappropriate levels of trust (i.e., too high or too low) can have serious implications for team functioning and outcomes. What is needed is proper calibration of team trust, which forms as team members work together over time (Fulmer and Gelfand 2012; Costa and Anderson 2017). However, it is important to note that trust is dynamic and fragile, so even though individuals or teams may have developed a level of trust, depending on the circumstances (e.g., stress, vulnerability, and risk), it can change, thereby impacting team performance. Therefore, being able to measure the dynamic and emergent nature of trust, as well as trust-based decisions and actions, will enable researchers to understand the factors that influence changes in trust, and when necessary, prescribe interventions that enable effective restoration of trust.

Trust assessments in human teams are typically accomplished through traditional methods using subjective measures such as questionnaires. However, given the dynamic nature of military and civilian operations involving intelligent agents and autonomy, it is necessary to expand trust measurement to incorporate methods that can capture trust flow and changes over an entire task or mission. This suggests a need for a multimodal measurement capability that can be linked to specific scenario events. For example, based on an extensive survey of the literature and data from laboratory and field studies, Krausman et al. (2022) identified several methods, in addition to subjective measures, that may be well-suited for inferring trust in teams composed of multiple humans and multiple autonomous systems. These measures include communication analyses that focus on communication flow and content, behavioral measures such as gestures and facial expressions, and physiological measures like heart rate and eye tracking. The advantage of using these novel measures is that they provide a rather unobtrusive, real-time measurement capability that allows researchers to capture data continuously, if desired, throughout the course of a mission. Coupled with subjective ratings, these data can provide a more comprehensive picture of team dynamics.

Although more research needs to be done to understand many aspects of human–autonomy teams, human–autonomy teams have shown improved performance on tasks compared with performance on the same tasks when autonomy was not part of the team (Cox et al. 2022). Improved performance and a reduction in the number of personnel needed make military operations one area that could benefit from the use of human–autonomy teams. To advance the research needed to integrate humans and autonomous systems within future combat operations, the US Army Combat Capabilities Development Command (DEVCOM) Army Research Laboratory (ARL) established the Human–Autonomy Teaming Essential Research Program (HAT ERP). The goal of the HAT ERP is to address the challenges associated with teaming humans and autonomous systems within a complex tactical environment to create synergistic teams that function effectively and adapt to the dynamic nature of combat. One specific area being addressed in Project 5 of the HAT ERP is how to effectively measure critical team processes such as trust and cohesion. The overall goal of HAT Project 5 is to develop novel, multimodal metrics of team trust and cohesion to effectively calibrate trust and improve human–autonomy team performance supporting the Next Generation Combat Vehicle (NGCV). Specific goals of HAT Project 5 include identifying unobtrusive and real-time, or near real-time, measures of trust that capture the dynamic nature of team trust and informing appropriate trust interventions to restore team trust, if necessary.

A recent study conducted at ARL examined how technology aids and crew size influenced mission performance in a notional armored vehicle (Cox et al. 2022). The study was conducted in a simulator, and subjects (Soldiers who had experience in armored vehicles) performed the roles of crew members in the notional armored vehicle. The subjects (i.e., friendly forces, also known as BLUFOR) performed 17 scenarios against an opposing force (OPFOR). For certain scenarios, subjects had access to three novel technology aids (i.e., automation systems) to assist them with their tasks. Crew size was also examined during the study. During the first week of the study, six subjects participated. Then a seventh subject joined them for the second week. The results of the study showed that use of the technology aids improved performance on mission tasks and reduced workload on the subjects. Specifically, the use of technology aids decreased the decision-making time for subjects to hit the OPFOR after initial detection of a shot. In addition, the technology aids reduced the physiological workload on both the six- and seven-person crews, and the reduction in physiological workload was more evident in the six-person crew. Crew size was also a factor in the cognitive workload reported by subjects. Cognitive workload was lower in the seven-person crew than the six-person crew.

This report focuses on analyses of a subset of the data collected during the study of technology aids and crew size. During that study, researchers collected many different types of data. These include subjective measures (e.g., questionnaires about trust, workload, and situational awareness) along with physiological measures, behavioral measures, and overall performance measures. The purpose of our analyses is to investigate relationships between team trust and various physiological and behavioral variables. By investigating these relationships, we expect to find variables that are indicators of team trust for members of a section in an NGCV. Having indicators of team trust that can be measured objectively and unobtrusively will allow us to identify when trust levels change without having to use questionnaires. This information can then be used to identify when interventions may need to occur to restore the proper balance of trust to the team.

2. Background

While trust can be measured using subjective experience, objective observations of human or agent behavior, as well as other data sources such as physiological and behavioral data, may provide additional insights regarding trust. Earlier work by Krausman et al. (2022) established a conceptual toolkit of potential measures or indicators of trust based on a comprehensive literature review as well as data from both laboratory and field studies examining trust in human–autonomy teams. There are several studies that examine these potential indicators of trust. A brief description of some of these studies follows.

Physiological measures such as heart rate, electroencephalogram (EEG) signals, electrodermal activity, and vocalics are influenced by trust. For example, Mitkidis et al. (2015) found that heart rate and shared heart rhythms are associated with building trust, and Zhou et al. (2019) found that heart rate correlates with user trust in predictive decision making using an intelligent decision aid. In a study of autonomous vehicle malfunctions, Seet et al. (2022) found that changes in EEG correlated with changes in trust. In a study by Schaefer et al. (2012), subjects who gave low trust ratings for their interactions with unreliable robots had high electrodermal activity. In examining qualities of speech that are distinct from verbal and linguistic content (i.e., vocalics), Elkins and Derrick (2013) found that voice pitch is related to trust.

Researchers have also investigated behavioral measures such as communication content, eye tracking, and facial expressions for their relationship to trust. Levels of trust may be reflected in communication patterns, as well as the content and quantity of verbal communication that human participants use during an interaction (Gonzales et al. 2010; Khawaji et al. 2013; Ghafari et al. 2020). In addition, a speaker's affect (or mood) can influence spoken communication patterns (Pennebaker et al. 2003). With respect to eye gaze, where a person looks in certain situations can be an indication of their level of trust. Team members who trust each other spend less time monitoring each other's behavior (Langfred 2004; Chen and Barnes 2012). Similarly, human operators monitoring a system may gaze less frequently on it if their trust in that system is high (Merritt et al. 2013; Hergeth et al. 2016). Emotions are often revealed in facial expressions. In a study of trust and facial expressions, Campellone and Kring (2013) found that facial expressions are a factor in an individual's decision to trust another person.

In our analyses of physiological and behavioral measures related to trust for this report, we use data collected during the technology and crew size study. Thus, the following subsections will provide a brief description of the methods used in that study. A full description is available in the technical report by Cox et al. (2022).

3. Methods Used in the Technology Aids and Crew Size Study

3.1 Participants

Seven US Army Soldiers participated in the technology aids and crew size study. Six of the Soldiers were from the Minnesota Army National Guard, and one was from the US Army Evaluation Center. They ranged in age from 29 to 57 years old (mean age: 38 ±10 years). They had 12 to 39 years of experience (mean experience: 20 ±10 years) in the Army, and their Military Occupational Specialties were 11B (Mechanized Infantry), 19D (Cavalry Scout), and 19K (Armor Crewman).

3.2 Facility and Equipment in the Simulators

The technology aids and crew size study was conducted in the Information for Mixed Squads (INFORMS) Laboratory at Aberdeen Proving Ground, Maryland. The INFORMS Laboratory contains two stationary vehicle simulators. Each simulator contains workstations for the members of the vehicle's crew. In the study, one simulator was used to represent one section of an armored vehicle platoon. The section consists of a manned control vehicle (MCV) and two robotic combat vehicles (RCVs). The workstations in the simulator were configured for a six- or seven-person crew. There was one workstation for the driver of the MCV and one workstation for the gunner of the MCV. There were also workstations for the driver and gunners of each RCV (i.e., a workstation for the driver of RCV1, a workstation for the gunner of RCV1, a workstation for the driver of RCV2, and a workstation for the gunner of RCV2). In the conditions when there was a six-person crew, the MCV gunner served as the section commander as well. For the conditions when there was a seven-person crew, the seventh person was the section commander, and they also had a workstation. The workstations included the following equipment for interfacing with the simulation: personal computer (PC; Rave Midtower, Rave Computer, Sterling Heights, MI); three touchscreen monitors (Elo 1502L, Elo Touch Solutions, Inc., Milpitas, CA) for visualizing the simulation environment and the controls for drivers and gunners; steering yokes (Fanatec ClubSport Steering Wheel and CSL Elite Wheel Base, Endor, AG, Landshut, Germany) for driving the vehicle or controlling the weapon system, as appropriate for the subject's role in the study; and pedals (Fanatec CSL Elite Pedals, Endor, AG, Landshut, Germany) for the driver to accelerate and brake the vehicle.

3.3 Independent Variables

During the study, two independent variables were manipulated: 1) technology aids available and 2) crew size. There were several different technology aids available and two crew sizes (six- or seven-person). The workstations in the simulators were set up with a baseline level of technology aids. These included an aided target recognition system and a common operating picture (COP) for all members of the section to view on their workstations. In addition, three novel technology aids were examined. They were the Commander's Interface, the Playbook, and Verbal Commands. The Commander's Interface is an automated system for providing the commander with a display of fuel and ammunition levels, crew casualties, and equipment status. It also gives the commander access to the display from any sensor that any of the section members has open at the time. The Commander's Interface automatically provides this information so that the section members are not burdened with providing it. The Playbook is a group of formations and movements that the commander can send to the crew member's display. This eliminates the need to verbally relay these commands to the section members. The Verbal Commands technology aid is something that allows each section member to accomplish certain actions via voice input rather than a touchscreen, keyboard, mouse, or other handheld device. With this technology aid, subjects could manipulate the map on the COP and add, modify, and categorize battlespace objects.

3.4 Dependent Measures and the Equipment used to Capture Them

All the measures described in the following were captured during the technology aids and crew size study. Most of the data required some type of preprocessing before analysis. The preprocessing procedures are described later in the report in the sections that document each of the analyses.

3.4.1 Subjective Measures

The following three questionnaires were among the eight questionnaires given to subjects after each of the missions. They were presented to subjects on a tablet PC to facilitate the collection of each subject's responses.

System Trustworthiness. Subjects rated six items on the System Trustworthiness questionnaire (Muir and Moray 1996). Five of the questions focused on the autonomous systems the subjects used. They were asked about the competency, predictability, dependability, responsibility, and reliability of those autonomous systems. The sixth question was about the subject's degree of faith that the system would cope with future events. Participants responded on a 100-point scale (0 = "Very Low"; 100 = "Very High").

Trust in Automation. The Trust in Automation questionnaire (Jian et al. 2000) contains 12 items for subjects to rate their feelings of trust and their impressions of the automation systems they are using. Participants responded on a 7-point ordinal scale (1 = "Not at all"; 7 = "Extremely"). Example items include, "the system is deceptive," "I am suspicious of the system's intent, action, and outputs," "the system has integrity," and "I can trust the system."

Team Trust. The Team Trust questionnaire (Costa 2003) is used for assessing human teams. It has four subscales. In the technology aids and crew size study, only two of the subscales were used: perceived trustworthiness and cooperative behaviors. Participants responded on a 7-point ordinal scale (1 = "Completely Disagree"; 7 = "Completely Agree").

3.4.2 Physiological Measures

Electrocardiogram (ECG) and Respirations. The ECG and respirations were collected with the Zephyr BiohHarness 3 (Zephyr Technology Corporation, Annapolis, MD). The BioHarness 3 contains three types of sensors in a lightweight (50 g) strap. The strap is worn around the upper torso and in direct contact with the subject's skin. Conductive pads in the strap passively capture electrical signals from the subject's beating heart. A pressure sensor in the strap captures expansion and contraction of the strap as the subject inhales and exhales. Although not used in the analyses for this report, a three-axis accelerometer measures torso accelerations.

Electroencephalogram (EEG). Electrical activity of the subject's brain was measured with an X24 EEG System (Advanced Brain Monitoring, Inc., Carlsbad, CA). This system has 20 sensors that lay on the subject's scalp and are held in place by an elastic headband. The EEG sensors used were P3, P4, O1, O2, FP1, FP2, F3, F4, F7, and F8. The sensors passively capture electrical signals from the subject's brain.

Pupillometry and Eye Tracking. Pupil size was collected using Tobii Pro Glasses 2 (Tobii AB, Stockholm, Sweden). These glasses contain a head-mounted IR-based eye-tracking device with cameras built into the frames. The cameras look at the subject's eyes to measure pupil size and eye movements.

3.4.3 Behavioral Measures

Communication Content and Linguistic Style Matching (LSM). Participants communicated with each other through one of three communication networks: 1) a Headquarters and Headquarters Company (HHC)-Commander Network, 2) Section Network, and 3) Vehicle Network. The HHC-Commander Network allowed communication between the headquarters and the section commander. The Section Network allowed communication between all members of the section. The Vehicle Network allowed communication between the two operators (i.e., gunner and driver) of each vehicle. To communicate on these networks, participants were equipped with JBL Quantum Headsets (Nuance Communications, Inc., Burlington, MA), which have built-in microphone booms. Speech recordings were captured and automatically transcribed in real time using Dragon Naturally Speaking (Harman International Industries, Inc., Stamford, CT). After the study ended, speech recordings were then transcribed using Google Speech-to-Text (Google, Inc., Mountain View, CA).

3.5 Scenarios and Missions within the Scenarios

The technology aids and crew size study took place over a two-week period. During the first week, six subjects participated, and they completed eight scenarios. Each scenario lasted approximately 1 h and consisted of approximately three missions. The missions were typical of the missions that Soldiers in armored units train for, such as movement to contact, hasty defense, attack, and react to indirect fire. In all of the scenarios, subjects had access to the baseline technologies: aided target recognition and the COP. In the first three scenarios, a different one of the novel technology aids was used in each mission. In two of the remaining scenarios, subjects had access to the baseline technologies plus all of the novel technologies (i.e., the All On condition). For the other three sessions, subjects only had the baseline technologies (i.e., the All Off condition). During the second week, the same six subjects from the first week, plus a Soldier playing the role of section commander, participated. They completed nine sessions. Five of the sessions were in the All On condition, and the other four sessions were in the All Off condition.

4. Overall Objective and Hypothesis for the Analyses of Team Trust and Potential Indicators of Team Trust

This exploratory analysis examines the relationship between physiological measures and self-reported trust responses from individuals participating in the technology aids and crew size study (Cox et al. 2022). We also examine the relationship between behavioral measures collected during the technology aids and crew size study and the same self-reported trust responses. The long-term goal is to enhance human–autonomy teaming. By monitoring physiological and behavioral variables related to trust, it may be possible to calibrate trust between humans and autonomy and then develop interventions if trust levels become too low or too high. Past studies that examined the relationship between trust and psychophysiological measures found evidence that changes in stress and emotional experience affect an individual's propensity to trust (Dunn and Schweitzer 2005; Potts et al. 2019). Thus, we hypothesize that self-reported measures of trust will vary across the missions in the technology aids and crew size study and that changes in physiological and behavioral measures will correspond to changes in self-reported measures of trust. The succeeding sections will cover each analysis completed for this report. If there was a specific hypothesis for any of the analyses, it will be described in that section.

