ARL-TR-9111 ● Nov 2020

DEVCOM ARMY RESEARCH LABORATORY

Human–Autonomy Teaming: Team Trust Metrics—Wingman Simulation Study

by Kristin E Schaefer, Ralph W Brewer, Anthony L Baker, Andrea Krausman, Catherine Neubauer, David Chhan, Evan Carter, Jonroy Canady, Alyssa Milner, Dae Han Seong, Robert Thomson, and Ericka Rovira

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-9111 ● Nov 2020

Human–Autonomy Teaming: Team Trust Metrics—Wingman Simulation Study

by Kristin E Schaefer, Andrea Krausman, Catherine Neubauer, David Chhan, Evan Carter, and Jonroy Canady
Human Research and Engineering Directorate, DEVCOM Army Research Laboratory

Ralph W Brewer
Vehicle Technology Directorate, DEVCOM Army Research Laboratory

Anthony L Baker
Oak Ridge Associated Universities

Alyssa Milner, Dae Han Seong, Robert Thomson, and Ericka Rovira
US Military Academy at West Point

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
November 2020	Technical Report	October 2019–October 2020

4. TITLE AND SUBTITLE
Human–Autonomy Teaming: Team Trust Metrics—Wingman Simulation Study

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Kristin E Schaefer, Ralph W Brewer, Anthony L Baker, Andrea Krausman, Catherine Neubauer, David Chhan, Evan Carter, Jonroy Canady, Alyssa Milner, Dae Han Seong, Robert Thomson, and Ericka Rovira

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
DEVCOM Army Research Laboratory
ATTN: FCDD-RLH-H
Aberdeen Proving Ground MD, 21005

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-9111

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 is unlimited.

13. SUPPLEMENTARY NOTES
ORCID IDs: Kristin E Schaefer, 0000-0002-1342-3446; Catherine Neubauer, 0000-0002-6686-3576; Andrea Krausman, 000-0003-1955-8867; Jonroy Canady, 0000-0003-3555-8779
Approved HRPP No.: ARL-19-175

14. ABSTRACT
This report presents the results from collaborative exploratory research designed to identify and assess different metrics of team trust and team cohesion that may be relevant to military human–autonomy teams. This was a simulation study using the Wingman Joint Capabilities Technology Demonstration simulation testbed to conduct manned–unmanned teaming scenarios on Table VI gunnery evaluations. Participants worked as a team to operate a simulated robotic ground vehicle from a simulated command-and-control vehicle to identify and engage targets on a simulated Army gunnery range. Findings suggest the importance of a multimethod approach to analyzing team trust and team cohesion. Traditional metrics based on team performance ratings or self-report are not indicative of the full team-trust relationship. This research provides valuable insights into how different measurement techniques can provide a more global understanding of the trust relationship.

15. SUBJECT TERMS
human–autonomy teaming, Wingman, trust, wearable technologies, team effectiveness

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Kristin E Schaefer-Lay
Unclassified	Unclassified	Unclassified	UU	61	19b. TELEPHONE NUMBER (Include area code) (410)278-5972

Contents

List of Figures

List of Tables

Executive Summary

US military leaders are exploring the implications of integrating manned–unmanned teams into combat-ready operations whereby a team consists of one or more crews interacting with multiple types of autonomy. One critical aspect of effectively integrating automation in a human–human or human–agent team is the effective development and assessment of trust. This simulation study examined team-trust and team-cohesion metrics building on a multimethod analytical approach including self-reported responses, behavioral and physiological indicators, and communication to explain team effectiveness during a manned–unmanned team gunnery exercise.

The Wingman simulation testbed is a software-in-the-loop version of actual real-world prototype vehicles using advanced mobility and weapon-system autonomy. Within this construct, participants work as a team to operate a simulated weaponized robotic ground vehicle from a simulated command-and-control vehicle to identify and engage targets on a virtual US Army gunnery range. While the original Wingman vehicle includes a 5-man team, the manned vehicle's driver and Long-Range Advanced Scout Surveillance System operator were simulated roles, and the vehicle commander was a confederate, a retired US Army Master Gunner, leaving the participants to fulfill the roles of a robotic vehicle's mobility operator and lethality operator. Our project recruited 28 cadets in groups of 2. Participants experienced two experimental conditions that varied in target exposure time of 100 s and 50 s (50 s is the standard amount of time allotted for manned vehicle gunnery-qualification exercises) and completed 5 target engagements (2 offensive and 3 defensive postures per condition).

Findings from this study provide a foundation for taking a multimethod analytical approach to quantifying team trust. Specifically, performance-based data, self-report, communication, physiology, and behavioral data all provide different yet complimentary indicators explaining the elements that impact team trust and cohesion and, in turn, the effectiveness of the team.

1. Introduction

The US Army seeks to identify current and emerging technologies and projections of technology-enabled concepts that could provide significant military advantage during operations in complex, contested, or congested environments between now and 2028. These include advanced technologies that support integration of joint human–autonomy teaming initiatives. As unmanned or robotic technologies advance from traditional teleoperation to more interdependent operations with advanced autonomous decision-making capabilities, it is essential to develop appropriate collaboration between the human and autonomy-enabled team members (Phillips et al. 2011). A driving reason for this focus on effective teaming is that appropriate use of the technology depends on the human's understanding of the system, its behaviors, and the reasoning behind those behaviors (Chen et al. 2014). If human expectations do not match system behaviors, people will question the accuracy and effectiveness of the system's action (Bitan and Meyer 2007; Seppelt and Lee 2007; Stanton et al. 2007). Such skepticism can lead to degraded trust which, in turn, can be directly linked to misuse or disuse of the system, even if it is operating effectively (Lee and See 2004; Schaefer and Straub 2016).

Research to date has exposed a number of metrics of trust and cohesion for either human-only teams, or small 1:1 ratios of human–robot or human–autonomy teams (Schaefer et al. Forthcoming 2020). However, there is limited research available on identifying key metrics of team trust and team cohesion, especially for these Army-relevant mission needs for human–autonomy teams. Further, it is critical to understand the coordination and cohesiveness of the team. Team cohesion is an emergent state that reflects the extent to which team members relate to each other as a group (Mullen and Copper 1994; Beal et al. 2003). It is affected by mission, environmental, and social context, and is a mediating factor in the relationship between team trust and team performance (Mach et al. 2010; Deortentiis et al. 2013). It is also a critical factor in emotional resilience following periods of stress (Neubauer et al. 2016). Because team coordination and performance are closely related, insights into the cohesiveness of the team can shed light on how the team ultimately performs (Salas et al. 2009).

Our earlier research has shown some promise that taking a multimethod approach (combination of subjective, performance, communication, and physiological measures) provides more information about the team than one measure alone (Schaefer et al. 2019a). This is supported by research that suggests that accurate team-trust and team-cohesion analyses are dependent on communication, physiological response, and emotional state evaluated through facial expressions, word choice, and heart-rate variability (Neubauer et al. 2016). Therefore, the goal

of this study is to conduct exploratory research to identify team-trust and team-cohesion metrics related to performance, behavior, communication, and physiological indicators.

1.1 Subjective Scales

There are many different potential state-based trust scales available. However, most of these scales were developed from either human–interpersonal or human–automation trust fields. Often, broad speculations of trust are made via self-reported items that range from a single question, "How much do you trust this robot?" to questionnaires that do not incorporate the complexities of human–robot collaborative interaction.

Previous research examining changes in trust in the US Army Wingman Joint Capabilities Technology Demonstration's (JCTD's) scenarios found quality differences in trust when looking at the following types of items: perceived level of intelligence of the vehicle, perceived level of automation, perceived trustworthiness, and perceived safety (items were initially developed and tested in Schaefer et al. 2012, used in a previous U.S. Army Research Laboratory research study [ARL-18-165], and reported in Schaefer et al. 2019a). Since a validated scale of cohesion is not available for human–autonomy teams, team cohesion was previously assessed through an experimenter-created questionnaire on team readiness specific to operations within the Wingman context (ARL-18-165; Schaefer et al. 2019a). In addition, measures of stress and workload were collected as both of these factors are variable and unpredictable and can lead to degradations in trust (Biros et al. 2004; Cosenzo et al. 2006; Schaefer and Scribner 2015).

1.2 Performance

Performance is a direct result of the team interoperability, where the manned command-and-control vehicle is often located at a remote position with respect to the robotic Wingman vehicle. Therefore, in support of prior research (Chen et al. 2014; Schaefer et al. 2017a), a technical solution for providing shared situation awareness* (SA) across the human–autonomy team is critical. Accomplishment of this goal rested with the development of the Warfighter Machine Interface (WMI), which provided interactive customizable displays for the commander, robot

*Team SA encompasses joint decisions and actions where each individual agent may have their own SA to carry out individual goals, but there is also a shared SA to enable the coordination needed such that accomplishment of individual subgoals supports the accomplishment of overall goals (Endsley 1995; Endsley and Jones 1997). This shared SA reflects the overlap of information upon which such team coordination is based.

operator, and robot-lethality operator to interact with the robotic vehicle. Each Wingman WMI has access to shared SA data, categorized by subsystem across the bottom of each display, including major subsystems such as map, sensor, alerts, and so on. The map screen provides an interactive aerial image, MIL-STD-2525B* symbols, mobility plans, sensor fields-of-view, and grid-reference lines. The sensor screen provides live video feeds with overlays providing SA such as azimuth, elevation, heading, and field-of-view. The commander and lethality operator use the sensor feeds to positively identify potential targets for engagement. Each WMI also has SA data available in a common toolbar and prioritized alerts visible as pop-ups at the top of the screen. The WMI is set up to record all button presses and associated timing on each display.

Objective measures associated with trust often relate to human behavior, such as number of interventions with the level of automation of a system or the amount of time interacting with or manually controlling a system (e.g., Spain and Bliss 2008). However, these measures, and their relationship with trust, are still being established. Therefore, we record specific interactions with the operation of the robotic Wingman vehicle, including frequency of interventions and duration of time relying on the vehicle, as well as provide standard scoring of the gunnery task using the Department of the Army (DA) Form 8265, the Common Crew Score Sheet.

1.3 Behavioral Indicators

There is a long history in psychological literature in understanding facial expression (Ekman and Friesen 1978). Facial expressivity has been shown to reliably relate to appraisal and coping mechanisms. Past studies by members of our team have found that automatic computations of facial expressivity are comparable to manual annotation of single action unit (AU) intensity (Neubauer et al. 2017) and have been used in a number of studies (Batrinca et al. 2013; DeVault et al. 2014; Chollet et al. 2015; Scherer et al. 2016; Parra et al. 2017). As such, there is evidence that automatic behavior trackers can provide researchers with much-needed objective assessments of behavioral indicators of stress, fatigue, or even trust through a nonintrusive measurement modality on a continuous time scale.

1.4 Communication

Effective communication results in improvements in other team processes and outcomes (Mathieu et al. 2000; Kozlowski and Ilgen 2006). Efficient team

* DOD's standard for common warfighting symbology, both for automated and hand-drawn graphic displays.

communication is critical for target engagement given the coordinated nature of gunnery operations. Since Wingman adds different types of autonomy to the equation, it is even more important to understand how team communication relates to performance, given that human–autonomy interaction still lacks the fluidity of human–human interaction (Bisk et al. 2016). As such, there is a need to test our methods for analyzing team communication to ensure they are applicable to the human–autonomy context. In general, understanding these changes in communication, as well as fluctuations in state, are important for human–autonomy team assessment because they provide continuous feedback regarding the nature of the interaction between the human and the system within a given environmental, mission, and social context (Marathe et al. 2020).