5. Methods, Results, and Discussion for Analyses of Data Related to Team Trust and Potential Indicators of Team Trust

In this report, we consider two kinds of teams, and we examine approximately 300 different variables for their relationship to team trust. One kind of team is the humans and the technology aids that were available to them (i.e., human–autonomy teams). The other kind of team is the typical human-to-human team. The measures of team trust can be categorized broadly as subjective, physiological, and behavioral measures. For the analyses, the physiological and behavioral measures are each divided into subcategories. The subcategories for physiological measures include the following:

- Cardiorespiratory Measures
- Generalized Linear Models of Multiple Physiological Measures
- Pupil Size as a Measure of Physiological Coherence and Team Trust

The subcategories for behavioral measures include the following:

- Communication Content and Linguistic Style Matching (LSM)
- Acoustically Derived Measure of Soldier Engagement
- Decision Times

5.1 Subjective Trust Measures

5.1.1 Methods

The subjective trust measures come from the System Trustworthiness, Trust in Automation, and Team Trust questionnaires. After two of the missions, several subjects answered the System Trustworthiness and Trust in Automation questionnaire twice. It is not clear why this occurred so we chose to use the first responses from these subjects because 1) the first responses seemed consistent with the responses from these subjects after the other missions; 2) in some of the second responses, it looked like some subjects just selected the same number for each question; and 3) for questions that require subjects to recall what they did or how they felt, it is generally best to ask questions as soon as possible after the subject finishes an activity. Because each of these questionnaires was administered on multiple occasions, yet each occasion only had six to seven respondents, we computed composite scores for each participant by averaging across their individual responses to these questionnaires. To assess whether we could justify averaging across responses, we tested whether item responses from the same questionnaire were correlated with each other. If responses to one or a few items were uncorrelated or negatively correlated with the rest, then those items were dropped. If the questionnaire was designed to measure a unitary construct but the correlation matrix indicated a different factor structure, we followed the data and computed composite scores for the apparent factors revealed by the matrices. Because these data came from the same participants responding over multiple occasions, we constructed correlation matrices using the "rmcorr" package (Bakdash and Marusich 2017) in R (R Core Team 2020).

5.1.2 Results

System Trustworthiness. Responses to this questionnaire were significantly correlated with each other (mean $r = 0.55$), so we averaged them to create a System Trustworthiness composite score.

Trust in Automation. Responses to this questionnaire appeared to separate along two dimensions, with negatively worded items correlating strongly with other negatively worded items (mean $r = 0.32$), but not at all with positively worded items (mean $r = –0.02$), and positively worded items were strongly correlated with each other (mean *r* = 0.52). Therefore, we computed two composite scores for each person—one representing “distrust in automation” (items 1–5) and the other representing “trust in automation” (items 6–12).

Team Trust. For the perceived trustworthiness subscale, all items correlated significantly with each other (mean *r* = 0.60). For the cooperation subscale, three items correlated significantly with each other (mean *r* = 0.436), whereas the item, “some people hold back relevant information in this team,” was uncorrelated with the others (mean *r* = 0.13), and therefore dropped from the composite score for cooperation.

5.1.3 Discussion

In support of the notion that trust is distinct from distrust (e.g., Lewicki et al. 1998), the Trust in Automation questionnaire separated into two factors—one representing trust and another representing distrust—and these different factors were unrelated to each other. These analyses further helped us identify one item to remove from a subscale, as it was unrelated to the other indicators and would have only served to increase noise in measurement when averaging with other item responses. Overall, these results provide some support for making composite variables from the different trust questionnaires.

5.2 Physiological Measures of Trust

5.2.1 Cardiorespiratory Measures

5.2.1.1 Methods

Some preprocessing of the data was done to calculate heart rate (HR), heart rate variability (HRV), and respiration rate (RR). These values were computed from the data collected by the BioHarness 3. The raw ECG time-series data were segmented into 30-s non-overlapping windows and processed using BioSPPy, the open-source Biosignal Processing package in Python (Carreiras et al. 2015) to extract R peaks. The R-to-R intervals represent the inter-beat intervals, which were used to compute HR. The HRV was computed from the root mean square successive differences in the HR data. The raw respiration waveforms were windowed into non-overlapping 30-s segments and bandpass filtered (0.1–0.35 Hz) before they were processed with BioSPPy to compute zero-crossing and obtain individual time-series RRs. To account for individual differences, each computed physiological measure (HR, HRV, and RR) was normalized by subtracting it from a resting state measure collected as participants sat quietly and watched a video of slow-moving objects for 5 min. (For more details, see Cox et al. 2022.) For each subject, in each mission, mean and standard deviations of the HR, HRV, and RR were computed. Then linear regression models were created to obtain measures of correlation between each physiological measure and individual self-reported trust responses as measured in the System Trustworthiness, Trust in Automation, Team Trustworthiness, and Team Cooperation questionnaires.

5.2.1.2 Results

The results revealed two opposite trends with weak but statistically significant relationships in that changes in physiological measures correspond to changes in trust levels. For the Team Trust questionnaire, for both the perceived trustworthiness and cooperative behaviors subscales, we observed lower HR, elevated HRV, and decreased RR as the levels of trust reported by the participants increased (Figs. 1 and 2 and Tables 1 and 2). These physiological changes are consistent with the idea that when participants trust their teammates, they appeared to be in less stressful situations with low cognitive demand and positive emotional experiences as reflected in the lowered HR and RR and increased HRV. However, the reverse trend is true for the trust in autonomy as measured by System Trustworthiness (Fig. 3 and Table 3) and Trust in Automation (Fig. 4 and Table 4) questionnaires. In those cases, HR and RR increased and HRV decreased as trust increased.

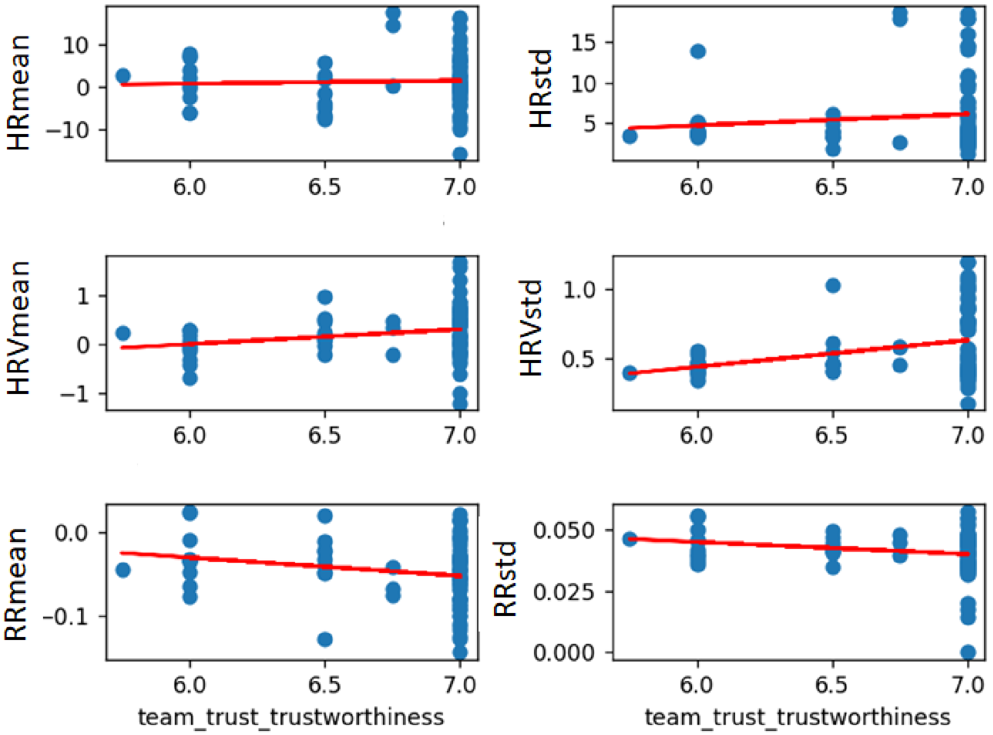
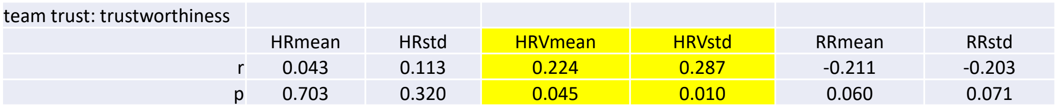


Fig. 1 HR, HRV, and RR means and standard deviations as functions of team trust: perceived trustworthiness. The perceived trustworthiness ratings go from 1 = "completely disagree" to 7 = "completely agree."

Table 1 The correlation (r) and p-value of HR, HRV, and RR means and standard deviations to team trust: perceived trustworthiness. Statistically significant results are highlighted in yellow.



team trust: trustworthiness						
	HRmean	HRstd	HRVmean	HRVstd	RRmean	RRstd
r	0.043	0.113	0.224	0.287	-0.211	-0.203
p	0.703	0.320	0.045	0.010	0.060	0.071

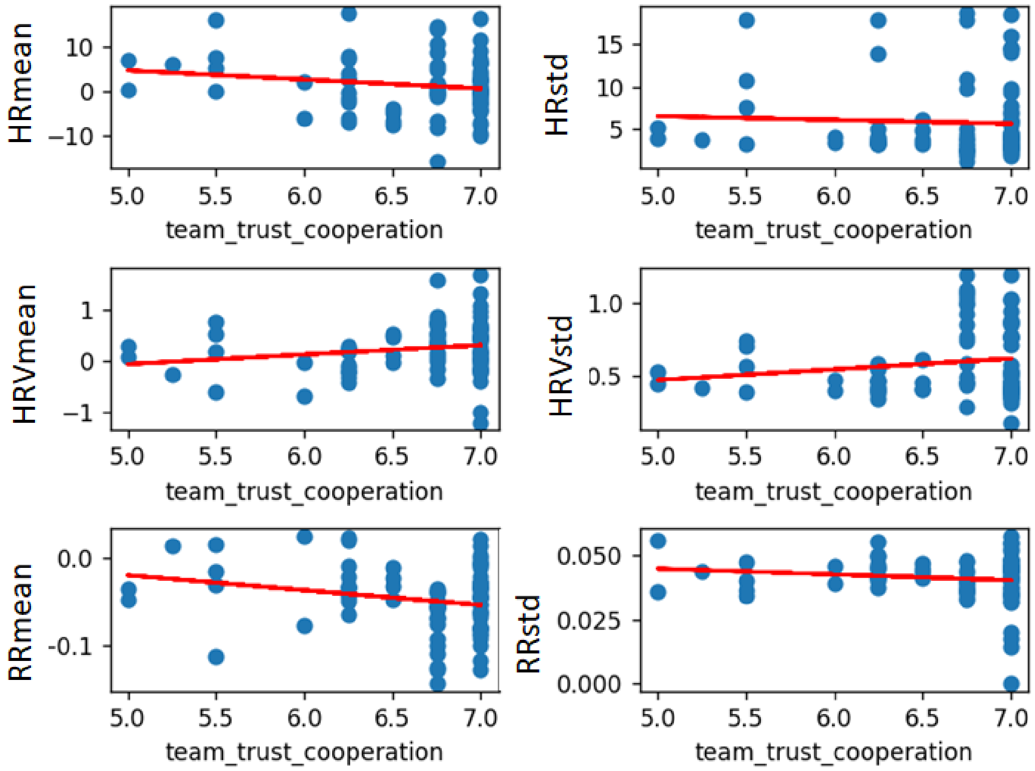
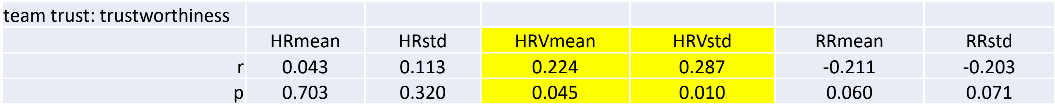


Fig. 2 HR, HRV, and RR means and standard deviations as functions of team trust: cooperative behaviors. The cooperative behaviors ratings go from 1 = "completely disagree" to 7 = "completely agree."

Table 2 The correlation (r) and p-value of HR, HRV, and RR means and standard deviations to team trust: cooperative behaviors. Statistically significant results are highlighted in yellow.



team trust: trustworthiness						
	HRmean	HRstd	HRVmean	HRVstd	RRmean	RRstd
r	0.043	0.113	0.224	0.287	-0.211	-0.203
p	0.703	0.320	0.045	0.010	0.060	0.071

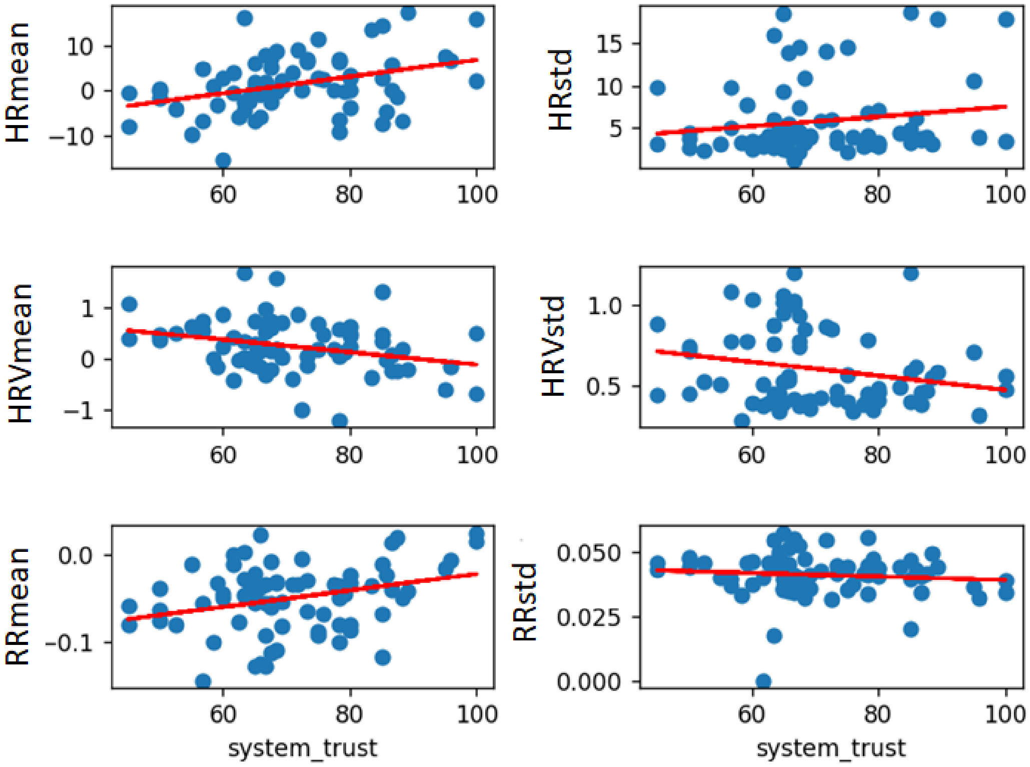
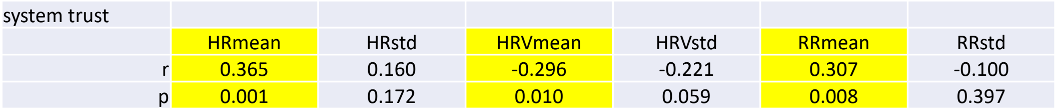


Fig. 3 HR, HRV, and RR means and standard deviations as functions of system trustworthiness. The system trustworthiness ratings go from 0 = "very low" to 100 = "very high."