Communication analysis is a critical part of the human–autonomy team analysis because effective communication undergirds team processes and successful outcomes (Mathieu et al. 2000; Kozlowski and Ilgen 2006). As a vehicle through which team members can distribute information, synchronize information from multiple sources, resolve disagreements, and align goals (Salas et al. 2005), communication is fundamental to team performance (Marks et al. 2001; Mesmer-Magnus and DeChurch 2009).

There are many approaches for evaluating team communication that can be applied to developing metrics for human–autonomy teams. Some initial metrics that are currently being explored for understanding team trust and team cohesion include latent semantic analysis (LSA; Gorman et al. 2003), language style matching (LSM; Gonzales et al. 2010), and turn-taking or communication flow (Baker et al. 2019, 2020). For this study, because the crew's communication largely comprised the commander's instructions to the team, these communication-analysis approaches would not be suitable since they rely on naturalistic, diverse communication content (LSA and LSM) or on the existence of varied patterns of who speaks to whom (communication flow). Therefore, an alternate approach was used that derives information about the team's gunnery performance from its communication during the gunnery process.

1.5 Wearable Technologies: Physiological Indicators

Previous research has suggested that there may be psychophysiological changes, as measured by electrodermal activity (EDA), heart rate (HR), and heart-rate variability (HRV), associated with a change in trust or the onset of a trust-based decision (Montague et al. 2014; Mitkidis et al. 2015). While this is new research, these types of measures have a long history of being associated with cognitive effort and decision-making capability that is critical to understanding interaction with

human–autonomy teams. Specifically, EDA, both tonic and phasic, is a measure of skin conductance and is a sensitive measure of emotional arousal such that levels increase during periods of anxiety and cognitive effort (Shi et al. 2007). Work by Bethel et al. (2007) was one of the first examples that used EDA during human–robot interaction studies. They found that tonic EDA measures increased with increased engagement with the robot. EDA measurement has also been shown to be generalizable across similar participants. Montague et al. (2014) found that individuals with similar, high ratings of trust in technology showed similar changes in EDA patterns. For example, if EDA levels of one subject were low, and both subjects were in a trust state, the second subject's EDA would be expected to be low as well. Conversely, patterns of high EDA corresponding to low subjective trust measures were reported during interaction with an unreliable robot (Sanders et al. 2012).

HR and HRV are often used in conjunction with EDA to infer the cognitive and affective response to a stimulus. Following a stimulus, an acute decrease in HRV, along with a simultaneous phasic response, has been associated with orienting behavior (Figner and Murphy 2011), which suggests that an event was salient to the subject. HRV may also be used in conjunction with EDA to infer levels of workload and trust (Matthews et al. 2005; Mehler 2009; Montague et al. 2014). Suggested from the previously cited literature, it can be posited that in a state of high trust it is unlikely one would feel anxious and, therefore, HRV would be high and HR low while tonic EDA levels would also be low. However, if there is an increase in cognitive workload and anxiety associated with the process of maintaining SA, such as what would occur in a state of low trust, it is likely HRV would fall and nonspecific phasic EDA levels would rise along with tonic EDA levels.

1.6 Current Work

This research was an exploratory study to identify possible team-trust and team-cohesion metrics associated with human–autonomy teams for Army operations. The task goals and outcomes supported the Wingman JCTD to provide technological advances and experimentation to increase the autonomous capabilities of robotic combat-support vehicles. A goal of this program was to advance human–autonomy teaming initiatives by iteratively defining and decreasing the gap between autonomous vehicle control and required level of human interaction (Schaefer et al. 2019b; Brewer et al. 2020). The larger research goals of team trust and cohesion were in line with the goals for the US Army Combat Capabilities Development Command (DEVCOM) Army Research Laboratory's (ARL's) Human–Autonomy Teaming Essential Research Program to

develop effective measurement techniques and metrics for team trust in human–autonomy teams (Schaefer et al. 2020). Finally, this joint research supported a Senior Capstone Research Project for two US Military Academy (USMA) cadets.

This simulation study used the software-in-the-loop Wingman simulation testbed that allows a human crew to interact with the actual robotic-vehicle autonomy on a realistic gunnery task. The original setup was designed to support a 5-man crew, but for this study, the manned vehicle driver and Long-Range Advanced Scout Surveillance System (LRAS3) operator were simulated roles and the vehicle commander was a member of the experimental teams. This allowed us to assess the robotic vehicle's mobility- and lethality-operator dyad while providing consistency and repeatability across participants. This study provides a means to assess and identify the interdependencies of using a multimethod approach to quantify team trust and cohesion.

2. Methods

Research was conducted under the oversight of DEVCOM ARL and the USMA institutional review boards and approved under ARL-19-175.

2.1 Participants

A total of 48 participants (36 from the original protocol and an additional 12 from an amendment) were recruited from the USMA cadet population enrolled in the Introduction to Psychology for Leaders course. After removing participants due to technical difficulties or no-shows of one or both team members, the results included the findings from 28 participants (14 dyads). Two of these dyads were unable to complete the study due to time constraints, and so their data sets were removed, resulting in a final data set of 24 participants (12 dyads).

All participants had completed basic military training; some cadets were former enlisted Soldiers. While we did not collect demographic data for this study, cadets could range from 18 to 24 years of age, could represent all 50 states, and 23% of the Corps of Cadets are women. Cadets signed up to volunteer for this study through the SONA system. SONA is an online participant pool management system that provides for scheduling and management of research projects, including providing cadets with a description of research projects.

2.2 Task

Participants conducted three engagement runs: a practice session (target exposure time unlimited), Simulated Gunnery Exercise 1 (target exposure time doubled from

standard Army doctrine: 100 s), and Simulated Gunnery Exercise 2 (target exposure time set to Army standard for a Table VI gunnery exercise: 50 s). Each run contained a minimum of five target engagements (i.e., five different sets of targets) and included both offensive and defensive operations on both stationary and moving targets. Two different courses were created, matched for consistency and difficulty, and counterbalanced so that participants did not get the same set of targets for each exercise. The order of the conditions was set, Exercise 1 followed by Exercise 2.

The purpose of the practice run was to familiarize participants with the task and role. All questions were answered during the practice run. Participants then completed the two simulated gunnery exercises in an assigned role. The commander role was fulfilled by a confederate who was trained in how to respond for a successful mission.

2.3 Crew Roles

The five crew-member roles were commander, robotic-vehicle lethality operator, robotic-vehicle mobility operator (Fig. 1), LRAS3 operator, and manned-vehicle driver. For the purpose of this experiment, the LRAS3 operator and manned-vehicle driver roles were simulated. The roles are described in detail as follows:

Commander (confederate role filled by a member of the experimental team, a retired Armor Master Gunner in the US Army) is responsible for mission objectives and operational outcomes; issues the crew a fire command when enemy contact is reported and then reports battle damage assessment to the tower.

Robot Lethality Operator (participant) is responsible for robotic gunnery operations, monitors autonomous targeting capability, and engages targets using a combination of touch commands and hand grip controls (e.g., initiates slew-to-cue gunnery movements, makes fine-grained gunnery movements to control for wind and terrain, and makes firing decisions).

Robot Mobility Operator (participant) is responsible for robotic vehicle's mobility, plans initial routes, monitors autonomous mobility of vehicle and terrain, adapts autonomy control allocation (switch between autonomous control modes, such as waypoint following or teleoperation), and pauses/resumes vehicle to support gunner.

Manned Vehicle Driver (simulated) controls the mobility of the manned vehicle so that the sensors (such as the LRAS3) onboard this vehicle can support the gunnery task.

LRAS3 Operator (simulated) is responsible for locating and lasing targets, which updates the WMI and supports the crew's SA and weapon-system autonomy features.

Fig. 1 Cadets in the role of Robot Lethality Operator (left) and Robot Mobility Operator (right)

2.4 Wingman Simulation Testbed

The Wingman simulation testbed is a software-in-the-loop simulation environment. This means it integrates all of the real-world Wingman vehicle software into a lab-based virtual setting. The design and development is available for review in ARL-TN-0830 (Schaefer et al. 2017b), ARL-TR-8254 (Schaefer et al. 2017c), and ARL-TR-8572 (Schaefer et al. 2018). It was designed to support the 5-man crew station on a command-and-control vehicle, where the roles could be manned roles or simulated. The virtual environment includes a virtual gunnery range from Camp Grayling, Michigan, and possible targets include both troop and vehicle (stationary and moving) targets.

The Wingman vehicle supports several levels of navigation autonomy, including teleoperation, waypoint finding, semiautonomous driving, and full autonomy. The operator's goal is to effectively switch between the different control modes and qualify in several different scenarios. The Wingman vehicle can help operators find potential targets and even keep weapons aimed on those targets. Ultimately however, it is the lethality operator's responsibility to decide whether to fire or not.

2.4.1 WMI

Participants interacted with a touchscreen computer. The software is the Wingman WMI developed by DCS Corporation for the Army. It provides an interactive platform that provides shared SA and cooperative team operations. The commander and two robotic operators each have a WMI display. This display provides a dynamic map showing the vehicle's placement in the world, identified targets, and sensor data. It also allows the robotic operators to interact with the vehicle and set different levels of autonomy.

2.4.2 Gamepad Controllers

Participants used a standard gamepad controller to teleoperate the robotic Wingman vehicle (Fig. 2) and operate the simulated weapon system (Fig. 3).

Fig. 2 Controller layout for robotic mobility operator's controls

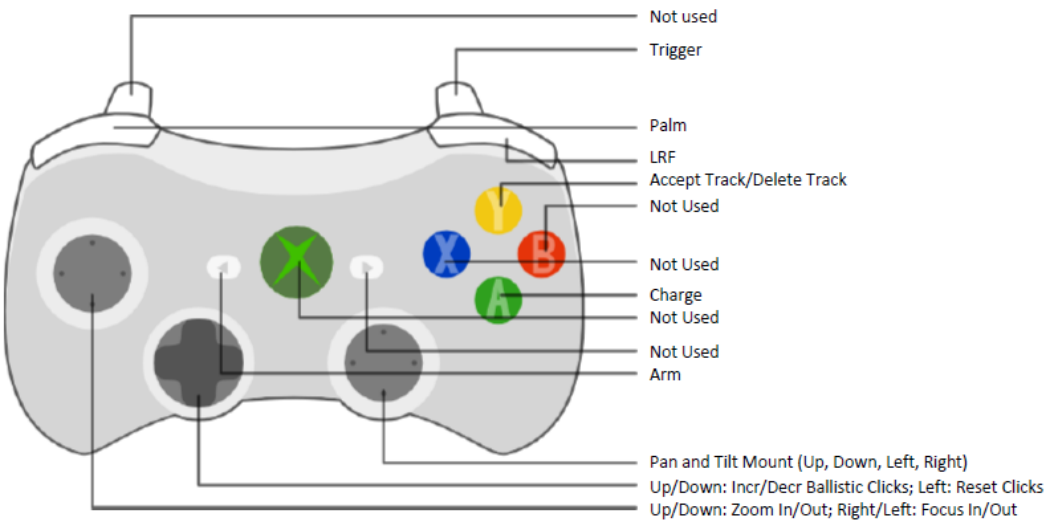

Fig. 3 Details of controller layout for robotic lethality operator's controls

2.5 Physiological and Behavioral Equipment

In order to assess potential physiological indicators of trust, the Empatica E4 sensor provided unobtrusive measures of HR and EDA. To assess additional behavioral indicators of trust (other than button presses on the WMI and autonomy use), the Logitech C920 HD Pro Web Camera was used to record changes in facial features.