Table 3 The correlation (r) and p-value of HR, HRV, and RR means and standard deviations to system trustworthiness. Statistically significant results are highlighted in yellow.



system trust						
	HRmean	HRstd	HRVmean	HRVstd	RRmean	RRstd
r	0.365	0.160	-0.296	-0.221	0.307	-0.100
p	0.001	0.172	0.010	0.059	0.008	0.397

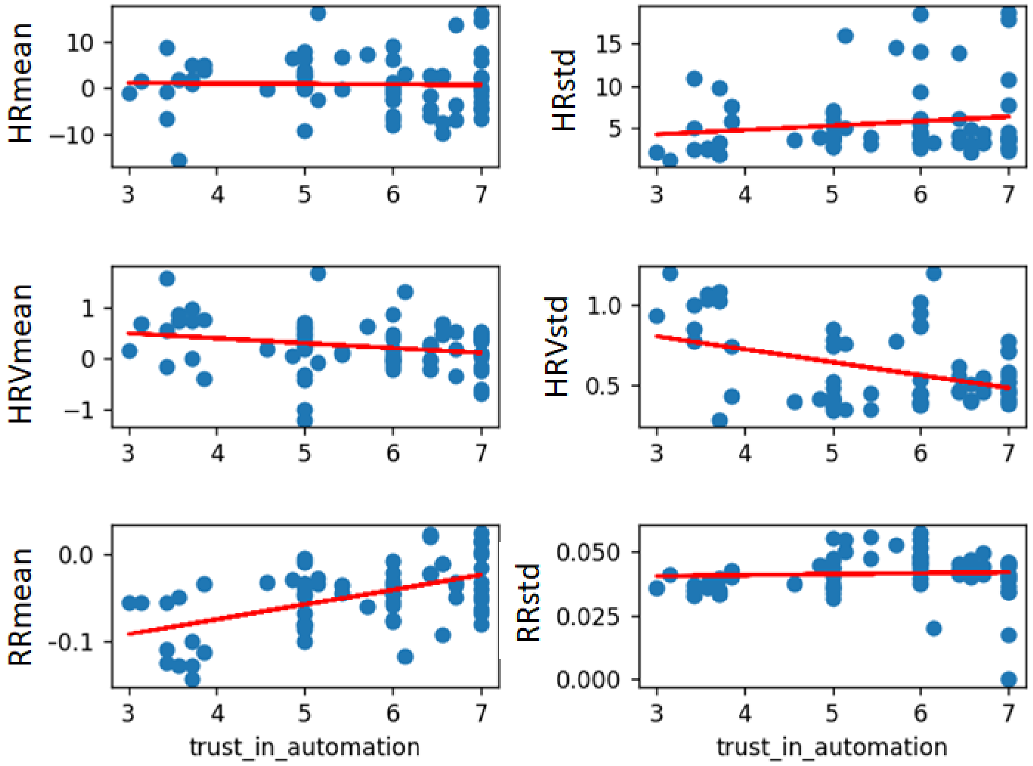
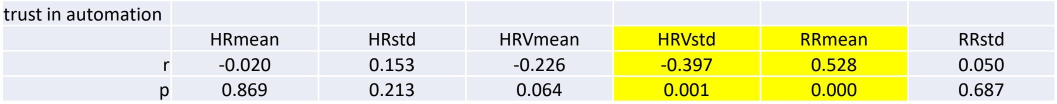


Fig. 4 HR, HRV, and RR means and standard deviations as functions of trust in automation. The trust in automation ratings go from 1 = "not at all" to 7 = "extremely."

Table 4 The correlation (r) and p-value of HR, HRV, and RR means and standard deviations to trust in automation. Statistically significant results are highlighted in yellow.



trust in automation						
	HRmean	HRstd	HRVmean	HRVstd	RRmean	RRstd
r	-0.020	0.153	-0.226	-0.397	0.528	0.050
p	0.869	0.213	0.064	0.001	0.000	0.687

5.2.1.3 Discussion

For team trust (i.e., human–human trust), both in the Perceived Trustworthiness and Cooperative Behaviors subscales of the Team Trust questionnaire, we observed lower HR, elevated HRV, and decreased RR as the levels of trust reported by the participants increased. These physiological changes are consistent with the idea that when participants trust their teammates, they appeared to be in less stressful situations with low cognitive demand and positive emotional experiences as reflected in the lowered HR and increased HRV. These results support the findings in a separate analysis on the same data set reported in Cox et al. (2022) that similar changes in HR and HRV were also observed when technology aids were introduced. The availability of more resources and helpful tools such as technology aids had a positive impact on the crew in reducing the level of workload, enhancing team trust, and coordination that resulted in increases in team performance.

For trust in technology as measured by the System Trustworthiness and Trust in Automation questionnaires, we found a reverse in the relationship between trust and physiology in that increases in trust correspond to increases in HR and RR and decreases in HRV. These results appeared to suggest trust in humans and trust in autonomy elicit different physiological responses. While it is tempting to draw conclusions on the trends observed here, this is only an exploratory analysis as more data with greater frequency of questionnaire results and additional analyses are needed to further validate the relationship between trust and physiology. Nonetheless, the ability to use physiological measures as indicators of team trust among other team behaviors could potentially be useful for the development of real-time team monitoring technology so that interventions could be provided at the appropriate time to enhance team performance.

5.2.2 Generalized Linear Models of Multiple Physiological Measures

5.2.2.1 Methods

5.2.2.1.1 Preprocessing of physiological data. Raw ECG data were cleaned using the Neurokit2 package for Python. ECG features were calculated over 30-s windows. The ECG features calculated were HR: min, max, mean, and variability. Raw EEG data were cleaned by low-pass filtering at 30 Hz and then mean-centering across the frontal and parietal channels of interest. EEG features calculated were power of alpha-band frequencies over parietal channels, power of theta-band frequencies over frontal channels, Pearson correlation coefficients between all EEG channel pairs, and functional connectivity features using the weighted phase lag index (WPLI) between alpha- and theta-band-passed channels. EEG features were calculated over 3-s windows. Eye-tracking data was preprocessed and parameterized using PyGazeAnalyzer v0.1.0 in 10-s windows. Eye-tracking features included pupillometry (mean and standard deviation of pupil size), fixation position, fixation duration, and saccades. All positional eye-tracking features were determined for each dimension independently (X and Y).

5.2.2.1.2 Overview of the algorithm. This exploratory analysis focused on determining which (if any) physiological parameters were associated with subjective questionnaire responses of trust in the prototype technologies. For this analysis, a three-step algorithm was created. First, pair-wise, nonparametric, Spearman rank correlations between a physiologic parameter and questionnaire response was conducted. This step served as a filter to identify which parameters might be related to questionnaire responses. Next, for some questionnaire responses, multiple physiological parameters were identified as being related to the same questionnaire response. Therefore, a principal component analysis (PCA) was performed to address correlation between the physiological parameters and for dimensionality reduction. Given this motivation, PCA was performed anytime more than one physiological parameter was identified as being related to a given questionnaire response. For physiological parameters that were singly related to a questionnaire response, no PCA was conducted. Next, a generalized linear model (GLM) was performed using the first three PCA components or the single physiological parameter as independent variables with the questionnaire response as the dependent variable. This final step was performed to infer how the identified variables were related to the questionnaire response.

5.2.2.1.3 Identification of relevant physiological parameters. From the 287 physiological parameters derived from the wearable sensors (4 from ECG, 16 from eye tracking, and 267 from EEG), we identified those that were closely related to the responses to questions in the three questionnaires. To do this, we performed a nonparametric Spearman rank correlation between each of the 287 physiological parameters independently and each question response. To control for multiple comparisons, we performed a Bonferroni correction for each questionnaire. For the System Trustworthiness, Trust in Autonomation, and Team Trust questionnaires, corrections were made for 1,723, 3,445, and 2,297 tests, respectively. Physiological parameters that remained statistically significant after this correction were used to model questionnaire responses in subsequent analyses.

5.2.2.1.4 PCA of physiological parameters. More than one physiological parameter was identified as being significantly related to participant responses in the Trust in Automation and System Trustworthiness questionnaires. PCA was performed on these physiological parameters to control for multicollinearity that arises from the selection procedure (i.e., iterative correlation testing). To mitigate this effect, PCA was performed to extract correlated variability between physiological parameters into a single component and negate any potential effects of having correlated independent variables.

5.2.2.1.5 Logistic regression of wariness ratings from the trust in automation questionnaire. The potential responses for the wariness measure in the Trust in Automation questionnaire ranged from 1 to 7 (i.e., “Not at all” to “Extremely”). However, the distribution of responses was bimodal (1 and 7) rather than uniform or normal (Fig. 5). Therefore, we first used k-means to binarize these data into two groups:

0 = “Strongly Disagree” and 1 = “Strongly Agree.” After binarizing the data, we used logistic regression to predict the probability of being associated with a particular grouping in lieu of predicting the actual response. In the logistic regression, the predictors were the first three PCA components that contained information from multiple physiological parameters or individual physiological parameters. This depended on the number of physiological parameters identified as being potentially related to questionnaire responses per the methods previously described.

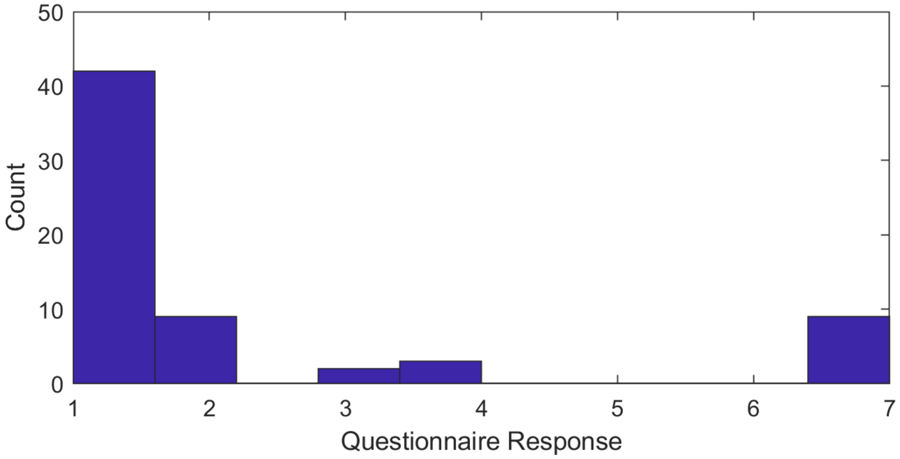


Fig. 5 Responses to the Trust in Automation questionnaire followed a bimodal distribution in contrast to a normal or uniform distribution. Therefore, logistic regression was used on these data after binarizing the responses using a k-means clustering with a group number = 2. Survey responses go from 1 = “not at all” to 7 = “extremely.”

5.2.2.1.6 GLM of remaining trust in automation and system trustworthiness data. Responses from the other three Trust in Automation ratings and the two from the System Trustworthiness questionnaire were consistent with a uniform distribution (Fig. 6) and therefore were more amenable to a traditional GLM approach. In this analysis, a GLM was developed that used either PCA components or an individual physiological parameter per the methods previously described.

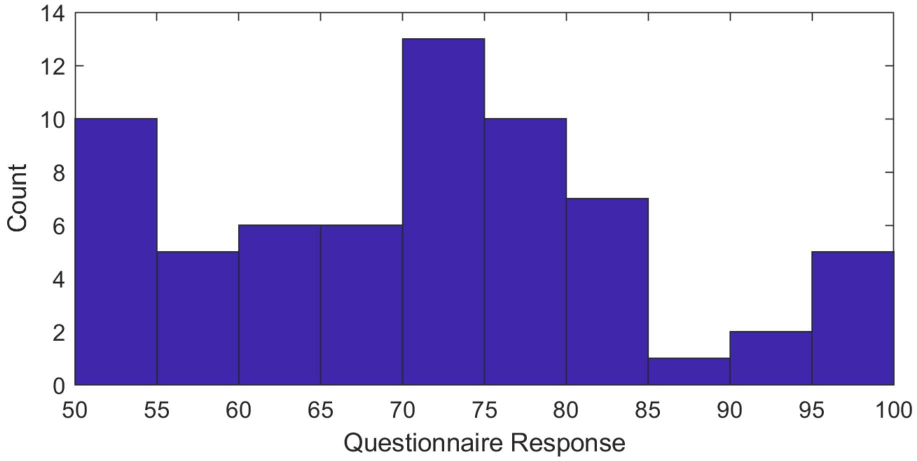


Fig. 6 Responses to the System Trustworthiness questionnaire followed a more continuous, uniform distribution, which permitted a GLM approach for analysis in contrast to the Trust in Automation data. Responses range from 0% = "not at all" to 100% = "extremely high."

5.2.2.2 Results

5.2.2.2.1 Correlation results. No significant relationships between physiological parameters and the Team Trust questionnaire were identified after correcting for multiple comparisons.

For the Trust in Automation questionnaire, 4 of 12 questions showed significant relationships between physiological parameters and questionnaire responses (Table 5). The four questions related to physiology were 1) wariness of the system, 2) dependability of the system, 3) trust in the system, and 4) familiarity with the system. The physiological parameters that correlated across the Trust in Automation questions included HR, posterior (P3 and O1) alpha power, time-locked correlation between frontal EEG channels (F4 and F8), and eye-tracking parameters including the standard deviation of pupil size and fixation position. The strengths of correlations ranged from –0.7 to 0.61 with a mean magnitude of 0.64. All p-values were less than 1e-5.

Table 5 Out of 12 responses, 4 showed a significant correlation with physiological parameters using the serial correlation approach. These parameters included multiple features from ECG, EEG, and eye tracking.

Trust Autonomy Survey	I am wary of the system	The system is dependable.	I can trust the system.	I am familiar with the system
Min Heart Rate				*0.6136*
P3 Alpha Power	*-0.691*			
O1 Alpha Power	*-0.6998*			
F4:F8 Correlation		*-0.67*		
Standard Deviation of Pupil Size		*0.6314*	*0.6037*	
Fixation Position (Y) Mean	*-0.5743*			

For the System Trustworthiness questionnaire, in two of six questions, there was a significant relationship between physiological parameters and questionnaire responses (Table 6). For system competency, seven physiological parameters were correlated, which included directed, functional connectivity in theta band for posterior channels (P3:P4 and P3:O2), and the clustering coefficient for multiple frontal and parietal channels in the alpha band. System reliability ratings were correlated with P3:O2 WPLI in the theta band. The correlations between physiological parameters and both competency and reliability were positive and ranged from 0.66 to 0.71 with an average of 0.69. All p-values were less than 2e-5.

Table 6 There were seven physiological parameters related to two out of six questions in the System Trustworthiness questionnaire. All of these parameters were derived from EEGs and included directed connectivity metrics and clustering coefficients from posterior and frontal electrodes, respectively.

System Trust	Competency	Reliability
P3_P4_WPLI_Theta	*0.6821*	
P3_O2_WPLI_Theta	*0.663*	*0.6724*
FP1_Clust_Alpha_Theta	*0.6931*	
F3_Clust_Alpha_Theta	*0.7121*	
F7_Clust_Alpha_Theta	*0.697*	
FP2_Clust_Alpha_Theta	*0.7117*	
F4_Clust_Alpha_Theta	*0.7061*	

5.2.2.2.2 PCA results. PCA was applied to two sets of physiological parameters that were related to Trust in Automation wariness and dependability. The loadings on the physiological parameters for each of the components are displayed in Fig. 7. As shown, in the left part of Fig. 7, there were complex relationships between the parameters and components. The first component for wariness demonstrated similar loadings across P3A, O1A, and the mean fixation position, which accounted for 73.3% of the variance in these independent variables. The second component, which accounted for 14.8% of the variance in these variables, showed a highly positive loading for P3A with negative loadings for O1A and fixation position. The third component had negative loadings for P3A and O1A and a strongly positive loading for fixation position. This component accounted for 11.9% of the variance in the independent variables. In the right of Fig. 7, the loadings for F4:F8 and the standard deviation for pupil size are shown. The first component demonstrates significant loading of F4:F8 with lower weighting of the standard deviation of pupil size, whereas the second component had the opposite pattern. The first and second components accounted for 99.26% and 0.73%, respectively, of the variance in F4:F8 and the standard deviation of the pupil size.

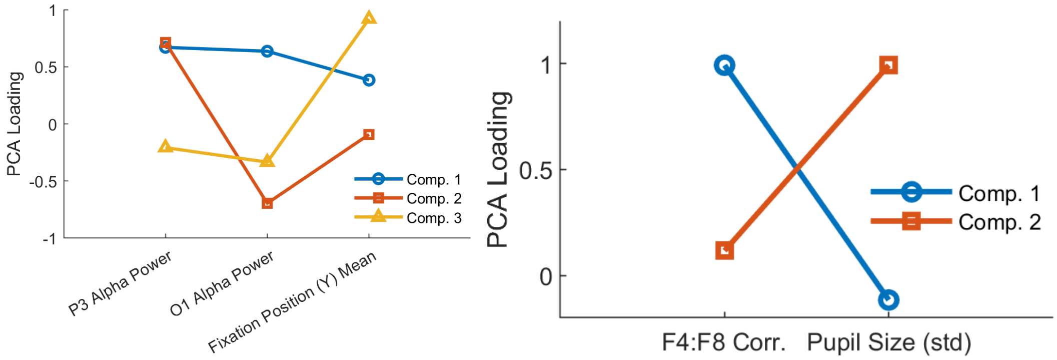


Fig. 7 The loadings of the first three components determined from the features that were correlated with system wariness in the Trust in Automation questionnaire (left). The loadings of the components determined from features that were correlated with system dependability in the Trust in Automation questionnaire (right).