2.5.1 Empatica

The Empatica E4 (Fig. 4) is a wrist-worn sensor with a photoplethysmography (PPG) sensor—providing blood volume pulse (BVP) from which HRV can be derived—and accelerometer, EDA sensor, and IR thermopile (skin temperature). For this study, we are interested in HR, HRV, and EDA. The Empatica E4 was cleaned with a sanitizing wipe after each participant and placed on the overnight charging device at the end of day.

Fig. 4 Empatica E4 Sensor

2.5.2 Digital-Camera Facial Video and Microphone Audio

Subject-oriented video and audio were continuously recorded during the experiment using a Logitech C920 HD Pro Web Camera.* The Logitech C920 is a commercially available, desktop-mounted stand-alone video camera and microphone from which metrics of posture, facial expression, eye blinks, brow raises, and speech semantics may be derived. As shown in Fig. 5, this system is mounted on a simulation computer monitor, and no part of the device is attached to the participant.

Fig. 5 Logitech C920 HD Pro Camera with autofocus and built-in stereo microphone; shown mounted on desktop monitor

2.6 Facilities

An 8- × 10-ft laboratory was used for data collection. The laboratory was equipped with lab tables and chairs and located on the 2nd floor of Thayer Hall, USMA, West

* https://www.logitech.com/en-us/product/hd-pro-webcam-c920; Logitech, Switzerland.

Point, New York. Researchers coordinated with the Department of Behavioral Sciences and Leadership's Research Psychologist to schedule laboratory use.

2.7 Questionnaires

Questionnaires' topics included trust in the robotic vehicle, team readiness, stress, and workload.

1) *Trust in the team/robotic vehicle*: Subjects rated their trust in the team and trust in the robotic vehicle. Items included 7-point Likert-type questions rating the perceived level of intelligence of the vehicle, perceived level of automation, perceived trustworthiness, and perceived safety. These items were adapted from Schaefer et al. (2012) and were used in prior Wingman studies (ARL-18-165; Schaefer et al. 2019a).

2) *Team Readiness*: Subjects rated their team readiness (7-point Likert-type questions). These items were developed for a previous protocol (ARL-18-165; Schaefer et al., 2019a):

    a. Team readiness to operate the robotic vehicle

    b. Self-confidence in operating the robotic vehicle or robotic weapon system.

    c. Confidence in the team to operate the robotic vehicle

    d. Confidence in the team to operate the weapon-system autonomy

    e. Trust in the weapon-system autonomy of the robotic vehicle

    f. Trust in the mobility-system autonomy of the robotic vehicle

    g. Trust in the team to conduct gunnery operations with the robotic vehicle

3) *Stress*: Because of the improved discriminant validity and the control of checking the response set, the Multiple Affect Adjective Checklist–Revised (MAACL-R; Lubin and Zuckerman 1999) form has been found to be particularly suitable for investigations that postulate changes in specific affects in response to stressful situations. This form consists of a list of 132 adjectives for which participants are instructed to check all of those words describing how they "feel right now" or "during the simulation." The MAACL-R assesses five affective dimensions: anxiety, depression, hostility, positive affect, and sensation seeking; also, two composite dimensions: dysphoria (DYS), and Positive Affect and Sensation Seeking

(PASS) (Lubin and Zuckerman 1999). Table 1 shows each of the dimensions and their associated adjectives. Raw scores are calculated for each dimension and converted to T-scores prior to analysis.

Table 1 Seven dimensions of stress from MAACL-R questionnaire

Dimension	Adjectives
Anxiety	Afraid, fearful, frightened, panicky, shaky, tense
Depression	Alone, destroyed, forlorn, lonely, lost, miserable
Hostility	Annoyed, critical, cross, cruel, disagreeable
Positive affect	Measures a state/trait of low arousal or calm; adjectives include happy, joyful, pleasant
Sensation seeking	Measures a state/trait of arousal or positive level of activation; adjectives include adventurous, daring, and energetic
DYS	Anxiety + depression + hostility
PASS	Positive affect + sensation seeking

4) *Workload*: The NASA-Task Load Index (NASA-TLX) six-item task-load index (Hart and Staveland 1988) provided workload assessment specific to mental demand, physical demand, temporal demand, performance, effort, and frustration. The NASA-TLX was used to evaluate participants' perceived workload level in these areas on 10-point scales.

2.8 Performance Metrics

Specific interactions with the operation of the robotic Wingman vehicle were recorded, including frequency of interventions and duration of time relying on the vehicle. These metrics are automatically collected in the WMI log file for each user station. Screen-capture software was used to record the state of each of the robotic operators' WMI screens to understand how each user interacted with their system.

Each run was also evaluated using the DA Form 8265, the Common Crew Score Sheet (Fig 6). Each engagement within a run is worth a maximum of 100 points. Shortcomings, errors, and inaccurate or incorrect responses during conduct of fire were recorded. There are four categories of crew-duty penalties: 1) immediate disqualification (i.e., extremely hazardous conduct), 2) automatic zero (i.e., disregard for announced actions, conditions, or standards), 3) 30 point (i.e., failure to adhere to basic safety or personnel-protection precepts), and 4) 5 point (i.e., failure to perform fundamental leader–crew tasks). These penalties are subtracted from the score.

TARGET ENGAGEMENT TIME		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
POINTS		99	97	96	94	93	91	89	88	86	85	83	81	80	79	78	77	75	73	72	70	65	60	55	50	45	40	35	30	25	20	15	10	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

TARGET	TARGET	15. TARGET 3	16. TARGET 4	17. MALFUNCTIONS	18. PENALTY CODES / REMARKS	NOTE: Assess DQ, AZ, and 30-point penalties and enter penalties in Block 20e. Assess 5-point penalties at the end of each day or night phase on the Common Crew Roll-Up, and enter points in Block 21c.			
a. KILL TIME A 15	a. KILL TIME 31	a. KILL TIME	a. KILL TIME	a. BREECH UP					
b. MODIFIER B (-) 8	b. MODIFIER (-) 10	b. MODIFIER (-)	b. MODIFIER (-)	b. CASE BASE		20a. NUMBER OF IMMEDIATE DISQUALIFICATION (DQ) PENALTIES:	0	21a. ENG SCORE (from block 20f):	94
c. DEFILADE C (-) 13	c. DEFILADE (-) 13	c. DEFILADE (-)	c. DEFILADE (-)	c. MISFIRE		b. NUMBER OF AUTOMATIC ZERO (AZ) PENALTIES:	0	b. TOTAL 5-POINT DEDUCTIONS (Multiply Block 20d times 5 points):	0
d. BREAK D (-)	d. BREAK (-)	d. DELAY / BREAK (-)	d. DELAY / BREAK (-)	d. STOPPAGE		c. NUMBER OF 30-POINT SAFETY PENALITES:	0	c. TOTAL LEADER OR FUNDAMENTAL CREW PENALTY POINT DEDUCTION ALLOWED AT END OF FIRING PHASE (Use the lesser of Block 21a or 21b):	0
e. ENG TIME E (*) -6	e. ENG TIME (*) 8	e. ENG TIME (*)	e. ENG TIME (*)	e. THERMAL FAIL		d. NUMBER OF 5-POINT PENALTIES:	0		
f. POINTS F 100	f. POINTS (+) 88	f. POINTS (+)	f. POINTS (+)	19a. TOTAL POINTS (*) 188	b. NO. OF TGTS (/) 2	c. BASE SCORE (*) 94	e. ENG PENALTIES (-) 0	f. ENG SCORE (*) 94	d. Did crew obtain 70 points per target? YES e. QUALIFIED ENG YES

Fig. 6 Common Crew Score Sheet

On the score sheet in Fig. 6 the kill time for each target is entered in (A). To calculate the modifier, the engagement modifier table is used, taking into account the platform, the vehicle posture, target type, target posture, and the range to the target. After entering that in (B), a defilade time is entered in (C) if the firing vehicle was in a defensive position. Then, the modifier and defilade time are subtracted from the kill time to determine the engagement time, which was then entered in (E). The points then are referenced from the target-engagement-time row on the top with the associated point underneath; that value is entered in (F). The total is divided by the total number of targets for determination of the base score. Then, any penalties are subtracted from the base score to get the engagement score. To qualify an engagement, the crew must score a minimum of 70 points per target and have an engagement score of at least 70 points (scores are provided in Appendix A).

2.9 Behavioral and Communication Measures

2.9.1 Facial Expressivity

The participant's face was continuously recorded throughout the task via a webcam mounted to the simulation screen. Features relating to emotional expression were quantified via the OpenFace software platform (Baltrušaitis et al. 2016). Overall, this platform provides automatic assessments of frame-by-frame, single-AU evidence following the Facial Action Coding System (FACS; Ekman and Friesen 1978). Following FACS, facial expressions relating to universal emotions (i.e., happiness, sadness, surprise, fear, anger, disgust, and contempt) were manually calculated, on a frame-by-frame basis separately for each task, using computations of single-AU evidence. Table 2 outlines the specific AUs needed to calculate each universal emotion.

Table 2 Facial-expression emotion calculation from single AUs (adapted from Ekman and Friesen 1978)

Emotion classification	Action units
Anger	4+5+7+23
Contempt	R12A+R14A
Disgust	9+15+16
Fear	1+2+4+5+7+20+26
Happiness	6+12
Sadness	1+4+15
Surprise	1+2+5B+26

2.9.2 Crew Audio

Audio data files were assessed to provide insights into how crews worked through the crew gunnery-engagement process known as DIDEA (detect, identify, decide, engage, assess). Using this process, the crew must rapidly acquire targets, identify them as potential threats, make a decision to engage or not engage a target, engage the target(s), and then assess the effects of each firing action (Schaefer et al. 2019b). Therefore, audio recording from each dyad's performance was used to identify times when crews a) first noticed targets, b) first acquired a target, c) first provided a fire command, d) first fired, and e) first registered a hit on a target. No participants opted out from audio data collection.