For the System Trustworthiness questionnaire, only one PCA was applied to the physiological parameters, which were correlated with competency ratings (first column of Table 6). The loadings for each component per physiological parameter are provided in Fig. 8. The first component, which accounted for 85.6% of the variance in the physiological parameters, had modest loadings for the directed connectivity metrics; however, it showed moderate and consistent loadings for the clustering coefficient at multiple sites. Conversely, both the second and third components demonstrated almost no loadings with the clustering coefficients; however, they had significant and varied loadings for the directed connectivity metrics. These two components accounted for 12.3% and 1.8% of the variance in these physiological parameters.

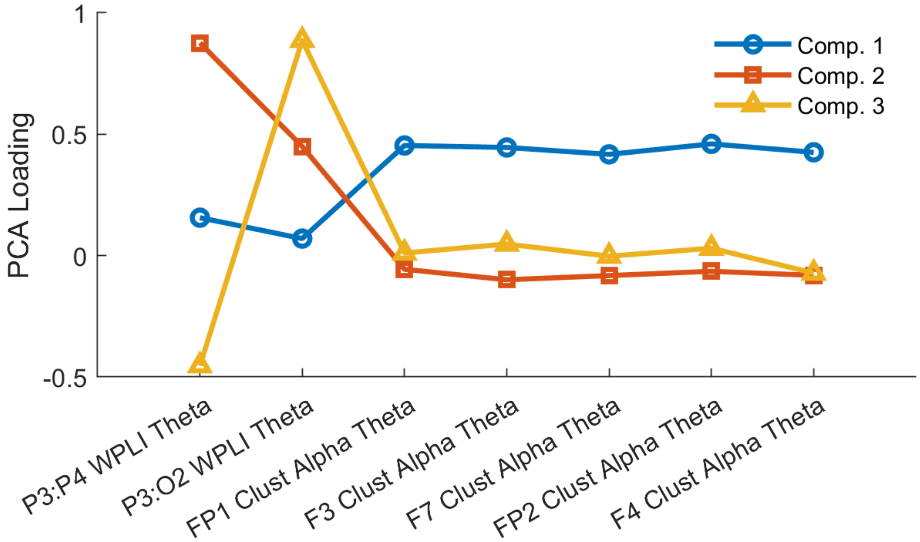


Fig. 8 Loadings for the physiological parameters identified as being related to the competency rating of the System Trustworthiness questionnaire

5.2.2.2.3 Logistic regression of trust in autonomy: wariness

Using the first three components of the physiological parameters related to wariness (see Fig. 7, left), the model was able to perfectly delineate high and low wariness. This was likely due to the unbalanced data set and potential overfit of the model. Therefore, we modeled each component separately and present those results in Fig. 9 and Table 7. Wariness of the technologies was well modeled by Components 1 and 3, whereas Component 2 did not relate to wariness.

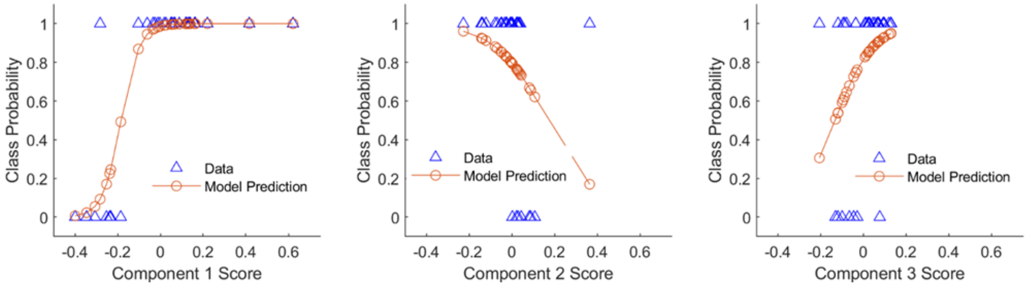


Fig. 9 Logistic regression model for Components 1, 2, and 3 in the left, middle, and right, respectively. An ensemble model using all three components perfectly separated classes; therefore, to better understand the prediction, we fit the data to each component separately.

Table 7 Logistic regression modeling results for wariness. Component 1 showed the strongest association between physiological parameters and system wariness. Component 2 was not statistically significant, whereas Component 3 demonstrated a moderate, positive, and statistically significant effect.

	est.	χ^2	p-value
Component 1	23.28	25.2	5e-7
Component 2	–8.107	2.98	0.0843
Component 3	11.19	4.56	0.0327

5.2.2.2.4 GLM of dependability, trust, and familiarity from the trust in automation questionnaire; competency and reliability from the system trustworthiness questionnaire. Dependability, as measured by the Trust in Automation questionnaire, was related to Components 1 and 2 (depicted previously in Fig. 7, right side). The model demonstrated a significant fit (F = 13.7, p = 6.3e-5). The first component had a coefficient estimate of –3.23 and associated p = 1e-4, while the second component was positively related with an estimate of 25.99 and associated p = 7e-3. In Fig. 10, a surface of the fitted model is shown against empirical data.

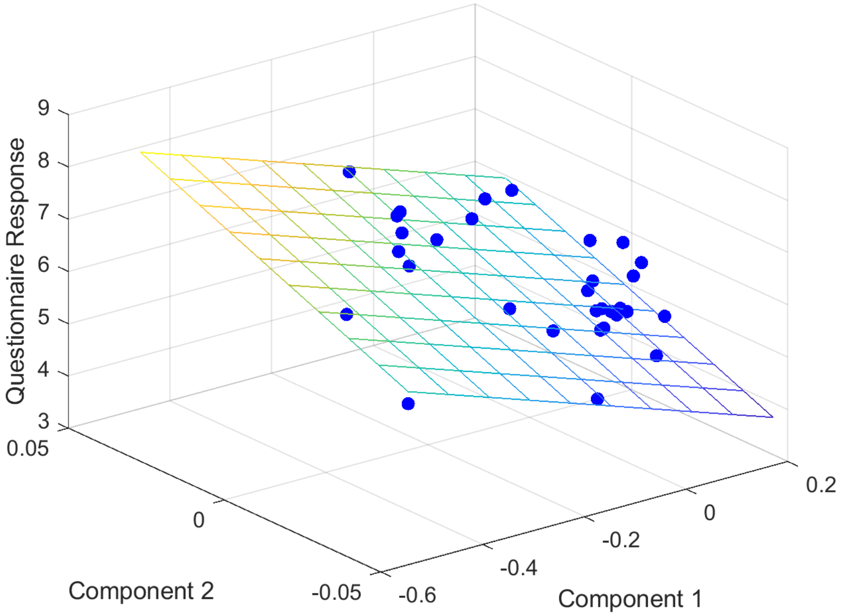


Fig. 10 Dependability, as modeled by the components derived from physiological parameters described in Fig. 7, right. The mesh surface shows the 2-D fit with blue points representing empirical questionnaire responses.

System trust was related to the standard deviation of the pupil size as identified by our correlation procedure depicted in Table 5 and Fig. 11 (left). The GLM fit to these data was significant: F = 26.2, p = 6.4e-6. The coefficient estimate for the predictor in this model was 0.12 with an associated p = 6.4e-6.

System familiarity, as measured by the Trust in Automation questionnaire, was related to HR minima derived in 30-s epochs as shown in Table 5 and Fig. 11 (right). The GLM fit to these data was significant: F = 26.4, p = 4.5e-6. The coefficient estimate for the predictor in this model was 26.4 with an associated p = 4.5e-6.

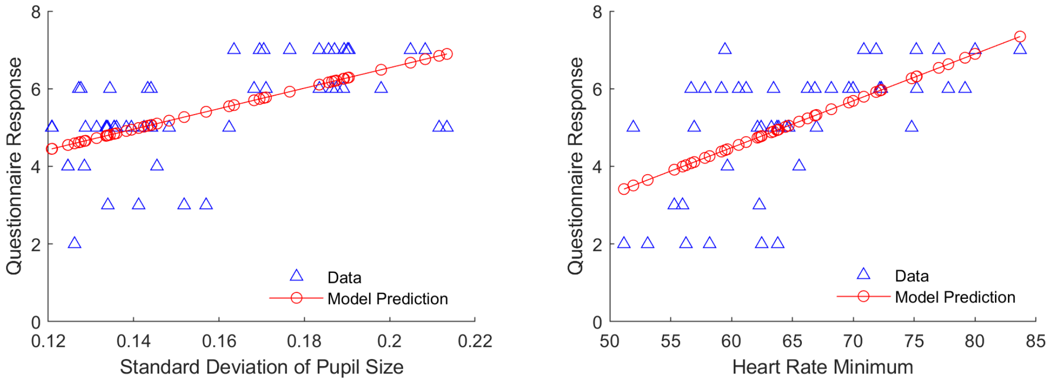


Fig. 11 System trust ratings as a function of the standard deviation of pupil size (left). System familiarity ratings as a function of HR minima in 30-s windows (right).

Belief in system competency, as measured by the System Trustworthiness questionnaire, was related to multiple physiological parameters, as depicted in Table 6. The first three principal components were fit to the questionnaire responses (see Fig. 8); the first two of these were significantly related to competency and are shown in Fig. 12. The overall model fit was statistically significant: F = 15.3, p = 4.4e-6. The coefficients for Components 1, 2, and 3 were 107.83, 351.86, and 168.28, respectively. The first two components were statistically significant with p = 2e-4 and 1.5e-5, respectively. The third component coefficient was not significant: p = 0.35.

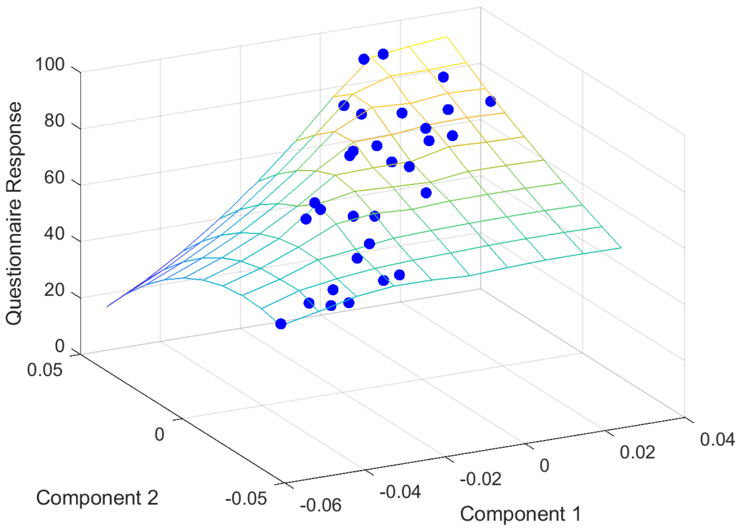


Fig. 12 Belief in system competency, as measured by the System Trustworthiness questionnaire, as a function of Components 1 and 2. Blue data points represent empirical data, and the mesh surface depicts the results of the model fit.

Belief in system reliability, as measured by the System Trustworthiness questionnaire, was related to the P3:O2 WPLI parameter, as depicted in Table 6. A GLM was created that modeled the questionnaire response as a function of P3:O2 WPLI, as shown in Fig. 13. The overall model fit was statistically significant: F = 34.3, p = 1.63e-6. The coefficient estimate for P3:O2 WPLI was 602.54, with an associated p = 1.63e-6.

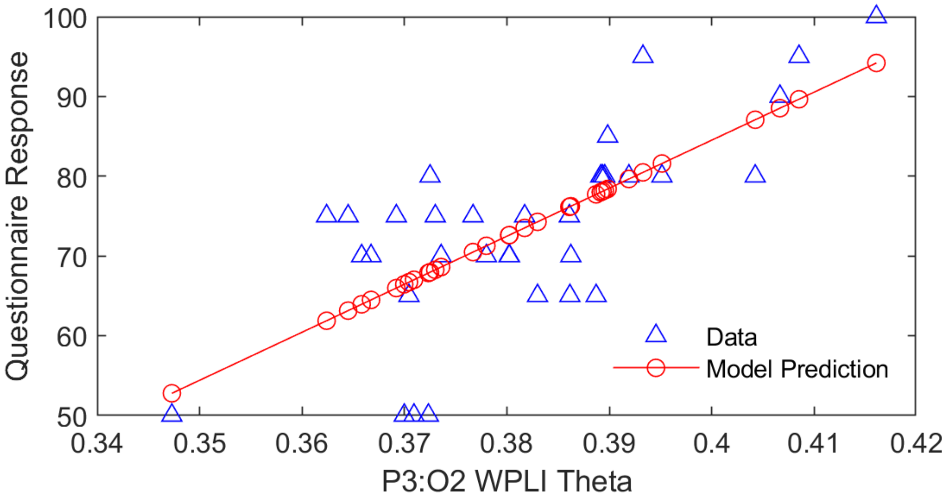


Fig. 13 Reliability, as determined by the System Trustworthiness questionnaire, showed a significant, positive relationship with the weighted phase lag index between the P3 and O2 electrodes in the theta band

5.2.2.3 Discussion

EEG. Previous studies demonstrated similar findings in the context of autonomous vehicle control (Seet et al. 2022) impacted by autonomous vehicle malfunction. This study demonstrated that as autonomous systems malfunctioned, signals from EEG showed significant differences that modulated with the human's trust of the autonomy and their belief in its competency and reliability. These specific EEG signals (frontal and posterior alpha modulated signals) have been related to motivational state and action planning (Wang et al. 2018; Seet et al. 2022). In this context, such an interpretation may make sense insofar as changes in these signals may reflect changes in motivation and action planning related to a person's use of the autonomous aid.

Eye Tracking. The standard deviation of pupil size showed relationships with system dependability and trust. Pupillary oscillations have been studied for quite some time and have been related to measures of cognitive workload and fatigue (Duchowski et al. 2018). In these studies, oscillations up to 0.8 Hz have been correlated with both fatigue and boredom (Tey et al. 2013; Duchowski et al. 2018). In our study, the results are consistent with this observation, as humans become bored or fatigued with reduced cognitive workload that is afforded by use of the autonomous systems.

Heart Rate. Interestingly, HR (specifically the minimum HR) in a 30-s window was positively correlated with familiarity of the autonomous systems. This is consistent with findings in the literature that show increases in HR in musicians as they played familiar music (Fontaine and Schwalm 1979). However, in self-paced exercise, performing familiar exercise demonstrates higher average HR compared to unfamiliar exercises (Lee and Mattison 2012). On the surface, our experimental paradigm is more closely similar to playing music (low physical and high cognitive tasks); however, it is evident familiarity has a complex relationship with cardiovascular physiology that may warrant additional investigation.

Somewhat surprisingly, measures of team trust were not correlated to any individual physiologic measure. This makes sense given that team activities require the interaction of more than one individual, so measures of correlated physiology and/or behavior may be needed to extract meaningful metrics of team trust. Indeed, this is previously documented in the literature with HR (Tolston et al. 2018) and is confirmed by our analysis of the eye-tracking data.

Collectively, our results provide preliminary evidence that trust is correlated with a subset of multiple physiological variables, which may be useful for developing human–autonomy teaming systems that modulate autonomous systems interaction.

5.2.3 Pupil Size as a Measure of Physiological Coherence and Team Trust

5.2.3.1 Methods

Pupil size is modulated by external factors like ambient light, but also influenced by internal mental events and cognitive processing (Mathôt 2018). In this study, we hypothesized a priori that pupil size would correlate over time positively between pairs of subjects in the experiment. This is because subjects were engaged as a cooperative team in simulated combat missions and experienced a similar flow of events within the missions. Further, we hypothesized that pairs manning the same combat vehicle (CV pair) would show a stronger correlation than individuals manning different vehicles (non-CV pairs) within the squad due to the increased coordination and similarity of their experiences in operating the vehicle. To summarize, in all pairs of subjects, we hypothesized a positive physiological coherence (i.e., correlation of pupil size) due to similarity of the general sequence of cognitive events experienced throughout the mission, but we also hypothesized that the correlation would be greater for CV pairs due to the requirement of closer coordination.

We used standard preprocessing procedures to correct for artifacts in the pupil size data due to blinks and other sources of signal dropout, including linear interpolation of epochs consisting of missing data (Thurman et al. 2021). We used a velocity-based approach to identify sudden and physiologically unrealistic (greater than four standard deviations) changes in pupil size and linearly interpolated these data points as well. Once interpolated, we identified a range of time for each mission in which every subject had viable eye-tracking data available because the start and stop points of the eye recording were manually set for each subject. After common start and end time points were determined, we resampled the pupil time series to have a common sampling rate (60 Hz) and total number of data points for all subjects in each mission. Data was resampled to allow analysis of correlations of pupil size between pairs of subjects at equivalent time points in the mission.