2.10 Physiological Metrics

The Empatica E4 uses a PPG sensor, which implements an optical measurement technique, to detect BVPs. While PPG-based, wrist-worn wearable sensors provide useful information related to the cardiovascular system, research has shown that its data are less reliable and more susceptible to noise when the wrist is moving. As a result, data reported here should be taken as more of a proof-of-concept and exploratory analysis on the use of wearables for objectively and unobtrusively tracking individual's physiological activity that could dynamically reveal insights into their physiological conditions during task performance. In processing the inter-beat-interval (IBI) and HR measures produced by the Empatica software package, we discovered that the two time-series data were incorrectly derived and the quality of these data was such that they could not be used for our analyses (see Appendix B for details). Therefore, we obtained IBI and HR metrics from the BVP signal instead with our processing pipelines. First, the 64-Hz sampled raw BVP signals were band-pass filtered between 0.8 Hz and 2.5 Hz, the frequency range that respectively corresponds to 48 and 150 beats per minute. The BioSPPy biosignal-processing package (in Python) was then used to process the filtered BVP signals to obtain IBI time-series estimates. Subsequently, IBI values considered

abnormal/artifacts (any IBI values that are more or less than 30% of its neighboring IBI values) and outliers (IBI values that are not within 3 standard deviations (SDs) of the entire IBI distribution) were removed before further analysis. HR data were calculated from the IBI time series (HR = 60000/IBI), and the time-domain HRV, quantified here using Root Mean Square of Successive Differences (RMSSD), was calculated using IBI values over time periods of resting, training, Exercise 1, and Exercise 2. Averaged HR and EDA measures with confidence intervals (CIs) were also calculated. Each individual HR, HRV, and EDA data were normalized by their resting measures before combining measures across all dyads (i.e., grouping lethality operators or robotic mobility operators).

2.11 Procedure

Upon the arrival of the participants, researchers greeted the participants and thanked them for their time and participation. Each participant randomly sat in either chair before them, which determined which role they would take. They were asked to fully read and fill out the two copies of consent forms in front of them. After researchers collected signed consent forms, further instructions were provided. Each participant was fitted with an Empatica E4 wrist sensor and if they agreed to being recorded visually and auditorily, Logitech cameras were turned on and put on record. Researchers explained to the participants that they were to use the Wingman protocol interfaces to practice and then complete two exercises with five target engagements each on a simulated gunnery course. The participants had a 5-min training session with the researchers to go over fundamental controls and job specifications.* The robotic lethality operator and mobility operator had a practice exercise where they could go through their tasks without time restrictions.

Each exercise started with one researcher giving a quick scenario and the participants verbalizing readiness to their confederate commander. Before beginning an iteration, both participants would reply with "REDCON 1" (Readiness Condition 1) to notify their confederate commander that they were ready to begin the iteration. Each exercise required the vehicle operator to maneuver the vehicle to specific battle positions. When targets eventually came up, the vehicle operator immediately stopped the vehicle and the lethality operator began to identify and shoot down targets. When the lethality operator had identified and lazed a target, they would notify the commander with "Target identified." Once the confederate commander verified, he would give the command to "Fire" and the lethality operator responded with "On the way" as he squeezed the trigger. The commander could see the results of each action and either say "Target, target

* Traditional gunnery training requires 6 months.

destroyed" or make corrections for the lethality operator to adjust their aim. The first exercise allotted 100 s of exposure time for the targets, while the second exercise gave only 50 s to engage and destroy the targets.

Through the different engagements, the researchers and cadet research assistants collected performance, physiological, auditory, visual, behavioral, and subjective data. The participants were graded based on qualification standards on a Table VI gunnery exercise. Subjective data were collected in the form of questionnaires that participants filled out on the completion of each exercise. Each participant completed three sets of questionnaires (after the training and each exercise) covering topics of trust in the Robotic Combat Vehicle (RCV), team readiness, stress, and workload. Upon the completion, the cadet dyads were released and thanked for their participation.

2.12 Design

This was a within-subjects design where dyads (+ confederate commander) completed the practice session (unlimited exposure time), Simulated Gunnery Exercise 1 (exposure time = 100 s), and Simulated Gunnery Exercise 2 (exposure time = 50 s). The order of conditions was fixed; however, the virtual test courses were matched for difficulty and counterbalanced prior to arrival in the study. Participants maintained their role throughout the entire experiment.

3. Results

These results provide an exploratory analysis of team trust for human–autonomy lethality teams. Analyses build on performance-based metrics for gunnery operations to explain team member perceptions of stress, workload, and trust as well as potential objective indicators of trust from communication, behavior, and physiological cues.

3.1 Performance

Performance scores included analysis from the standard scoring using the Army DA Form 8265, the Common Crew Score Sheet. Gunnery performance was measured as a crew and reported in Table 3 (see also Appendix A). Even though the length of the exercises was different, the standard for qualification remained the same.

Table 3 Qualification scores

Dyad	Exercise 1			Exercise 2		
	No. qualified	Total score	Avg Score	No. qualified	Total score	Avg score
2	3/5	272	54.4	3/5	340	68
3	2/5	265	53	5/5	478	95.6
4	2/5	309	61.8	3/5	314	62.8
5	3/5	384	76.8	3/5	277	55.4
6	3/5	393	78.6	0/5	191	38.2
7	3/5	316	63.2	3/5	420	84
8	1/5	200	40	3/5	324	64.8
9	1/5	236	47.2	2/5	284	56.8
11	2/5	305	61	3/5	385	79
12	3/5	398	79.6	2/5	338	67.6
13	3/5	347	69.4	2/5	281	56.2
14	1/5	233	52.4	2/5	262	46.6

Note: Performance scores were calculated using Army guidelines in Training Circular 3-20.31. Each engagement had a maximum score of 100 points, where 70 are considered a qualifying score.

A two-way analysis of variance (ANOVA) was conducted on the performance scores for exercise (Exercise 1 vs. Exercise 2) and posture (offensive vs. defensive). There was no significant difference in performance between Exercise 1 (M = 61.45, SD = 34.597) and Exercise 2 (M = 64.58, SD = 33.066), p = 0.613. There was a significant difference in posture, $F(1, 116)$ =10.70, p = 0.001, where performance scores for defensive operations (M = 70.72, SD = 3.78) were significantly higher than offensive operation (M = 51.46, SD = 4.78). This is in line with standard gunnery findings where engagements in a defensive position were easier to qualify since timing for offense began once targets were locked in position. A significant interaction, $F(1, 116) = 8.58$, p = 0.004, showed that the difference between postures was only significant for Exercise 1 when participants had more time. Manned-platform crews are given a standard of 6 months to train on their system as a crew prior to qualification. This enables the members to gain trust in their crew and their weapon system. These findings suggest the cadets were able to show these same patterns with only two exercises totaling 55 min.

3.2 Self-Reported Ratings

Self-reported ratings of trust, stress, and workload can provide critical insights into performance-based scores.

3.2.1 Self-Reported Differences between Exercises 1 and 2

Paired-samples *t*-tests were conducted to assess differences in subjective responses on trust, stress, and workload between Exercises 1 and 2. The only significant difference was in workload, $t(23) = 2.30$, $p = 0.030$, where participants experienced higher mental demand ($M = 43.96$, SD = 19.166) in Exercise 1 than Exercise 2 ($M = 35.42$, SD = 22.745). This result reinforces the performance findings. In particular, the workload score reinforces the effects of the minimal training time on the first exercise. The significant drop in mental demand suggests the technology and team dynamics for manned–unmanned gunnery operations is transparent and successful with minimal exposure to the task and autonomy.

3.2.2 Self-Reported Differences between Lethality and Mobility Operators

Paired-samples *t*-tests were conducted to assess the impact of the operator role (lethality or mobility) on self-reported ratings of trust, stress, and workload. When analyzing stress, we found that positive affect for the mobility operator was significantly higher, $t(39.79) = -2.40$, $p = 0.021$, whereas the lethality operator scored significantly higher on sensation seeking, $t(44.39) = 2.08$, $p = 0.044$. With respect to the DYS scale—a composite of anxiety, depression, and hostility scales—lethality operators scored significantly higher, $t(34.82) = 2.35$, $p = 0.025$, which may suggest the presence of emotional distress (Fig. 7)

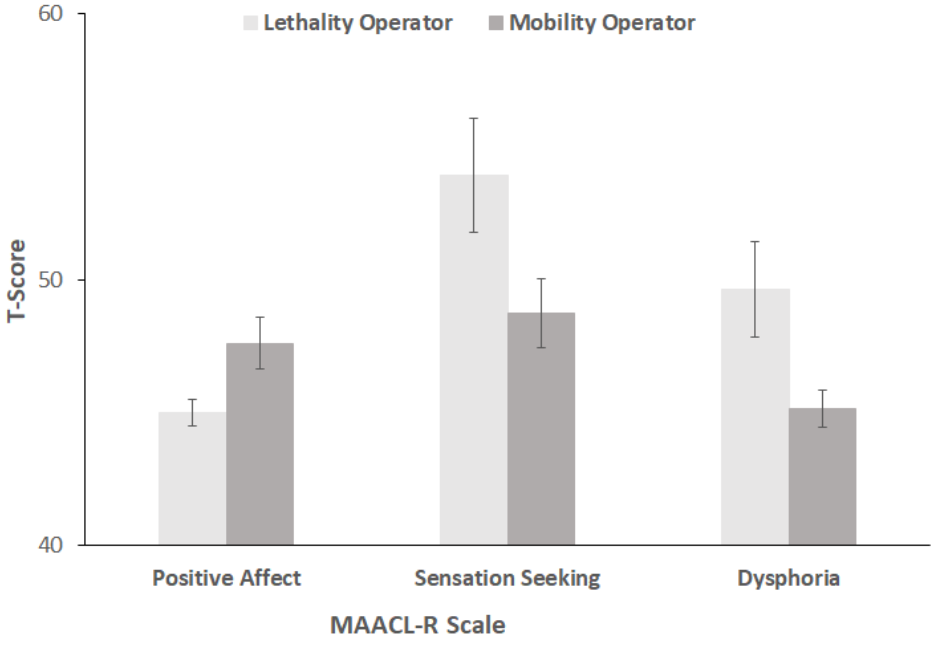

Fig. 7 Stress by role for positive affect, sensations seeking, and dysphoria

This matched with findings from the workload assessment, whereby the lethality operator had higher workload than the mobility operator—a significant main effect of role on total workload, $F(1, 44) = 19.75$, $p < 0.001$; mental demand, $F(1, 44) = 6.61$, $p = 0.014$; temporal demand, $F(1, 44) = 14.46$, $p < 0.001$; frustration, $F(1, 44) = 5.51$, $p = 0.24$; and effort, $F(1, 44) = 15.38$, $p < 0.001$.

3.2.3 Team Readiness

To understand the level of cohesion between the operators, a Pearson correlation analysis was conducted for the Team Readiness questionnaire (TRQ) items—assessing readiness, confidence, and trust when interacting with autonomy—for the different operator roles (Table 4) of the RCV. Positive correlations among team members suggest cohesion among the team members.