Pupil size is a physiological signal that reflects changes in both internal mental events (e.g., cognitive load, decision making, emotion, and attention) and external events (i.e., environmental brightness and other features; Sirois and Brisson 2014). Pupil size can change rapidly on a short timescale (on the order of seconds) to reflect sudden changes in mental processes and/or sudden changes in visual properties of the environment (i.e., pupillary light reflex). On a longer timescale (on the order of minutes), however, pupil size modulations reflect more sustained changes in mental state (i.e., arousal and vigilance) and prevailing changes in the visual environment. For this reason, temporally smoothing pupil size data is an effective way of capturing larger-scale changes in the intrinsic state of the subject in response to the task and environment, while filtering out (or “smoothing over”) the more rapid and idiosyncratic variability that occurs due to rapid luminance changes and sudden task changes on a shorter timescale.

Hence, the next step in preprocessing involved applying a temporal smoothing filter to pupil size data for each subject to capture more robust and long timescale state fluctuations prior to computing person–person pupil correlations. We used a Savitsky-Golay filter (order 3) with a smoothing factor of 5120 data points, corresponding to 85.3 s. Savitsky-Golay filtering uses local least squares fitting with polynomials to smooth data to increase precision without distorting the signal tendency, preserving better signal-to-noise ratio than other smoothing methods (e.g., convolution with Gaussian kernel).

Finally, we computed the correlation of artifact-corrected and temporally smoothed pupil size data between every pair of individual subjects in each mission. There were seven roles total, including one commander and three pairs of subjects designated to operate either the manned control vehicle (MCV) or one of two robotic combat vehicles (RCV1 and RCV2). We categorized each subject-pair as representing subjects operating either the same vehicle (up to three pairs per mission depending on whether there was missing eye-tracking data), labelled “CV pair,” or operating different vehicles (up to 18 pairs per mission) labelled “non-CV pair.” There were some missions in which there were technical issues with eye-tracking data collection, or a subject was not available for the commander role (e.g., 8 out of 16 missions are missing a commander), so the total number of CV pairs and non-CV pairs available for analysis varied by mission.

5.2.3.2 Results

We first examined whether pupil correlations tended to be positive, indicating a general tendency for physiological synchrony. We found that the distribution of mean correlations of CV pairs across 16 missions was positive and significantly different from the null hypothesis of zero, t(15) = 3.86, p = 0.0016, with a mean = 0.25, std = 0.26. For non-CV pairs, the correlation was also positive and significantly different from zero, t(15) = 5.7, p<0.001, with a mean = 0.14, std = 0.10. Consistent with our hypothesis, most pairs of subjects tended to show a positive correlation greater than zero, and CV pairs tended to have a stronger correlation than non-CV pairs (mean = 0.25 and 0.14, respectively).

To test this second hypothesis more directly, we compared mean pupil correlations of CV pairs to non-CV pairs across the 16 missions using a t-test. Results revealed a significant difference between the two groups, t(15) = 2.11, p = 0.05, indicating that CV pairs had a higher physiological pupil correlation than non-CV pairs (Fig. 14). Taken together, these results are consistent with both of our a priori hypotheses showing significant positive physiological synchrony among pairs of subjects and significantly greater correlations for CV pairs that performed more closely as a team in controlling the same CV (Fig. 15).

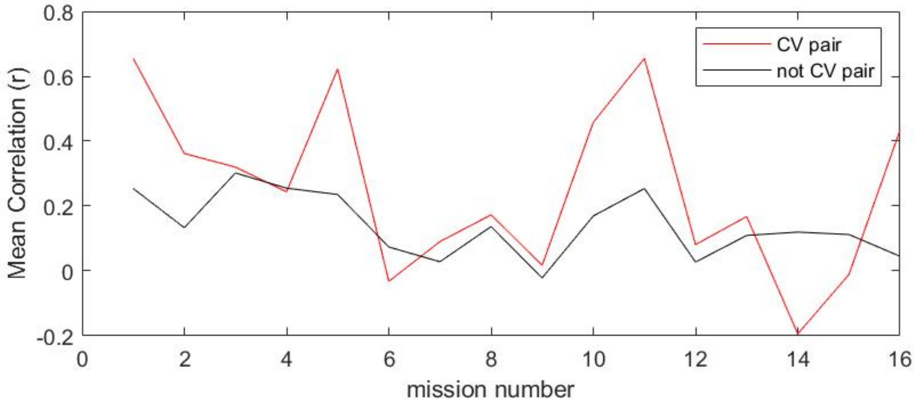


Fig. 14 Mean correlation of all pairs of subjects in each of 16 missions (x-axis) separated by group. The red line represents the mean physiological synchrony of CV pairs (individuals controlling the same CV), and the black line represents non-CV pairs (all other pairings on individuals in the section).

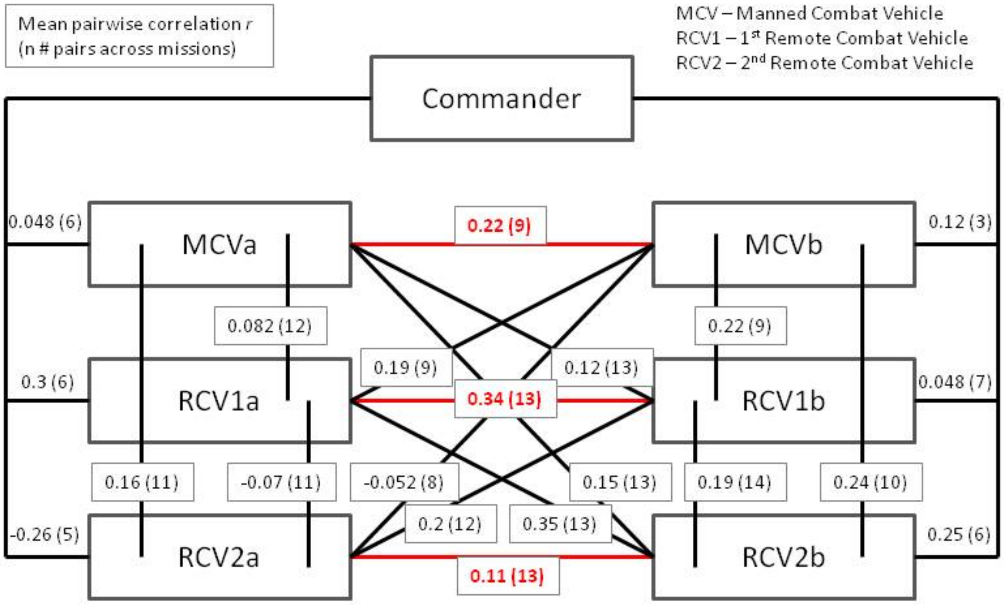


Fig. 15 Mean correlation for each set of possible pairs in the NGCV (with n # of missions of data available in parentheses). Each of the seven large boxes represents an individual subject with a specific role. For example, MCVa and MCVb were the two individuals tasked with operating the MCV. Correlations are shown in red for those classified as "CV pairs." The commander was not specifically paired with any other individual and was only present for 7 of the 16 missions.

Next, we examined the correlation between physiological synchrony and subjective ratings of team cohesion and team cooperation. The mean correlation of physiological synchrony with team cohesion ratings across 16 missions was positive but not statistically significant: r(14) = 0.38, p = 0.14. Similarly, the correlation of physiological synchrony with team cooperation was positive but not statistically significant: r(14) = 0.18, p = 0.5. While these results were in the expected direction of our hypothesis of a positive relationship, we reasoned that sample size was a limitation due to there being only 16 missions total. To increase the sample size and resolution of our analysis, we decided to examine the correlation between these factors at the lowest level of our unit of analysis—individual pairs of subjects (irrespective of which mission it was).

For every pair in each mission, we computed the correlation of their pupil size as well as their mean subjective rating of team cohesion and cooperation. In total, there were 31 CV pairs and 172 non-CV pairs of subjects in this analysis (Fig. 16). Across all pairs we found that there was a significant positive correlation between physiological synchrony and cohesion ratings, r(201) = 0.23, p = 0.001 (Fig. 16, left), as well as a significant positive correlation between physiological synchrony and team cooperation ratings, r(201) = 0.15, p = 0.035. Both of these results are consistent with our hypothesis that variability in subjective ratings of team behavior would be associated with our measure of physiological synchrony (i.e., the correlation of pupil size between a pair of individuals).

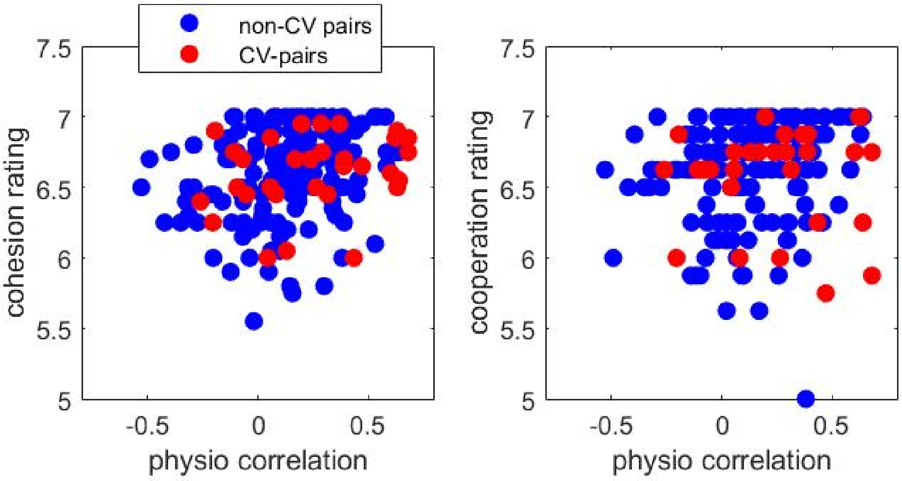


Fig. 16 Scatterplot showing the relationship between pupil (physio) correlation (x-axis) and cohesion ratings (left) and team cooperation ratings (right). Each data point is a pair of subjects in 1 of the 16 missions. There were 31 CV pairs (red) and 172 non-CV pairs (blue) across all the missions.

5.2.3.3 Discussion

Together, these results suggest that the correlation of pupil size between two individuals as a representation of physiological synchrony is a significant indicator for which individuals were working most closely together in a teaming context (i.e., responsible for operating the same CV as a team), and a significant predictor of how well the individuals thought subjectively that they worked well together as a team in terms of team trust (cohesion) and cooperation throughout the varied combat missions. This is a very interesting result that signifies the potential value of tracking pupil size of one or more individuals to predict team-based roles (who was working more closely with whom) and team trust relationships in complex, Army-relevant scenarios.

5.3 Behavioral Measures of Trust

5.3.1 Communication Content and LSM

5.3.1.1 Methods

The speech recordings were transcribed by two different software programs depending on the information needed for analyzing communication content. During the missions, crew members' speech on the different channels was transcribed by Dragon Naturally Speaking, which enabled us to store each transcription along with information about the crew member and on which network they were speaking (i.e., vehicle vs. section), thus providing valuable information for examining speech similarity. After the study concluded, we transcribed the complete audio files from each crew station using Google Speech-to-Text, which we found to be more accurate, but we could not easily parse which communication channel (network) was being used. Therefore, we used the less accurate Dragon transcriptions to examine the similarity between crew members' language usage and the more accurate Google transcriptions to examine content patterns in each crew member's speech. Both transcriptions were processed through a dictionary-based linguistic word count using the Linguistic Inquiry and Word Count (LIWC) process with the LIWC 2015 standard English dictionary (Pennebaker et al. 2015) and, for content analysis, also a custom dictionary. These data were then analyzed for their correlation with responses to the System Trustworthiness, Trust in Automation, and Team Trust questionnaires. In addition, they were correlated with the Trustworthiness, and Cooperation subscales of the Team Trust questionnaire using repeated measures correlations (Bakdash and Marusich 2017). This method allows us to examine relationships within each individual participant and tests for commonalities of these relationships across participants. This method also allows for the control of individual differences in personality, typical speaking patterns, and general tendency to trust. The repeated measures correlation coefficient (r_{rm}) can range from –1 to 1, indicating the strength and direction of the relationship (Bakdash and Marusich 2017). Calculations were performed in R using the R package "rmcorr" (Marusich and Bakdash 2022).

We examined six different aspects of communication content: affect, verbosity, affiliative language, agreement language, first-person language, and descriptiveness. Additionally, we examined one aspect of linguistic similarities between crew members: LSM. The following is a brief description of each aspect and our expectations for the correlations.

5.3.1.1.1 Affect. Trust is largely an emotional (affective) experience (e.g., Dunn and Schweitzer 2005). For example, without trust in one's teammates, a person might feel worried about the performance of the task, irritated at the perceived incompetence of their teammates, and/or stressed about anticipated difficulties. With more trust, a person is likely to feel more confident about the task performance, and they additionally may experience more positive affect arising from more amicable social interactions or stronger social connectedness during the task. For this reason, it is fairly common to include linguistic content indicators of positive and negative affect when constructing machine models of trust (e.g., Beigi et al. 2016; Ghafari et al. 2020). A more positive affect is expected to be associated with greater trust (as found by Khawaji et al. [2013] in dyadic interactions, and as predicted by Licorish and MacDonell [2014] for larger teams). We therefore predict that participant affect in our experiment, as reflected in the emotional tone of their word choices, would be more positive when trust is higher. We predict this to be the case for both human–human and human–autonomy trust.

5.3.1.1.2 Verbosity. Increased verbosity has been associated with higher levels of trust (Lukin et al. 2018). Increased verbosity in team interactions has correlated with stronger team cohesion (Gonzales et al. 2010) and serves as an indicator of team functioning (Fischer et al. 2007). We expect verbosity would likewise indicate team trust levels. Trust is likely to partially guide the amount of communication a participant feels is necessary or appropriate to complete a mission. A participant may feel more comfortable speaking to trusted teammates, for example, and may speak more when trust levels are higher.

In the context of a complex task, we might also predict the opposite. In this case, a participant may feel that an untrusted teammate needs additional explanations and instructions to perform the task and may speak more under this condition. If the untrusted teammate is the machine, additional instructions and explanations may need to be provided to the human teammates to compensate.

5.3.1.1.3 Affiliative language. Feelings of trust may be associated with a sense of camaraderie or team friendliness (as discussed in Licorish and MacDonnell [2014]), and this may be expressed through increased use of affiliative language (language related to companionship, sharing, and togetherness). For example, a participant experiencing a high level of trust in their teammates might be more likely to address them as "bro" or to make references to loyalty or collaboration. We therefore predict trust to correlate with percent affiliation words in the transcripts.

5.3.1.1.4 Agreement language. As with affiliative language, agreement language (assent) may be an indicator of levels of trust. If a participant trusts their teammates, we expect they would be more likely to agree with their decisions. Active agreement, and a mindset of agreeability, would be reflected in greater use of agreement or assent-related language. Increased use of assent language predicts trust levels in conversational dyads (Khawaji et al. 2013) and success in group decision-making tasks (Fischer et al. 2007). We predict the prevalence of assent words to correlate positively with trust for human–human trust. Agreement with the machine teammate is expressed through button presses, so we do not expect human–machine trust levels to impact agreement-related language use.

5.3.1.1.5 First-person language. Pronoun usage can reveal the degree to which a participant is thinking of themselves more as a lone actor during a team activity, or if they are regarding themselves as part of a whole (Stone and Pennebaker 2002). As in Gonzales et al. (2010) and Licorish and MacDonell (2014), we expect participants experiencing high levels of trust would be in more of a collective mindset than an individualistic mindset, and this would be reflected in their relative usage of first-person plural pronouns (such as "we" or "us") versus first-person singular pronouns (such as "I" or "me"). We predict this to be the case for both human–human and human–machine trust.

5.3.1.1.6 Descriptiveness. When conducting a team task in a complex environment with diverse simulated terrain, targets, and hazards, it can be useful to share detailed and up-to-date descriptions among teammates as the mission evolves. Trust may impact the extent of this descriptiveness. One result of low trust may be an increased desire to describe conditions, objects, and circumstances in greater detail. If a participant is experiencing low levels of trust, they may feel the need to be more descriptive when relaying information to their teammates because they do not trust that the teammate will perceive the situation independently or accurately. We predict descriptiveness, as measured by the use of adjectives, to be higher when trust is lower, and we predict this to be the case for both human–human and human–machine trust. While direct verbal communication to the machine teammate is minimal, lower trust in that machine teammate may necessitate greater descriptiveness to the human teammates to make up for perceived or anticipated gaps in performance.

Descriptiveness was scored as the percentage of adjectives (as counted by the standard LIWC 2015 dictionary) less the percentage use of the adjective "red." Because "red" is part of the platoon call signs in this simulation (e.g., "red one" and "red two" represent vehicles), it is spoken extremely often, and a raw count of adjectives would fail to capture a speaker's general descriptiveness. A custom dictionary was therefore created to allow this calculation.

5.3.1.1.7 LSM. When effective team members communicate with each other in service of reaching common goals, they typically converge on certain linguistic norms. In fact, several word categories are designed to take on almost any meaning, so long as those communicating have established what those words should mean. These are called function words, or content-free words, because these words perform specific functions to make complex ideas easier to communicate over time. In contrast to content words such as nouns and verbs, function words do not contain semantic information. Instead, the meaning of function words is based on a shared understanding between speakers. Despite there being only around 400 function words, they make up more than 70% of the most commonly used words in English (Bird et al. 2002), making them abundant and easy to measure across all situations.

To track the rate at which speakers coordinate their use of language, Niederhoffer and Pennebaker (2002) computed the similarity in rates between speakers' function word usage, which represents the rate at which speakers are building common knowledge in the conversation. Later, researchers found that this type of LSM corresponds to greater levels of trust between speakers (Scissors et al 2009; Gonzales et al. 2010). Therefore, we predicted that LSM scores would correspond with higher levels of self-reported trust measures.

LSM is computed using the proportion of function words in speech. LIWC (Pennebaker et al. 2015) defines nine function word categories: auxiliary verbs (e.g., to be, to have), articles (e.g., an, the), common adverbs (e.g., hardly, often), personal pronouns (e.g., I, they, we), indefinite pronouns (e.g., it, those), prepositions (e.g., for, after, with), negations (e.g., not, never), conjunctions (e.g., and, but), and quantifiers (e.g., many, few).

To compute LSM between a speaker and one or more speakers, we use the following formula:

$$LSM = \frac{\sum_{i=1}^{9} 1 - \frac{|s1.prop_i - grp.prop_i|}{s1.prop_i + grp.prop_i}}{9}$$

where i is one of nine function word categories (as mentioned previously), $s1.prop_i$ is the proportion of one speaker's words that fall into category i, and $grp.prop_i$ is the proportion of the rest of the group's words that fall into category i. If there is no difference between speakers in the proportion of words in a category, then synchrony for that category is considered perfect and is valued at 1. On the other hand, if there is a maximal difference in which all words from a category come from one speaker, then speakers are considered completely out of sync in that category and valued at 0. For each mission, we computed two LSM scores for each person: one to compute the similarity between a person's communication with the other communications on the vehicle channel (LSMvehicle) and one to represent similarity on the section channel, where everyone could communicate (LSMsection). Because there was limited communication on the team-only channel, we had twice as many LSMsection scores as LSMvehicle scores, thus limiting our statistical power when examining vehicle channel communications.

5.3.1.2 Results

5.3.1.2.1 Affect. Contrary to predictions, greater trust (as measured by the Task Team Trust perceived trustworthiness subscale) was correlated with a more negative emotional tone ($r_{rm}(102) = -0.20$, 95% confidence interval (CI) [–0.383, –0.009], p = 0.039) (Fig. 17). Human–machine trust measures were not related to affect.

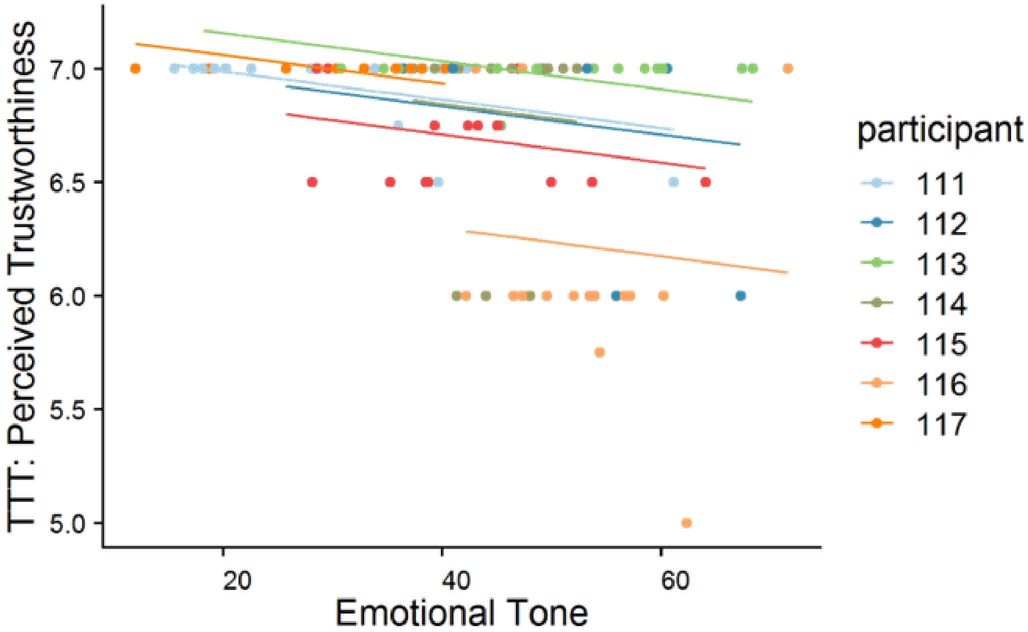


Fig. 17 Emotional tone in spoken language content (higher values indicate relatively more positive affect) vs. Task Team Trust perceived trustworthiness subscale. Each dot represents a single observation, color coded by participant ID.

The negative relationship found between affect and interpersonal trust was quite unexpected and may suggest that task-specific factors can influence the relationship between trust and affect. For example, it is possible that increased interpersonal trust made participants feel more comfortable "venting" about difficulties they encountered during this lengthy and highly complex experimental task. Future work could employ detailed manual annotation of transcripts to uncover whether this is the case.

5.3.1.2.2 Verbosity. Task Team Trust (overall) was significantly positively correlated with word count ($r_{rm}(102) = 0.20$, 95% CI [0.007, 0.381], $p = 0.041$) (Fig. 18), while a trend in the same direction was found for the Task Team Trust perceived trustworthiness subscale ($r_{rm}(102) = 0.19$, 95% CI [–0.004, 0.371], $p = 0.053$). Other trust measures were not significantly related to word count. Our findings support the hypothesis that interpersonal trust may increase levels of comfort or social ease, thus leading to increased quantity of verbal communication. The hypothesis that increased verbal quantity may stem from lack of trust was not supported, neither for human nor agent teammates.

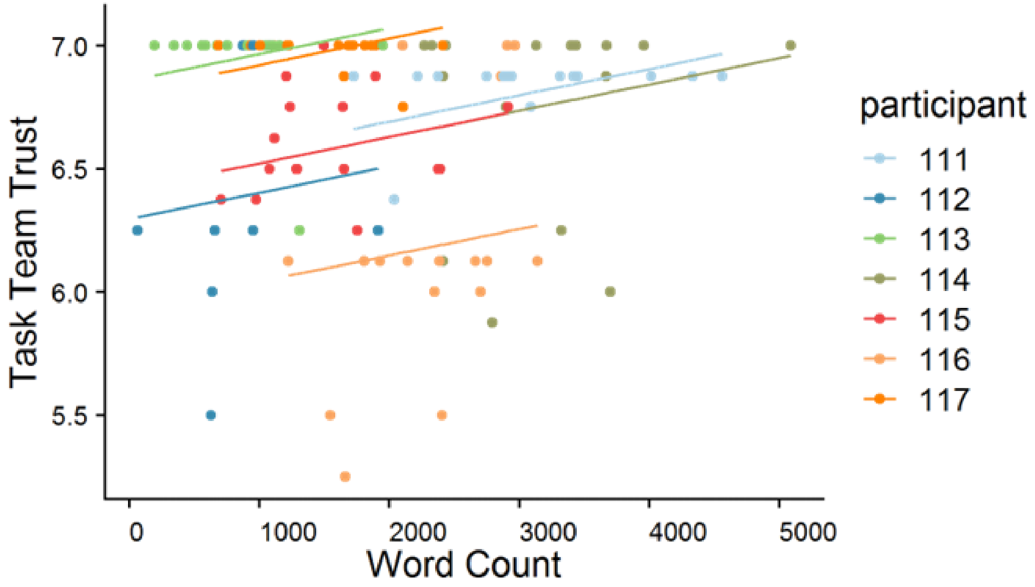


Fig. 18 Word count vs. Task Team Trust (overall). Each dot represents a single observation, color coded by participant ID.

5.3.1.2.3 Affiliative language. A positive trend was found for affiliative language versus the Task Team Trust perceived trustworthiness subscale, $r_{rm}(102) = 0.18$, 95% CI [–0.016, 0.361], $p = 0.069$). No other measure showed a relationship with affiliative language. This suggests interpersonal trust may be associated with greater feelings of group camaraderie as expressed by verbal terms of affiliation, but human–autonomy trust has no such effect.

5.3.1.2.4 Agreement language. No measure of trust was found to be significantly correlated with use of assent language. This was expected for human–autonomy trust measures, but not for human–human trust measures. Our experimental task involved a variety of computer-based map labeling and simulated driving actions, and it is possible that assent was often expressed practically rather than verbally. For example, an assenting participant may have marked an object on the map as instructed by a teammate, in which case the original speaker could see via the computer interface that this action had been taken. It is possible extra assent language was superfluous in this scenario. Examination across diverse experimental task scenarios may reveal whether this is the case.

5.3.1.2.5 First-person language. Contrary to our prediction, first-person language (singular vs. plural) did not correlate with human–human or human–autonomy trust. This suggests that human–autonomy trust may not have engendered a feeling of collective togetherness, at least not as expressed in word choice when speaking with the human teammates.

5.3.1.2.6 Descriptiveness. In line with our prediction, Trust in Automation was found to be related to use of descriptive language, with greater trust being associated with use of fewer descriptive words ($r_{rm}(102) = –0.22$, 95% CI [–0.4, –0.03], p = 0.023) (Fig. 19). This pattern suggests that when the automated features are trusted, information can be successfully conveyed to the agent and to human teammates via the machine interface, reducing the need for verbal descriptions of objects and events. This was the only significant language content-trust relationship found for human–autonomy trust in this analysis. No measures of human–human trust were found to be related to descriptiveness, perhaps because the participants relied heavily on the interface to convey descriptive details, regardless of their trust feelings toward their human teammates.

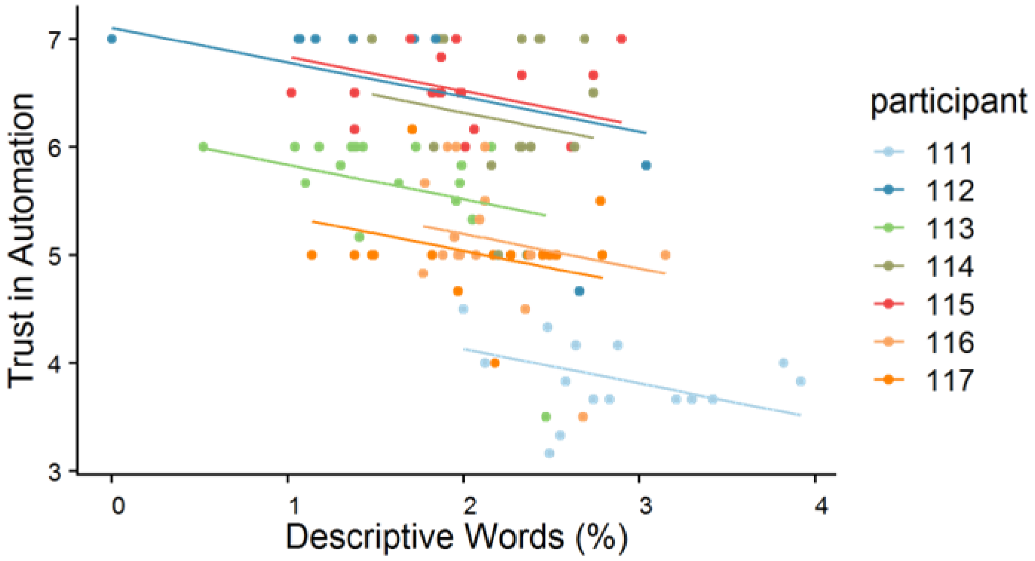


Fig. 19 Average percent adjectives (excluding "red") spoken per session vs. Trust in Automation. Each dot represents a single observation, color coded by participant ID.

5.3.1.2.7 LSM. $\mathrm{LSM}_{\mathrm{section}}$ scores were positively related to self-reported cooperative behaviors, $r_{\mathrm{rm}}(84) = 0.22$, 95% CI [0.01, 0.42], $p = 0.041$. If this result is accurate, it may indicate that LSM promotes trust through enabling crew members to cooperate more effectively. Conversely, this finding may be the result of participants viewing LSM as a form of cooperation.

We found no other significant relationship between LSM scores and other measures of trust. However, we did find three other small but nonsignificant correlations that seem worth mentioning. First, $\mathrm{LSM}_{\mathrm{section}}$ scores also had a positive, albeit nonsignificant, relationship with perceived trustworthiness, $r_{\mathrm{rm}}(84) = 0.16$, 95% CI [–0.05, 0.37], $p = 0.133$, and an apparently negative relationship with Trust in Automation, $r_{\mathrm{rm}}(84) = –0.17$, 95% CI [–0.37, 0.04], $p = 0.112$, which $\mathrm{LSM}_{\mathrm{vehicle}}$ scores seemed to positively correspond with Trust in Automation, $r_{\mathrm{rm}}(47) = –0.21$, 95% CI [–0.08, 0.47], $p = 0.147$.

5.3.1.3 Discussion

As anticipated, several spoken-language indicators showed promise in their ability to predict self-reported trust measures in our simulated CV team task scenario.

Verbal descriptiveness was the only language content measure found to be related to human–autonomy trust. The percentage of words used that were (non-call sign) adjectives significantly negatively correlated with Trust in Automation. This suggests that when human users do not trust the machine interface to correctly record or convey situation information, they compensate by being more descriptive in their verbal communication to their human teammates.

Emotional tone showed predictive value for the percent trustworthiness subscale of Task Team Trust, although the direction of the relationship (negative) was the opposite of that expected and warrants further study. Specific elements of the task itself may be critical to whether, and how, emotional tone or other possible measures of affect are related to trust.

As in Gonzales et al. (2010) and Lukin et al. (2018), higher trust was associated with higher verbosity (word count) for the Task Team Trust overall and nearly significantly with the percent trustworthiness subscale of Task Team Trust. Also similar to Gonzales et al.'s (2010) findings, we found trust was associated with greater LSM between crew members.

With only seven participants, it is not surprising that many comparisons yielded no statistically significant results. It is possible that a larger sample size in future studies may reveal additional effects.

The content, quantity, and linguistic style similarity of verbal communication that human participants use during a mission can provide insights about the participants' internal state, including about their feelings of interpersonal trust and human–machine trust. Transcripts from speech recordings collected during a team task, analyzed post hoc or potentially in real time, can be used as one behavioral indicator of trust levels. Unlike questionnaires, speech measurement can be noninvasive and nondisruptive. Insight from these measurements can inform future interface designs to achieve desired levels of trust and/or could be used to trigger real-time interventions to maintain appropriate levels of trust.

5.3.2 Acoustically Derived Measure of Soldier Engagement

5.3.2.1 Methods

This analysis involves VocSTR (Vocal model of Stress), and a composite trust score. VocSTR is a 3-D convolutional recurrent neural network (CRNN) model that predicts workload from the acoustic features of speech. VocSTR's feature input consists of three vectors: the static log-Mel filterbank energies (Lyons 2017) and their derivatives, delta (magnitude of change), and delta-delta (rate of change). This input provides both spectral and temporal information about the speech signal.