Table 4 Correlations among items on the TRQ

Role	TRQ item	Operate RCV	Operate weapon	Self-confidence	Team confidence	Confidence weapon	Trust weapon	Trust mobility	Team trust
Lethality operator	Operate RCV	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
	Operate weapon	0.872[a]	. . .	. . .	. . .	. . .	. . .	. . .	. . .
	Self-confidence	0.732[a]	0.837[a]	. . .	. . .	. . .	. . .	. . .	. . .
	Team confidence	0.786[a]	0.814[a]	0.863[a]	. . .	. . .	. . .	. . .	. . .
	Confidence weapon	0.721[a]	0.836[a]	0.933[a]	0.896[a]	. . .	. . .	. . .	. . .
	Trust weapon	0.418[a]	0.455[a]	0.496[a]	0.570[a]	0.572[a]	. . .	. . .	. . .
	Trust mobility	0.462[a]	0.280	0.442[a]	0.580[a]	0.502[a]	0.528[a]	. . .	. . .
	Team trust	0.662[a]	0.614[a]	0.766[a]	0.833[a]	0.807[a]	0.554[a]	0.553[a]	. . .
Mobility operator	Operate RCV	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
	Operate weapon	0.652[a]	. . .	. . .	. . .	. . .	. . .	. . .	. . .
	Self-confidence	0.646[a]	0.569[a]	. . .	. . .	. . .	. . .	. . .	. . .
	Team confidence	0.624[a]	0.585[a]	0.857[a]	. . .	. . .	. . .	. . .	. . .
	Confidence weapon	0.718[a]	0.889[a]	0.627[a]	0.669[a]	. . .	. . .	. . .	. . .
	Trust weapon	0.274	0.414[a]	0.200	0.203	0.431[a]	. . .	. . .	. . .
	Trust mobility	0.349[b]	0.365[b]	0.365[b]	0.458[a]	0.412[a]	0.438[a]	. . .	. . .
	Team trust	0.494[a]	0.661[a]	0.666[a]	0.671[a]	0.661[a]	0.424[a]	0.471[a]	. . .

[a] Correlation is significant at the 0.01 level (2-tailed)

[b] Correlation is significant at the 0.05 level (2-tailed)

Paired-samples *t*-tests were conducted to determine if there were any differences in lethality and mobility operators' ratings of readiness, confidence, and trust when using the mobility autonomy and weapon autonomy. For the mobility autonomy, results showed no significant differences for team readiness ($p = 0.512$), confidence ($p = 0.765$), or trust ($p = 0.365$). Similar results were found for the weapon autonomy: no significant differences in ratings for readiness ($p = 0.903$), team confidence ($p = 0.894$), or trust ($p = 0.636$). Further, mean ratings for readiness, confidence, and trust for the mobility and weapon autonomy ranged from 5.3 to 5.8, with the exception of trust ratings for the weapon-system autonomy, which was 4.42 (SD = 1.61) for the lethality operator and 4.63 (SD = 1.64) for the mobility operator, indicating a relatively high level of readiness, confidence, and trust. Taken together, the lack of significant differences, coupled with relatively high mean ratings on the TRQ items, is promising and indicates that both operators were able to perform their tasks and developed confidence and trust in the mobility and weapon autonomy. This is critical because the operators only had a short time interacting with the systems (e.g., a total of 55 min, with 5 min of training) and no prior exposure, further reinforcing the performance findings and providing some initial insights for integrating autonomous assets into military teams.

Since confidence can directly impact trust and readiness to team with autonomy, a 2 × 3 repeated-measures ANOVA was conducted to understand if there were any significant effects of Role (lethality vs. mobility) and Condition (practice, Exercise 1 vs. Exercise 2) on ratings of self-confidence and team confidence with the mobility autonomy and weapon autonomy. For ratings of self-confidence with the mobility and weapon autonomy, there was only a main effect of condition, $F(2, 20) = 16.76$, $p < 0.001$. Pairwise comparisons showed significantly lower self-confidence ratings for Practice compared with Exercise 1 (mean difference = –1.36, $p < 0.001$) and Exercise 2 (mean difference = –1.64, $p < 0.001$); however, self-confidence ratings did not significantly differ between Exercise 1 and 2 ($p = 0.480$). These findings indicate self-confidence with the mobility- and weapon-system autonomy increased with usage, which is an important finding when teaming humans and autonomous systems.

For ratings of team confidence with the mobility autonomy, there was only a significant main effect of condition, $F(2, 22) = 11.74$, $p < 0.001$. Pairwise comparisons showed significantly lower team-confidence ratings for Practice compared with Exercise 1 (mean difference = –0.917, $p = 0.010$) and Exercise 2 (mean difference = –1.42, $p = 0.003$), and significantly lower team-confidence ratings for Exercise 1 compared with Exercise 2 (mean difference = –0.500, $p = 0.026$). Similarly, ratings of team confidence when using the weapon system

differed by Condition, $F(2, 22) = 5.81$, $p = 0.002$. Pairwise comparisons showed significantly lower-team confidence ratings for Practice than Exercise 1 (mean difference = –0.792, $p = 0.020$), and Exercise 2 (mean difference = –0.917, $p = 0.023$), but no differences between Exercise 1 and 2 ($p = 0.586$). Lastly, a paired-samples t-test was performed to compare team confidence operating the RCV and team confidence operating the weapon. Results showed a significant difference in team confidence between the two systems, $t(47) = 2.86$, $p = 0.006$, with significantly higher team-confidence ratings for the mobility autonomy (M = 5.63, SD = 1.00) than weapon autonomy (M = 5.40, SD = 1.14). These findings suggest that team confidence also increased with system usage for the mobility autonomy, but team members were able to reach a stable level of team confidence with the weapon system more quickly than with the mobility autonomy. Considering the total time interacting with the team was 55 min, these results are noteworthy and should be investigated further with a larger sample size to confirm these findings.

3.3 Detect, Identify, Decide, Engage, Assess (DIDEA) Communication Analysis

Audio recordings were assessed to provide insights into how crews worked through the DIDEA process. The audio was used to identify times when crews a) first detected targets, b) first acquired a target, c) first provided a fire command, d) first fired, and e) first registered a hit on a target. These times were used to determine how long it took the crews to go from first noticing a target to firing on it, which is an indicator for how well they performed crew gunnery and how quickly they moved through the DIDEA process.

Figure 8 represents the average time between when a target was first detected and when the target was either first fired upon (blue dots) or hit (orange dots). The data are averaged across all crews and represent the time-to-fire for each engagement. The data are trending downward, indicating crews reduced their time-to-fire as they completed more engagements and providing evidence for a training effect over time. This trend also supports the finding that participants experienced higher mental demand in Exercise 1 than Exercise 2, indicating participants became more comfortable with the gunnery process, further reinforcing the finding from both the performance and self-reported analyses.

Note the decrease in the difference between fire time and hit time as crews proceed through the engagements. A larger difference between the “fire time” (blue dots) and “hit time” (orange dots) indicates the shots likely missed and the crew had to readjust their aim in order to successfully hit the target. A smaller difference indicates the crew was able to hit the target in fewer attempts. Therefore, the

decrease in the difference between "fire times" and "hit times" between the first few engagements and the last few engagements suggests that crews became more accurate with their shots and efficient at the gunnery process.

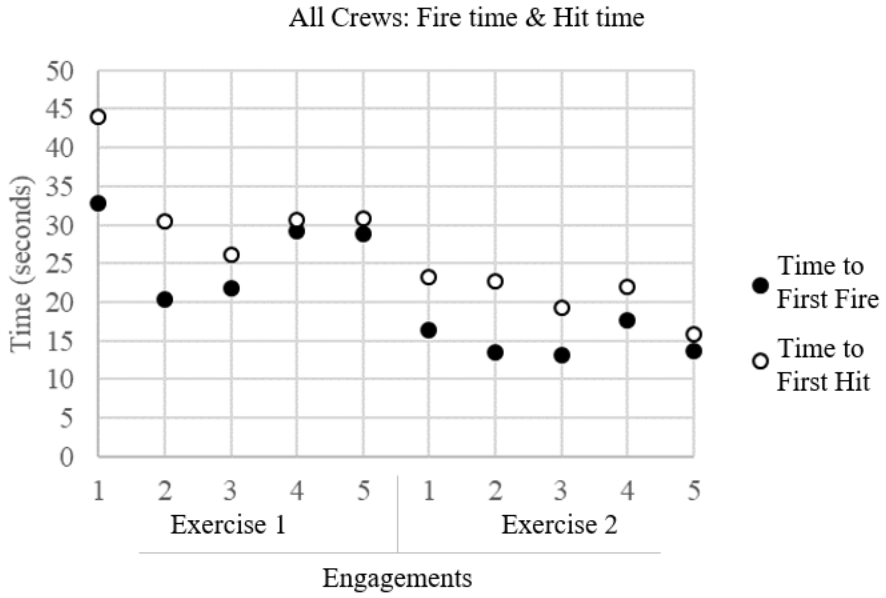

Fig. 8 Average times for all crews to fire on (black) and hit (white) targets; all 10 target engagements are in order of exposure, 5 engagements in Exercise 1, and 5 engagements in Exercise 2

There was a significant negative correlation between crews' average hit time in Exercise 2 and their self-reported trust in the weapon autonomy (r = –0.627, p= 0.060, n = 12). In other words, for Exercise 2, as crews reduced their hit times (thus hitting targets more quickly), their trust in the weapon autonomy increased. This suggests that a primary driver of crews' trust in the weapon autonomy is their ability to use it effectively to complete the mission. This also supports the earlier finding that crews reported a higher mental demand in Exercise 1 than in Exercise 2, suggesting these trends reveal crews' improvement in gunnery and in the DIDEA process itself.

CI analysis (Table 5) suggests no significant differences in time-to-first-fire or time-to-first-hit between engagements in Exercise 1, and no significant differences in engagements in Exercise 2 (Fig. 9). However, CI analysis suggests the time-to-first-fire was significantly longer for Exercise 1 than Exercise 2 for Engagement 1 and nearing significance for Engagement 4. For time-to-first-hit, Exercise 1 was significantly longer for Engagement 1 (Fig. 10). The other point of note is that the variability in teams is much smaller for Exercise 2 than Exercise 1.

Table 5 Descriptive statistics for average crew time-to-first-fire and time-to-first-hit

Crew time	Exercise	Engagement	M	SD	CI low	CI high
First fire	1	1	37.75	13.03	30.38	45.12
		2	25.42	26.40	10.48	40.35
		3	26.83	15.51	18.06	35.61
		4	34.17	15.97	25.13	43.20
		5	33.92	33.05	15.22	52.62
	2	1	21.42	6.07	17.98	24.85
		2	18.58	14.64	10.30	26.87
		3	18.17	8.35	13.44	22.89
		4	22.67	11.49	16.16	29.17
		5	18.67	12.97	11.33	26.01
First hit	1	1	48.92	22.78	36.03	61.81
		2	35.50	25.27	21.20	49.80
		3	31.17	20.19	19.74	42.59
		4	35.67	15.93	26.65	44.68
		5	35.75	32.13	17.57	53.93
	2	1	28.17	9.93	22.55	33.78
		2	27.67	15.48	18.91	36.43
		3	24.33	12.82	17.08	31.58
		4	27.00	11.88	20.28	33.72
		5	20.92	13.13	13.49	28.35

Note: Time in seconds. M = mean.