The CRNN architecture for VocSTR is an extension of a model developed to predict emotion from speech (Chen et al. 2018) and trained on data labeled with workload levels (for full details see Scharine and Schaefer 2021). An earlier version of VocSTR was trained with recordings from the Wingman Joint Capability Technology Demonstration exercises held at Ft Benning, Georgia in 2018 and 2019 and the NGCV Phase I Soldier Operational Exercise held at Ft Carson, Colorado in 2020 (Scharine 2021; Scharine and Schaefer 2021). Because the earlier data contained armored vehicle noise, it was necessary to retrain VocSTR on the quiet data.

To create the acoustic feature matrix required for training, all recordings were preprocessed and segmented into 5-min segments sampled at 16 kHz with 16-bit quantization. The filterbank energies were computed for each segment as described in Scharine and Schaefer (2021).

Audio segments with no speech were coded as 0. Audio segments containing speech had a baseline level of 1. To obtain estimates of workload or stress, the time-stamped information about target visibility and distance and hostile engagements were used. Each factor was operationalized and added to the baseline value as a measure of the team's workload (Table 8). The maximum workload value was set to 5, meaning that even if the factors added to greater than 5, the highest possible level was 5. Audio segments not containing speech (labeled 0) were not used to train the model because they constituted the majority of the recorded data and were therefore a source of bias during training. This retrained version of VocSTR showed an accuracy rate of 90.5% on novel test data.

Table 8 Weights used for engagement values

Base	1	2	3	4	Total
Speech no/yes	BLUFOR visible no/yes	OPFOR visible yes/no	2/distance (0.5 km)	Suppressed? yes/no	Base + sum(1,2,3,4)[a,b]
1	0/1	0/1	0/2	0/1	0/5

[a] Values that exceeded 5 were rounded down to 5.

[b] These were rounded to the nearest integer.

A composite trust score was calculated using the ratings from the Trust in Automation and System Trustworthiness questionnaires, which showed some variation in scores across participants and across scenarios. For this analysis, we evaluated six scenarios from the All On condition (i.e., Scenarios 5, 9, 10, 13, 15, and 17). A composite score was created for each of these measures and averaged. Thus, scores of the Trust in Automation questionnaire were averaged for each scenario and each participant, resulting in a score between 1 and 7. Similarly, the scores on the System Trustworthiness questionnaire were averaged for each scenario and participant, resulting in a score between 0 and 100. The scores on the Trust in Automation questionnaire were rescaled to have a maximum value of 100 by multiplying the original scores by (100/7). This score was averaged with the System Trustworthiness score and then the composite score was rescaled to a value between 0 and 5 by multiplying the composite by (5/100). This last rescaling was done to allow an easy comparison to the workload scores.

5.3.2.2 Results

The composite scores for trust and workload are shown in Table 9, and Fig. 20 gives a visual comparison of scores. Some of the corresponding data were unavailable for participants 2 and 6.

Table 9 Composite trustworthiness (T) and workload (W) scores for each participant in Scenarios 5, 9, 10, 13, 15, and 17

Scenario	111		112		113		114		115		116		117	
	T	W	T	W	T	W	T	W	T	W	T	W	T	W
5	1.8	1.1	. . .	. . .	2.4	1.9	3.1	2.7	3.7	3.2	. . .	. . .	2.8	4.4
9	1.9	1.3	4.0	1.9	3.2	3.1	3.0	2.8	3.8	3.3	. . .	. . .	3.8	4.1
10	1.0	1.6	4.8	3.2	3.0	2.7	2.8	2.3	3.7	3.3	2.3	3.8	2.9	4.2
13	1.2	1.5	. . .	. . .	2.3	2.2	2.1	2.8	4.4	3.1	3.1	3.8	2.8	4.5
15	1.4	1.4	. . .	. . .	3.0	2.4	3.3	2.6	4.1	2.5	2.4	3.9	2.8	4.1
17	1.2	1.2	5.0	2.5	1.9	2.6	2.9	2.5	4.3	2.5	1.7	3.9	2.9	4.1

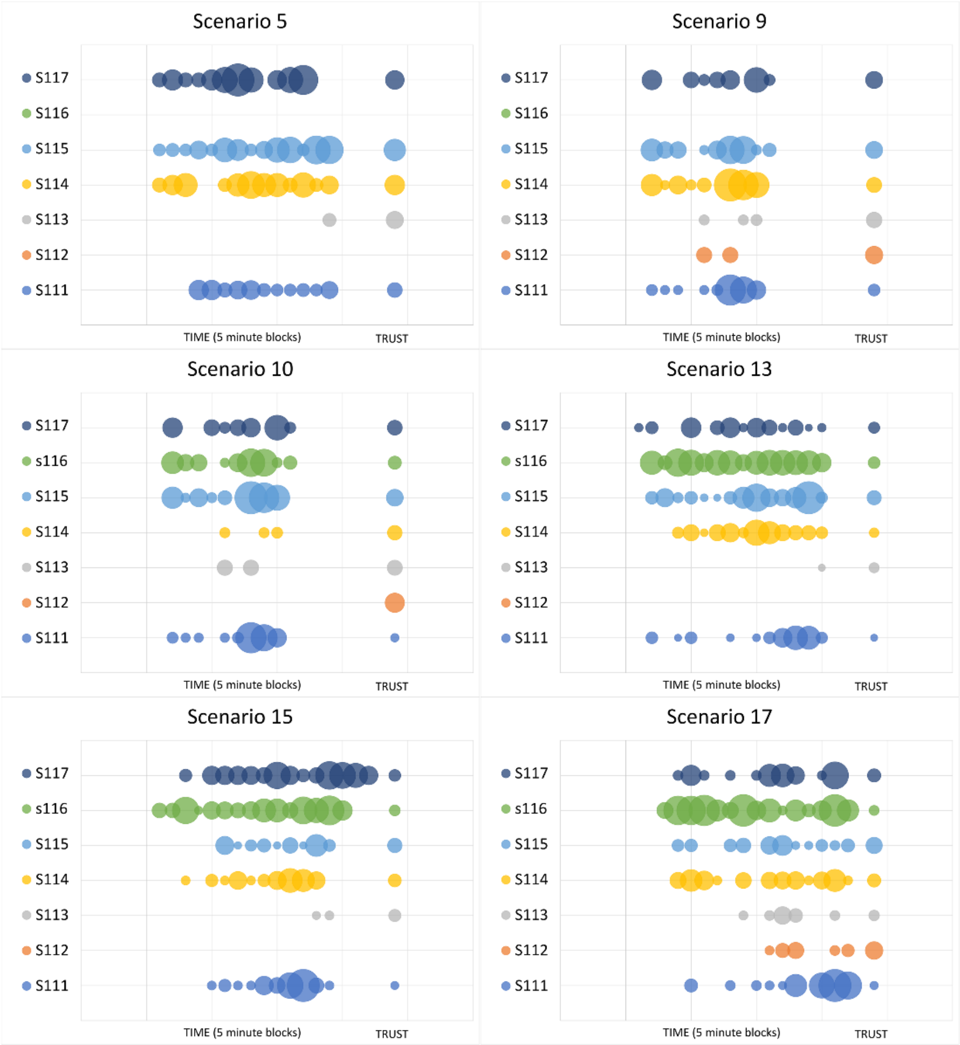


Fig. 20 Frequency and intensity of high workload scores shown in bubbles representing 5-min time intervals for each scenario. The bubbles on the right show the composite trust scores.

It is difficult to make a definitive statement about the relationship between trust and workload, as measured by vocal features. First, not every crew member speaks equally often. Second, the trust measures were administered once per scenario, whereas the workload analysis is a continuous measure. Further, there is no reason that a team characterized by high trust should not experience high workload at times. Rather, the hypothesis is that high-trust teams are better able to manage high workloads. For this data, this might mean that periods with higher workload levels (high stress) will reduce trust relative to a baseline, and conversely, a low-workload period may result in higher levels of trust.

In Fig. 20, the bubbles on the left represent the number and workload level of the utterances for utterances that VocSTR labeled as 3 or greater in a 5-min period. The larger bubbles mean there were more utterances, indicating higher workload. No bubbles indicate that speech was neutral, or no speech occurred. On the right side, the bubbles labeled Trust indicate the composite trust level. This composite trust score is a measure of trust in the autonomous systems. A larger bubble indicates higher trust. No bubble means the participant did not participate in that scenario, or there was no audio data for that scenario.

With respect to the hypothesis mentioned previously, scenarios with longer strings of large bubbles did seem to result in lower levels of trust; for example, the trust levels measured for Scenarios 13, 15, and 17 for S116 resulted in a lower trust score than measured in Scenario 10, where S112 does not register any level of workload and shows the highest level of trust across the scenarios. However, it is by no means clear this is a real trend.

To determine statistically whether it is possible to predict trust from workload, the data were fitted to a simple neural net model. This approach was chosen over linear regression because it was suspected there might be a nonlinear relationship between the clustering of high workload scores and intensity of those scores. The structure of the model is shown in Table 10. Of the datapoints used, 60% were for training, 20% for validation, and 20% for testing.

Table 10 Structure of neural network used to fit workload to composite trust scores

Layer	Output shape	Parameters
Normalization	(None, 37, 18)	. . .
Input	(None, 18)	37
Fully connected layer	(None, 128)	2432
Fully connected layer	(None, 64)	8256
Fully connected layer	(None, 1)	65

When training a neural net model, the algorithm adapts the parameters to minimize loss for the training set, and then tests those parameters on the validation set during each epoch. The objective is to optimize performance for the training set in such a way that performance generalizes well to the validation set. After training is completed, model performance is tested by using the trained parameters to predict performance on novel data.

Figure 21 shows the training history for the model. The blue line is the loss computed for each epoch of the training. The orange line shows the corresponding loss for the validation set. The loss values decrease at approximately the same rate. When tested with the test data, the trained model predicted trust from workload with 28% loss, meaning that on average, predictions were within ±28% of the actual value. While this is a relatively high rate loss, given the small sample size and varying conditions across scenarios, there is a numerical relationship between workload and trust.

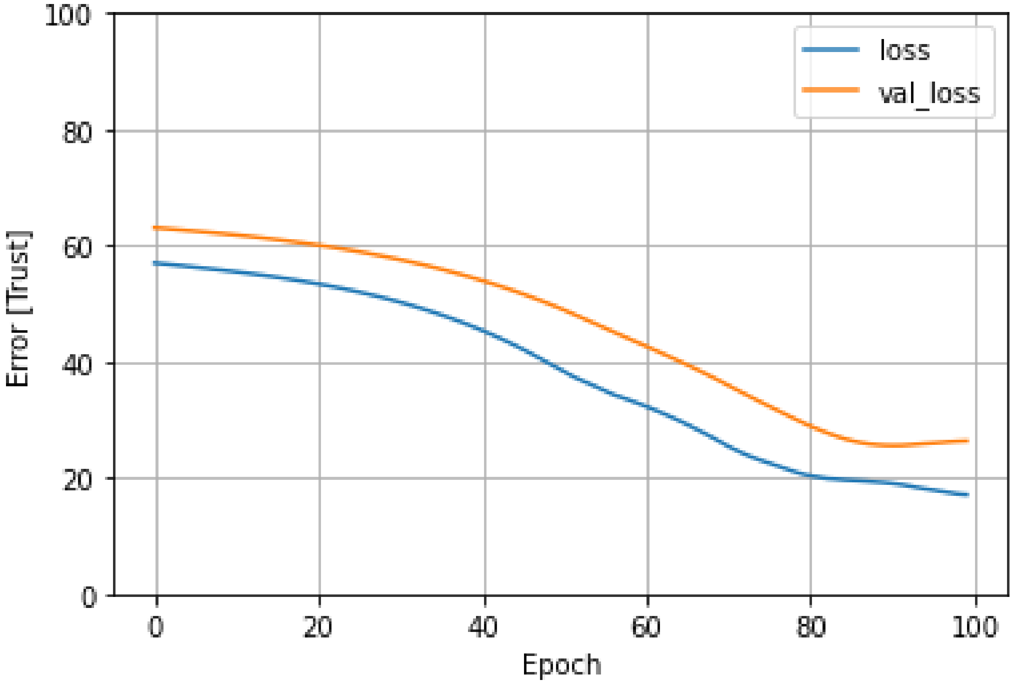


Fig. 21 Training history modeling the relationship between workload and trust. Performance on a novel test set showed 28% error in predictions.

5.3.2.3 Discussion

Some of the participants spoke more than others, resulting in average scores that are based on more speech samples than those of others. It is likely that the low workload scores observed for participant 111 are a consequence of the fact that, as commander, this participant spoke more than others and therefore, had more neutral speech segments. This could skew things for an intervention system. Although the model has promise for identifying periods of high workload, it is not possible at this point to verify whether there is a relationship to trust. It may, however, serve as a trigger for automated interventions that will enhance team trust.

5.3.3 Decision Times

5.3.3.1 Methods

To understand whether self-reported trust measures corresponded to greater team performance, we correlated self-reported trust scores with how quickly crew members identified and responded to threats in their vicinity. These decision times were computed based on the timing of each BLUFOR vehicle's first shot that hit a specific OPFOR target and when that specific target first appeared within the BLUFOR vehicle's vicinity.

To understand how quickly participants responded to OPFOR targets in the simulated environment, we needed to determine when each BLUFOR vehicle (MCV, RCV1, and RCV2) crew first saw each OPFOR vehicle, if at all, and when each BLUFOR vehicle first caused damage to each OPFOR vehicle, if at all. To do so, we first determined which vehicles were within view of OPFOR vehicles, only for instances when the OPFOR was detected by the automated detection system (ADS). The ADS did not provide information about the OPFOR vehicle's identification; to obtain this information, we used the "distm" function in the "geosphere" package (Hijmans 2021) in R version 4.0.2 (R Core Team 2020) to compute each OPFOR vehicle's closest position to the location recorded in the ADS. Then, from these vehicle positions, we selected the vehicle whose timestamp was closest to the timestamp recorded in the ADS. To determine which BLUFOR vehicles were within view of the specific OPFOR vehicle, we created viewsheds for every point in the simulated environment, where each viewshed was a 200 × 200 matrix (representing a radius of 300 m in the virtual environment) and each cell represented whether an object in that location would be visible (0 = not visible; 1 = visible) to someone at the center of the matrix [101,101]. The ADS produced the position of the detected OPFOR vehicle; therefore, to determine which BLUFOR vehicles were visible to the OPFOR at that time (and vice versa), we examined each BLUFOR vehicle's position at the time of the OPFOR's detection, then recorded whether that position was within view of the OPFOR's position. Next, we examined every instance when that OPFOR vehicle was damaged and, working backward from that point, which within-range BLUFOR vehicle's weapon fired closest in time to each of that OPFOR's damage events. Once the time of each BLUFOR vehicle's first hit was recorded, we computed response times for each vehicle by subtracting the first time each BLUFOR vehicle was within range of the OPFOR, if at all, from the first time of each BLUFOR vehicle's first hit, if at all.

Finally, to correlate these data with the self-reported trust questionnaires, for which we only had one measure per person, per mission, we averaged each BLUFOR vehicle's decision times within missions. Further, because decision times were computed at the level of the vehicle, we averaged trust responses across crew members from the same vehicle. Then, to account for the repeated measures nature of these data, which violates the assumption of independence required for Pearson's correlation, we computed repeated measures correlation using the "rmcorr" package (Bakdash and Marusich 2017) in R v. 4.0.2 (R Core Team 2020).

5.3.3.2 Results

We found no relationship between responses to our self-reported trust measures and the average time it took each BLUFOR vehicle to respond to threats. Trust in Automation was unrelated to decision times, $r_{\text{rm}}(55) = –0.017$, $p = 0.899$, nor were cooperative behaviors, $r_{\text{rm}}(55) = –0.086$, $p = 0.525$, nor was task trust, $r_{\text{rm}}(55) = –0.085$, $p = 0.527$.

5.3.3.3 Discussion

These results may suggest the aspects of team performance that trust promotes are not apparent in this metric. Alternatively, this could be due to poor measurement qualities of either the self-reported measures of trust, the measure of team performance based on decision times, or both. To be sure, further research is necessary on the qualities of the self-reported measures, as well as various performance measures in addition to decision times.