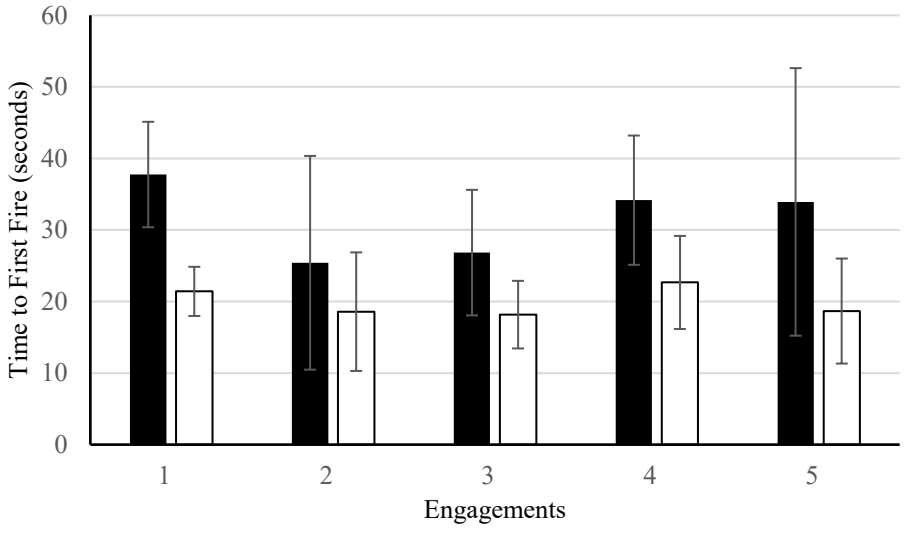

Fig. 9 CI analysis of time-to-first-fire; black bar is Exercise 1, white bar is Exercise 2

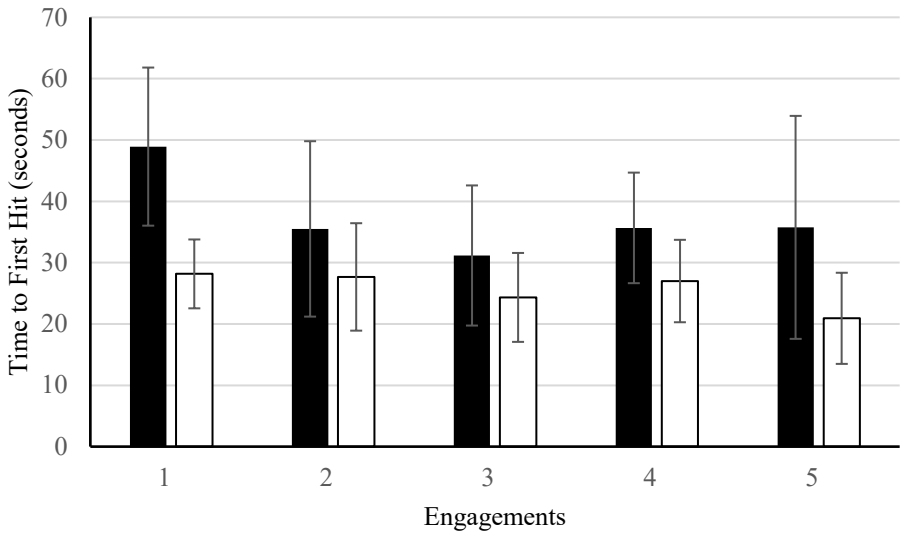

Fig. 10 CI analysis of time-to-first-hit; black bar is Exercise 1, white bar is Exercise 2

3.4 Behavioral Indicators

Performance scores alone may not be enough to understand the crew's interactions during the task. While subjective assessments provide inferences about what a person is consciously aware of, behavioral responses can be indicative of subconscious indicators of trust. Here, facial expressions are assessed as a metric of a person's affective experiences in real time, which can build on and extend beyond the retrospective self-reported response of emotional perceptions.

3.4.1 Facial Expressivity

This analysis focused on analyzing mean facial expressivity as a function of exercise (i.e., Exercise 1 vs. Exercise 2) and role (i.e., mobility vs. lethality operator) to determine if there were affective differences between the exercises or according to the role the participant was engaged in. The facial expressions analyzed reflect changes in the seven universal emotions (i.e., happiness, sadness, surprise, fear, anger, disgust, and contempt). A 2 (exercise) × 7 (emotion) repeated-measures ANOVA found no significant effect of exercise, $F(1, 22) = 2.46$, $p = 0.131$, but inspection of the means indicated the evidence for emotional expression did increase somewhat from Exercise 1 to Exercise 2 for all emotions, except for the emotion of "Surprise" (Fig. 11). Additionally, the main effect of emotion was significant, $F(6, 17) = 16.65$, $p < 0.001$, indicating the evidence values for the seven universal emotions significantly differed from one another and the time × emotion interaction was significant, $F(6, 17) = 3.08$, $p = 0.031$.

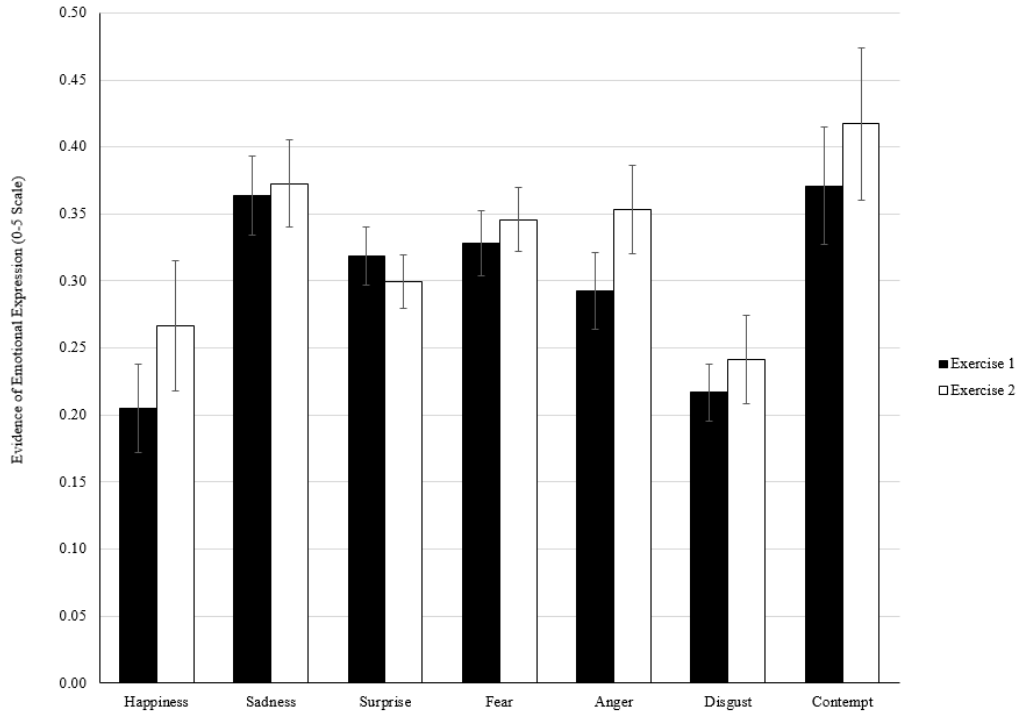

Fig. 11 Facial-expression evidence for seven universal emotions as function of exercise: means are somewhat higher in Exercise 2 than Exercise 1 (except for "Surprise"); error bars are standard errors

Additionally, a series of *t*-tests were performed to determine if there was a significant difference in the means of the facial expressivity according to role. From Fig. 12, we see that evidence for facial expressions was somewhat higher for the lethality operator (e.g., this operator was more outwardly expressive) for all emotions except "Surprise" and "Disgust".

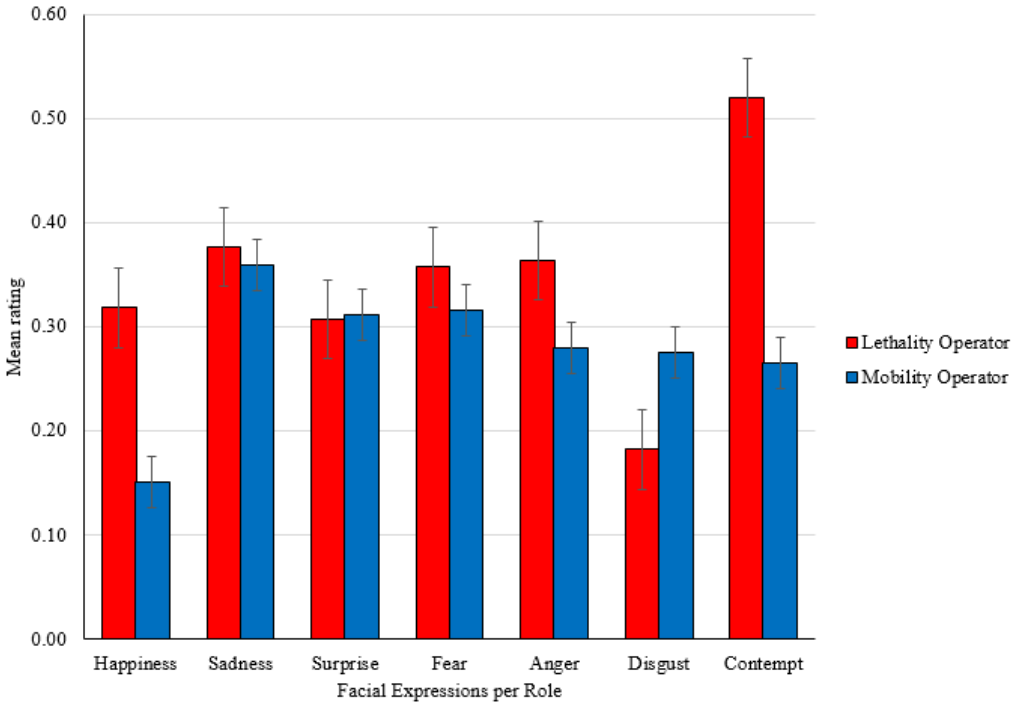

Fig. 12 Facial-expression evidence for seven universal emotions as function of role: means for all emotions are somewhat higher for lethality operator (except "Surprise" and "Disgust"); error bars are standard errors

More specifically, when analyzing the difference in means we found the following emotions were significantly different from one another: For both Exercises 1 and 2, "Happiness" and "Contempt" were significantly higher for the lethality operator than the mobility operator (see Table 6). From this, we can infer the significantly higher happiness and contempt emotions could be due to the level of responsibility and level of difficulty required of the lethality operator.

Table 6 "Happiness" and "Contempt" findings

	Lethality operator		Mobility operator			
Emotion	**M**	**SD**	**M**	**SD**	***t***	***p-value***
Exercise 1 Happiness	0.27	0.19	0.14	0.11	2.19	0.039
Exercise 1 Contempt	0.48	0.20	0.26	0.17	3.14	0.004
Exercise 2 Happiness	0.37	0.29	0.16	0.10	2.35	0.029
Exercise 2 Contempt	0.56	0.30	0.23	0.22	2.99	0.007

3.4.2 Correlates of Facial Expressivity

Additionally, bivariate correlations among 1) the seven facial expressions, 2) the seven MAACL dimensions, and 3) the physiological variables mentioned previously were performed to determine if any directional relationships existed among the variables, according to participant role, across both exercises.

For the lethality operator, no relationships were found during Exercise 1; however, there were some significant relationships among the facial expressions and previously mentioned variables during Exercise 2 (Table 7). From this, it appears positive relationships exist among the emotional expression "Contempt", the MAACL dimension "Sensation seeking", and the combined "PASS" dimension. The negative relationship between the anxiety dimension on the MAACL and the facial-expression data indicates higher nervousness was not outwardly expressed as disgust for the lethality operator.

Table 7 Correlations among facial expression and MAACL variables for lethality operator, Exercise 2

MAACL self-reported ratings	Facial data	
	Disgust	Contempt
Anxiety	–0.688[a]	. . .
Sensation seeking	. . .	0.666[a]
PASS	. . .	0.591[a]

[a] Correlation is significant at the 0.05 level

Note: No relationships were found in Exercise 1 for the lethality operator

For the mobility operator there were several significant relationships among the facial expressions and other variables in both Exercise 1 and 2 (Table 8). Negative relationships between the MAACL dimension "Sensation seeking" and "Happiness" were found to exist. Additionally, higher depression ratings were also related to more outward expressions of "Contempt". Finally, positive relationships with the RMSSD measure (both resting baseline levels and Exercise 1) exist for the expression "Surprise".

Table 8 Correlations among facial expression, MAACL, and physiological variables for mobility operator, Exercises 1 and 2

	Exercise 1			Exercise 2	
	Contempt	**Happiness**	**Surprise**	**Surprise**	**Disgust**
Depression	0.586[a]	…	…	…	…
Sensation seeking	…	–0.629 [a]	…	…	…
Positive affect	…	…	…	…	0.608 [a]
PASS	…	…	…	…	0.604 [a]
Resting RMSSD	…	…	0.591[a]	…	…
RMSSD	…	…	0.557[a]	0.578[a]	…

[a] Correlation is significant at the 0.05 level

In Exercise 2, significant relationships were found among the emotion "Disgust" and "Positive affect" and the combined "PASS" dimension. This may be indicative of frustration at having to participate in the monotonous nature of the task for this particular role. Finally, RMSSD and "surprise" were positively correlated.

3.5 Physiological Indicators

Results from our physiological analyses suggest that, relative to their resting baseline measures, both the mobility and lethality operators' averaged HR, EDA, and HRV increased during training and while they engaged targets in both exercises (see Figs. 13–15). These results are consistent with the idea that as the participants performed the tasks in those periods, they were more engaged and alert. Increases in EDA for both roles indicated the dyad experienced higher arousal and increased cognitive workload. While the levels of EDA increases were similar in both exercises for the two roles, the mobility operator had significantly higher EDA during training (Fig. 13). Coupled with the mobility operator having higher HR than the lethality operator in all conditions (Fig. 14), these results suggest role differences in that the operator of the vehicle was more alert and cognitively burdened as (s)he constantly provided safe and secure mobility operation. Parallel increases over time in vagally mediated HRV (noninvasive measure of parasympathetic activity) for both roles (Fig. 15) suggest each dyad became more familiar with each other. Since research in animal models and humans suggest the parasympathetic nervous system is particularly important for the expression of social emotions and affiliative behaviors (Porges 2007), these findings suggest there may be some indication of positive social emotions and affiliative behaviors.

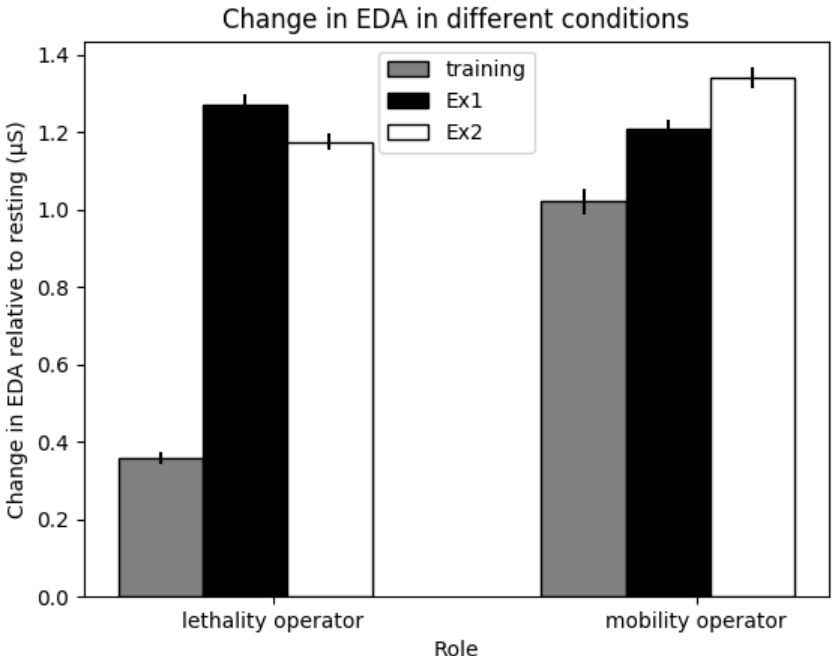

Fig. 13 Change in lethality and mobility operators' EDA during training and Exercises 1 and 2 shown as averages with CIs

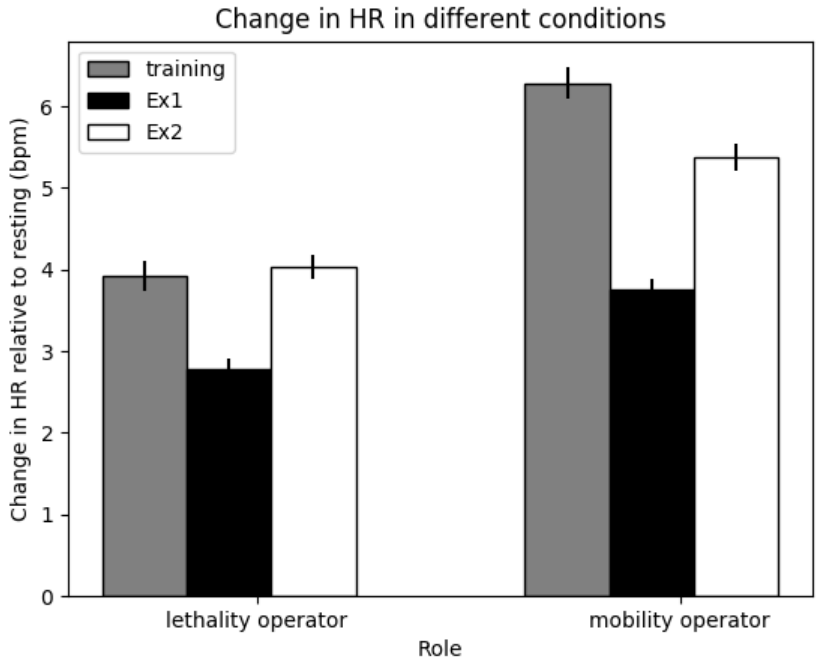

Fig. 14 Change in lethality and mobility operators' HR during training and Exercises 1 and 2 shown as averages with CIs

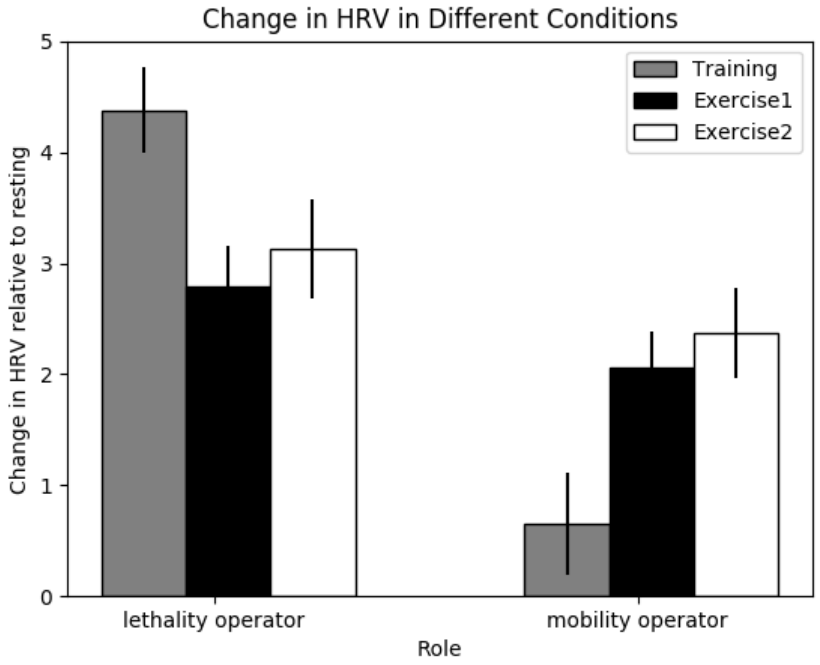

Fig. 15 Change in lethality and mobility operators' HRV during training and Exercises 1 and 2 shown as averages with CIs

These physiological results were also consistent with other measures of team trust and cohesion, in that both lethality and mobility operators appeared to engage and stay alert (evidenced by increased HR and EDA) while performing the tasks and appeared to be relaxed (evidenced by decreased HRV) when between high-risk tasks, as evidenced in both exercises. While additional data and analysis are required to validate the use of physiological signals to infer team trust and cohesion, these results demonstrate its importance and application, as continuous tracking of these physiological signals could provide real-time estimates of team trust and cohesion that could be positively impacted with various interventions.

3.6 Time Series Classification (TSC)

The goal here was to build an algorithm to determine whether participants were engaged with a target on the basis of physiological signals measured from the wrist-worn Empatica. In principle, this kind of algorithm could provide commanders with a binary outcome reflecting crew state that would be passively available at all times. Therefore, this goal was conceptualized as one of TSC; that is, given a short time series of physiological data measured from a participant, determine whether it was taken from a time when the participant was engaging with a target or not. To do so, data from seven signals (x, y, and z accelerometer signals, EDA, skin temperature, BVP, and IBI) over 14 sessions were divided into 3-s time series "epochs" to build

a model that maximally distinguished target-engaged and target-not-engaged epochs from one another.

Using TSC techniques is an active area of work in machine learning; however, the best methods tend to be very time-consuming to train and tend not to scale to larger data sets (e.g., Fawaz et al. 2020). Recent work focusing on methods from the field of deep learning has identified several approaches that are equal to or better than these methods while being far less costly to implement. For example, prior research has conducted systematic studies of TSC deep-learning architectures that use the UCR/UEA archive (the largest time-series data repository; Chen et al. 2015, Bagnall et al. 2017) as a benchmark (Fawaz et al. 2019, 2020; Dempster et al. 2020).

Therefore, in exploratory analyses we adopted and tuned the most promising of these methods to this specific data set. Initial findings clearly indicated that ROCKET (**R**and**o**m **C**onvolutional **Ke**rnel **T**ransform; Dempster et al. 2020) performed best compared to other state-of-the-art methods. However, ROCKET is applied specifically to univariate time series, so there was an option of applying this method to each physiological time series individually or creating an ensemble of ROCKETs to be applied to each time series simultaneously. In the former approach, ROCKET is applied exactly as it was originally conceived to each of the seven available signals separately. In the latter, the outputs of the individual ROCKET models are used as inputs to a single standard feedforward network trained to do the same classification task; that is, the feedforward network should learn an optimal mix of the original classifiers. The ensembling network architecture was designed using a Bayesian optimization algorithm implemented in KerasTuner (O'Malley et al. 2020). Furthermore, the models were trained on data sets either split by role (i.e., lethality operator vs. mobility operator) or not, as the prior analyses suggested it may be that different roles show different physiological profiles in the presence of a target. Thus, models were trained on four possible subsets of the data that followed from the four unique combinations of subsetting by single-signal versus ensemble and split-by-role versus roles-combined.