6. Conclusions

The purpose of our analyses was to investigate the relationships between team trust and various physiological and behavioral variables. We examined team trust in terms of human–human teams based on responses to the Team Trust questionnaire (Costa 2003) and human–autonomy teams using responses to the System Trustworthiness questionnaire (Muir and Moray 1996) and the Trust in Automation questionnaire (Jian et al. 2000). These questionnaires were collected during the study of technology aids and crew size conducted by Cox et al. (2022). We also used the physiological and behavioral data collected in that study. In our analyses, we found the following variables to be correlated to human–human trust:

- Heart rate
- Heart rate variability
- Respiration rate
- Pupil size for pairs of subjects controlling the same vehicle
- Speech communication content
  - Affect
  - Verbosity
  - Affiliative language
- Linguistic style matching

We found the following variables to be correlated to human–autonomy trust:

- Heart rate
- Heart rate variability
- Respiration rate
- Frontal and posterior alpha modulated EEG signals
- Posterior theta band EEG signals
- Standard deviation of pupil size
- Eye fixation position
- Speech communication content
    - Descriptiveness

We found many correlations between trust and the physiological and behavioral variables that were significant; however, there were some limitations to our analyses based on the study design, sample size, and when the questionnaires were given. The data used in our analyses came from a study of how technology aids and crew size influence the performance of a section controlling armored vehicles. With the focus of the study on technology aids and crew size, and questionnaires about trust only given at the end of each scenario, it is not possible to know about changes in trust that may have occurred within the scenario. Also, it is unclear if changes in trust over the various scenarios caused the changes in the subjects' physiology or behavior. Additionally, there were only six or seven subjects per scenario. Perhaps with more subjects, some of the variables that were trending toward statistical significance would have shown statistically significant differences. Addressing these limitations in future studies will produce clearer results about which physiological and behavioral measures are the best indicators of human–autonomy and human–human team trust. This will have an important impact on the development of systems to monitor the level of team trust and implement interventions if trust levels become too high or too low.

The path forward for this work will involve analyzing data from other studies, conducting studies specifically focused on assessing trust, and examining additional physiological and behavioral measures. Two studies conducted in August 2022 will provide additional data for analysis. Both studies were similar to the technology aids and crew size study. The scenarios were similar, and the same types of physiological and behavioral data were collected. However, the recent studies had more subjects, which should produce more robust results. The data from those studies are being processed, and results will be published in the near future.

With regard to future studies, it will be important to conduct studies designed specifically to examine team trust. In future studies, it will be useful to include instruments that were not part of the technology aids and crew size study, particularly ones that show promise for providing good measures of physiology and behavior related to trust. For example, instruments that measure galvanic skin response and facial expressions should be included. It will be useful to conduct two types of studies. In the first type of study, trust will be manipulated while controlling other factors that can influence physiology and behavior. This type of study can be used to identify more specifically which variables are most important to measure when addressing levels of team trust. The second type of study will be directed toward the development of interventions. This type of study will involve situations in which subjects over-trust or under-trust the autonomous systems. Such studies can verify that the most relevant physiological and behavioral variables are being measured and this information can be used to guide the design of intervention systems.

7. References

Bakdash JZ, Marusich LR. Repeated measures correlation. Front Psychol. 2017;8:456. https://doi.org/10.3389/fpsyg.2017.00456.

Beigi G, Tang J, Wang S, Liu H. Exploiting emotional information for trust/distrust prediction. In: Proceedings of the 2016 SIAM International Conference on Data Mining; 2016. p. 81–89. https://doi.org/10.1137/1.9781611974348.10.

Bird H, Franklin S, Howard D. 'Little words'—not really: function and content words in normal and aphasic speech. J Neurolinguistics. 2002;15(3–5):209–237.

Campellone TR, Kring AM. Who do you trust? The impact of facial emotion and behaviour on decision making. Cogn Emot. 2013;27(4):603–620, doi:10.1080/02699931.2012.726608.

Carreiras C, Alves AP, Lourenço A, Canento F, Silva H, Fred A, et al. BioSPPy - Biosignal processing in Python. 2015 [accessed 2022 Aug 18]. https://github.com/PIA-Group/BioSPPy/.

Chen JYC, Barnes MJ. Supervisory control of multiple robots: effects of imperfect automation and individual differences. Hum Fact. 2012;54(2):157–174.

Chen M, He X, Yang J, Zhang H. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Process Lett. 2018;25(10):1440–1444. https://doi.org/10.1109/LSP.2018.28

Costa A. Work team trust and effectiveness. Pers Rev. 2003;32:605–622. doi:10.1108/00483480310488360.

Costa AC, Anderson N. Team trust. Wiley Blackwell handbook of the psychology of team working and collaborative processes. In: Salas E, Rico R, Passmore J, editors. Wiley; 2017. doi:10.1002/9781118909997.

Cox KR, Rexwinkle JT, Gremillion MG, Perelman BS, Brooks JR, Dyer P, Pollard KA, Forster DE, Chhan D, Carter EC, et al. Effects of technology aids and crew size on a Next Generation Combat Vehicle platoon section. DEVCOM Army Research Laboratory (US); 2022. Report No.: ARL-TR-9403.

Demir M, Likens AD, Cooke NJ, Amazeen PG, McNeese NJ. Team coordination and effectiveness in human–autonomy teaming. IEEE Trans Hum Mach Syst. 2019;49(2):150–159. doi:10.1109/THMS.2018.2877482.

Duchowski AT, Krejtz K, Krejtz I, Biele C, Niedzielska A, Kiefer F, Raubal M, Giannopoulos I. The index of pupillary activity: measuring cognitive load *vis-à-vis* task difficulty with pupil oscillation. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems; 2018 Apr 21–26; Montreal QC Canada. Association for Computing Machinery; c2018. p. 1–13. doi:10.1145/3173574.3173856.

Dunn JR, Schweitzer ME. Feeling and believing: the influence of emotion on trust. J Pers Soc Psychol. 2005;88(5):736–748. doi.org/10.1037/0022-3514.88.5.736.

Elkins AC, Derrick DC. The sound of trust: voice as a measurement of trust during interactions with embodied conversational agents. Group Decis Negot. 2013;22(2013):897–913.

Fischer U, McDonnell L, Orasanu J. Linguistic correlates of team performance: toward a tool for monitoring team functioning during space missions. Aviat Space Environ Med. 2007;78(5):10.

Fontaine CW, Schwalm ND. Effects of familiarity of music on vigilant performance. Percept Mot Skills. 1979;49:71–74.

Fulmer CA, Gelfand MJ. At what level (and in whom) we trust: trust across multiple organizational levels. J Manage. 2012;38(4):1167–1230. doi:10.1177/0149206312439327.

Ghafari SM, Beheshti A, Joshi A, Paris C, Yakhchi S, Jolfaei A, Orgun MA. A dynamic deep trust prediction approach for online social networks. In: Proceedings of the 18th International Conference on Advances in Mobile Computing & Multimedia (MoMM2020); 2020 30 Nov–2 Dec; Chiang Mai, Thailand. p. 11–19. doi.org/10.1145/3428690.3429167.

Gonzales AL, Hancock JT, Pennebaker JW. Language style matching as a predictor of social dynamics in small groups. Commun Res. 2010;37(1):3–19. doi.org/10.1177/0093650209351468.

Grossman R, Feitosa J. Team trust over time: modeling reciprocal and contextual influences in action teams. Hum Resour Manag Rev. 2018;28(4):395–410. doi.org/10.1016/j.hrmr.2017.03.006.

Hergeth S, Lorenz L, Vilimek R, Krems JF. Keep your scanners peeled: gaze behavior as a measure of automation trust during highly automated driving. Hum Fact. 2016:58(3):509–519.

Hijmans RJ. Geosphere: spherical trigonometry. R package version 1.5-14. CRAN.R-project.org; 2021. https://CRAN.R-project.org/package=geosphere.

Jian JY, Bisantz AM, Drury CG. Foundations for an empirically determined scale of trust in automated systems. Int J Cogn Ergon. 2000;4(1):53–71.

Khawaji A, Chen F, Marcus N, Zhou J. Trust and cooperation in text-based computer-mediated communication. Proceedings of the 25th Australian Computer–Human Interaction Conference on Augmentation, Application, Innovation, Collaboration (OzCHI '13); 2013 Nov 25–29; Adelaide, Australia. Association for Computing Machinery; c2013. p. 37–40. doi.org/10.1145/2541016.2541058.

Krausman A, Neubauer C, Forster D, Lakhmani S, Baker AL, Fitzhugh SM, Gremillion G, Wright JL, Metcalfe JS, Schaefer KE. Trust measurement in human–autonomy teams: development of a conceptual toolkit. Trans Hum Robot Interact. 2022;11(3). doi.org/10.1145/3530874.

Langfred C. Too much of a good thing? Negative effects of high trust and individual autonomy in self-managed teams. Acad Manage J. 2004;47:385–399. doi:10.2307/20159588.

Lee L, Mattison M. The effects of task familiarity on performance and strain during a self-paced lifting and lowering task. Ergonomics SA. 2012;24:31–43.

Lewicki RJ, McAllister DJ, Bies RJ. Trust and distrust: new relationships and realities. Acad Manage Rev. 1998;23(3):438–458.

Licorish SA, MacDonell SG. Understanding the attitudes, knowledge sharing behaviors and task performance of core developers: a longitudinal study. Inform Soft Technol. 2014;56(12):1578–1596. doi.org/10.1016/j.infsof.2014.02.004.

Lukin SM, Pollard KA, Bonial C, Marge M, Henry C, Artstein RA, Traum D, Voss CR. Consequences and factors of stylistic differences in human–robot dialogue. In: Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue; 2018 July 12–14; Melbourne, Australia. Association for Computational Linguistics; c2018. p. 110–118.

Lyons J. python_speech_features documentation, release 0.1.0. 2017 Sep 30.

Marusich LR, Bakdash JZ. rmcorr: repeated measures correlation (0.5.2). github.com; 2022. https://github.com/lmarusich/rmcorr.

Mathôt S. Pupillometry: psychology, physiology, and function. J Cogn. 2018;1(1).

Merritt SM, Heimbaugh H, LaChapell J, Lee D. I trust it, but I don't know why: effects of implicit attitudes toward automation on trust in an automated system. Hum Fact. 2013;55(3):520–534. doi.org/10.1177/0018720812465081.

Mitkidis P, McGraw JJ, Roepstorff A, Wallot S. Building trust: heart rate synchrony and arousal during joint action increased by public goods game. Physiol Behav. 2015;149:101–106.

Muir BM, Moray N. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. Ergonomics. 1996;39(3):429–460.

Niederhoffer KG, Pennebaker JW. Linguistic style matching in social interaction. J Lang Soc Psychol. 2002;21(4):337–360.

Pennebaker JW, Mehl MR, Niederhoffer KG. Psychological aspects of natural language use: our words, our selves. Ann Rev Psychol. 2003;54(1):547–577. doi.org/10.1146/annurev.psych.54.101601.145041.

Pennebaker JW, Boyd RL, Jordan K, Blackburn K. The development and psychometric properties of LIWC2015. University of Texas at Austin; 2015. http://hdl.handle.net/2152/31333.

Potts SR, McCuddy WT, Jayan D, Porcelli AJ. To trust, or not to trust? Individual differences in physiological reactivity predict trust under acute stress. Psychoneuroendocrinology. 2019 Feb;100:75–84. doi: 10.1016/j.psyneuen.2018.09.019.

R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing; 2020. https://www.R-project.org.

Salas E, Sims DE, Burke CS. Is there a "big five" in teamwork? Small Group Research. 2005;36(5):555–599.

Schaefer KE, Sanders TL, Yordon RE, Billings DR, Hancock PA. Classification of robot form: factors predicting perceived trustworthiness. Proceedings of the Human Factors and Ergonomics Society 56th Annual Meeting; 2012 Oct 22–26; Santa Monica, CA. Human Factors and Ergonomics Society; c2012. p. 1548–1552.

Scharine A. Development of a neural network algorithm to detect Soldier load from environmental speech models of speech emotion recognition (SER). Advances in Human Factors and Systems Interaction; 2021.

Scharine A, Schaefer KE. Adapting a model of emotional state recognition to detect stress in a high-noise environment. DEVCOM Army Research Laboratory (US); 2021. Report No.: ARL-TR-9137.

Scissors LE, Gill AJ, Geraghty K, Gergle D. In CMC we trust: the role of similarity. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; 2009 Apr. p. 527–536.

Seet M, Harvy J, Bose R, Dragomir A, Bezerianos A, Thakor N. Differential impact of autonomous vehicle malfunctions on human trust. IEEE Trans Intell Trans Syst. 2022;23(1):548–557. doi:10.1109/TITS.2020.3013278.

Sirois S, Brisson J. Pupillometry. Wiley interdisciplinary reviews: cognitive science. 2014;5(6):679–692.

Stone LD, Pennebaker JW. Trauma in real time: talking and avoiding online conversations about the death of Princess Diana. Basic Appl Soc Psychol. 2002;24(3):173–183.

Tey F, Lin ST, Tan YY, Li XP, Phillipou A, Abel L. Novel tools for driving fatigue prediction: (1) dry EEG sensor and (2) eye tracker. In: Schmorrow DD, Fidopiastis CM, editors. Foundations of augmented cognition. Springer; 2013. p. 618–627. doi:10.1007/978-3-642-39454-6_66. (Lecture notes in computer science; vol. 8027.)

Thurman SM, Hoffing RAC, Madison A, Ries AJ, Gordon SM, Touryan J. “Blue Sky Effect”: contextual influences on pupil size during naturalistic visual search. Front Psychol. 2021;12.

Tolston MT, Funke GJ, Alarcon GM, Miller B, Bowers MA, Gruenwald C, Capiola A. Have a heart: predictability of trust in an autonomous agent teammate through team-level measures of heart rate synchrony and arousal. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting; 2018 Sep. Vol. 62, No. 1. p. 714–715. SAGE Publications; c2018.

Wang M, Hussein A, Rojas RF, Shafi K, Abbass H. A. EEG-based neural correlates of trust in human–autonomy interaction. 2018 IEEE Symposium Series on Computational Intelligence (SSCI); 2018 Nov 18–21; Bengaluru, India. p. 350–357. doi:10.1109/SSCI.2018.8628649.

Zhou J, Hu H, Li Z, Yu K, Chen F. Physiological indicators for user trust in machine learning with influence enhanced fact-checking. Machine Learning and Knowledge Extraction. CD-Make; 2019. p. 94–113.

List of Symbols, Abbreviations, and Acronyms

2-D	two-dimensional
3-D	three-dimensional
ADS	automated detection system
ARL	Army Research Laboratory
BLUFOR	friendly forces
CI	confidence interval
COP	common operating picture
CRNN	convolutional recurrent neural network
CV	combat vehicle
DEVCOM	US Army Combat Capabilities Development Command
ECG	electrocardiogram
EEG	electroencephalogram
GLM	generalized linear model
HAT ERP	Human–Autonomy Teaming Essential Research Program
HHC	Headquarters and Headquarters Company
HR	heart rate
HRV	heart rate variability
ID	identification
INFORMS	Information for Mixed Squads
IR	infrared
LIWC	Linguistic Inquiry and Word Count
LSM	linguistic style matching
MCV	manned control vehicle
NGCV	Next Generation Combat Vehicle
OPFOR	opposing force
PC	personal computer

PCA	principal component analysis
R	free software environment for statistical computing and graphics
RCV	robotic combat vehicle
RR	respiration rate
VocSTR	Vocal model of Stress
WPLI	weighted phase lag index

1 DEFENSE TECHNICAL
(PDF) INFORMATION CTR
DTIC OCA

1 DEVCOM ARL
(PDF) FCDD RLB CI
TECH LIB

8 DEVCOM ARL
(PDF) FCDD RLA FA
D FORSTER
FCDD RLA FB
H CROWELL
S THURMAN
FCDD RLA FC
K POLLARD
FCDD RLA FD
A KRAUSMAN
S LAKHMANI
A SCHARINE
FCDD RLA FE
D CHHAN