A leave-one-session-out cross-validation approach was used whereby each session served as the test data while the remaining sessions served as the training data. As a result, these models are assessed in terms of an average of balanced accuracy over 14 test cases (Table 9). The best-performing model was created by subsetting on both participant roles and making use of only a single physiological signal. Specifically, a model trained on only z-axis accelerometer data from lethality operator outperforms all other options.

Table 9 Balanced accuracy, averaged over 14 sessions (and SDs) for each model trained on each unique combination of role and signal subsetting

		Signals	
		Single M (SD)	**Ensemble M (SD)**
Roles	**Split**	0.55 (0.04)	0.54 (0.06)
		0.63 (0.06)	0.51 (0.02)
	Combo	0.54 (0.02)	0.53 (0.03)

Note: In the case of role splits, the mobility operator is given first and then the lethality operator. In the case of the single signals, only results from the model trained on the best signal are shown. The best performance result is bolded.

4. Discussion

Identifying possible trust-related metrics for effective human–autonomy team performance is difficult and complex. Performance scores alone do not provide any information about trust or the cohesion of the team. Even the traditional, self-reported rating only provides a conscious state response to trust that is often confounded by the task and crew and suffers from reporting bias. This work provides further support that a multimethod approach is critical for quantifying team-trust and team-cohesion analyses. These results bring us one step closer to identifying key metrics for supporting human–autonomy team effectiveness.

4.1 Review of Main Findings

Traditional gunnery evaluations use the Common Crew Scoresheet to determine if a crew can qualify on a Table VI gunnery exercise. While this is traditionally used for human-only teams, this performance metric has more recently been applied to manned–unmanned lethality teams (Schaefer et al. 2019a; Baker et al. 2020). However, for this simulation study these metrics showed no difference between Exercise 1, where manned–unmanned lethality teams had twice as long to fire on target than normal, and Exercise 2, where the teams had the 50 s that is traditionally allowed. This finding contradicted expectations as Exercise 1 was twice as long, which should have led to better performance.

However, prior research has suggested that in quantifying team trust and team cohesion for these manned–unmanned lethality teams, performance metrics alone do not tell the whole story. Specifically, a multimethod analytical approach is needed to understand the full implications on effective performance (Schaefer et al. 2019a; Milner et al. Forthcoming 2021). Taking this approach, the data showed it is possible to glean information from very different measurement approaches to

understand the implications of role dynamics as they affect team trust, cohesion, and subsequently performance.

Specifically, self-reported assessment occurs after the fact and takes into account a person's conscious reflection about a task. These self-reported findings suggested higher mental demand for Exercise 1. While on the surface this could be a reflection of the limitations of training time, the significant drop in mental demand between the two exercises provides promising support in the current design of the autonomy and transparency of the WMI for successful team operations. Since the Wingman simulation testbed is a software-in-the-loop version of a real prototype system, this finding provides valuable feedback to the engineering team related to the current and future transparency of the system.

Second, self-reported assessments brought to light critical similarities and differences between the lethality and mobility operators. An important finding was the presence of emotional distress and higher workload by the lethality operators were indicative of the differences in role responsibilities of the given task. However, the absence of significant differences between operators for team readiness, confidence, and trust in autonomy, coupled with relatively high mean ratings of these items, is promising and indicates both operators were able to perform their tasks and developed confidence and trust in the mobility and weapon autonomy. This is critical because the operators only had a short time interacting with the systems (e.g., a total of 55 min, with 5 min of training) and no prior exposure. This further reinforces the performance findings and provides some initial insights for integrating autonomous assets into military teams.

Objective assessment of team communication with behavioral and physiological data can provide indicators of team trust and cohesion that occur during the interactions. Specifically with this data set, communication analysis demonstrated that crews reduced their time-to-fire as they completed more engagements. This is in line with the self-reported results of higher mental demand in Exercise 1 than Exercise 2 and indicates participants became more comfortable with the gunnery process. Further, crews became more accurate with their shots and efficient at the gunnery process. As their gunnery performance improved, they reported higher trust in the weapon autonomy, suggesting that a key to crews' trust in the weapon autonomy is their ability to use it effectively to complete the mission. This links back to the gunnery score we discussed earlier by providing deeper context for why teams performed the way that they did.

The facial and physiological analyses add additional layers on top of this to help explain the crew members' physical and affective responses to what was happening on the task. Overall, the facial-expression analyses and physiological results are

somewhat corroborative. Here, HRV values increased overall, which may indicate that participants were able to better regulate their emotions and, thus, exhibit outwardly low facial expressions. HRV was somewhat higher for the lethality operator and somewhat lower for the mobility operator, further supporting previous self-reported findings related the divergence in the amount of effort for each operator. That said, one operator outwardly expressed high levels of emotional expression (i.e., facial expression values were comparatively low in this data set on a scale of 0–5). This may be due to the nature of the task, which was not particularly emotionally evocative; the task roles (e.g., the mobility operator had a very low task load and therefore may not have responded to task-related demands); or the subject population (e.g., cadets who may possess more self-control and emotional-regulation abilities due to training). Further, the expressions were somewhat higher during Exercise 2, which may reflect the time constraints imposed on the team. These results will be investigated further in more realistic operational environments to better understand the linkages between behavioral and physiological measures.

Together, all of these approaches give us a clearer perspective on the task as a whole and a better understanding of how the crew members functioned and interacted to achieve their goals on the task. In both exercises, with just 5 min of training to explain the features of the WMI and mission operations, teams almost reached qualification scores, and self-reported data related to team cohesion identified that team members were cohesive in their responses. This finding is promising in that the WMI and the capabilities of the autonomy lend themselves to appropriate and effective team dynamics. Additionally, the cohesive nature of ratings on the self-reported measures could also indicate the "adoption" of autonomy as a team member rather than just as a tool to augment team performance. These claims will be verified and validated with larger data sets, including multiple team members both human and autonomous.

4.2 Path Forward

Trust-and-cohesion measurement alone is not enough to fully explain team performance or better facilitate effective team performance. Thus, the first step is to build on these approaches to identify and validate the key metrics needed to properly assess human–autonomy teaming with multiple humans and multiple types of autonomy for a variety of military operations. The second step is to determine how to use this type of data to provide trust-calibration interventions. More specifically, how can these different types of measures provide key markers as to when, how, and what changes should be implemented—from training, to machine-learning behaviors, to transparency displays or communication, to after-

action reviews—to facilitate appropriate team dynamics. The third step is to develop and test the algorithms that classify these team dynamics.

In line with the first two steps, approaches that use communication to understand trust and cohesion are promising; however, the structure of the crew communication from working with a novice and untrained team resulted in a hierarchical, commander-directed task allocation rather than a fully diversified team structure. Therefore, additional research is needed that investigates the relationships between human–autonomy team communication patterns with trust and cohesion. Better communication-based measures of trust and cohesion will allow for a unique window into the experiences of the team and will offer a means to identify when interventions are necessary as a result of communication breakdowns. In addition, future work examining physiological synchrony, behavioral coupling, and their association with self-reported measures is critical to inform near real-time measures of trust and cohesion that will support human–autonomy teams in dynamic, uncertain environments and help inform appropriate and timely interventions to ensure properly calibrated trust within human–autonomy teams. While these methods are currently only exploratory in nature, results described here and in previous work show promise for trust and cohesion measurement over time and will help fill an existing gap in the literature that needs to be addressed.

In line with the third step, additional testing and development of engagement classification is needed to better understand the interplay between autonomy and the human as a sensor within a larger team. For this study, the choice of physiological equipment with the limited training scenario impacted the accuracy of classification approach for engagement classification. It was most likely due to insufficient data (e.g., limits of wrist-worn physiological sensors and limits of time synchronization of this data), insufficient task parameters (e.g., this simulated gunnery training did not elicit strong-enough responses during an engagement vs. baseline), limitations of the classification pipeline, or a combination of those. To address these limitations, near-term plans include 1) training and testing the classifier on upcoming data sets with more expansive physiologic data and different operational task scenarios and 2) evaluating the effects of changes to classifier parameters (e.g., epoch size).

5. Conclusions

Overall, this study advanced general understanding of human–autonomy team trust and cohesion for lethality operations. It first reinforced that the current state of the actual prototype vehicle autonomy and transparency of the Warfighter Machine

Interface were promising for developing rapid team dynamics that resulted in performance scores that were in the range of gunnery-qualification metrics after only 5 min of training. This is a critical finding because traditional gunnery qualification occurs after six months of training. While additional training would ensure crews are well versed in the behaviors of the autonomy, WMI, and controllers, to allow Soldiers to become accustomed to the unmanned asset and more effectively facilitate appropriate trust in the team, the current findings are promising in that the inclusion of transparency-based displays and behaviors may drastically reduce this training time.

The second key finding was that measuring team trust and cohesion for human–autonomy teams is complex. Results from this laboratory study provided some initial insights regarding the utility of using a multimethod measurement approach. Further work is being done to determine how these methods translate to a more operationally relevant setting, adding to the perception of actual risk, given the challenges of data collection during field experiments. Specifically, understanding the interplay among mission context, environmental context, and social context will provide a more complete vision of the reasoning behind team decision-making. Thus, refinement of novel and exploratory measures, such as physiological synchrony and facial analysis, may provide near real-time indicators of trust and cohesion. In addition, communication can give a good window into understanding trust and cohesion in the team as well as providing additional reasoning behind team gunnery performance, leading to more realized implications for expanding these methods to understand these team constructs in larger crews (e.g., larger than dyads, triads).

Finally, the pipelines developed and lessons (as well as limits) learned through the physiological analysis and time-series classification of these data help inform what is currently feasible in these scenarios. This lays the groundwork for further investigation into using continuous physiology-based methods of state estimation to enable intelligent-technology adaptions and interventions. Going forward, we will iterate on these methods, apply them to other data sets, and test the efficacy of adapting systems in response to our real-time state estimations.

6. References

Bagnall A, Lines J, Bostrom A, Large J, Keogh E. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Min Knowl Discov. 2017;31(3):606–660.

Baker AL, Schaefer KE, Hill SG. Teamwork and communication methods and metrics for human–autonomy teaming. Aberdeen Proving Ground (MD): DEVCOM Army Research Laboratory (US); 2019 Oct. Report No.: ARL-TR-8844.

Baker AL, Fitzhugh SM, Forster DE, Brewer RW, Krausman A, Scharine A, Schaefer KE. Team trust in human-autonomy teams: analysis of crew communication during manned-unmanned gunnery operations. Aberdeen Proving Ground (MD); DEVCOM Army Research Laboratory (US); 2020 May. Report No.: ARL-TR-8969.

Baltrusaitis T, Robinson P, Morency, LP. OpenFace: An open source facial behavior analysis toolkit. 2016; p. 1–10. https://www.cl.cam.ac.uk/research/rainbow/projects/openface/wacv2016.pdf.

Batrinca L, Stratou G, Shapiro A, Morency LP, Scherer S. Cicero—towards a multimodal virtual audience platform for public speaking training. International Workshop on Intelligent Virtual Agents; 2013 Aug. Springer, Berlin, Heidelberg. p. 116–128.

Beal DJ, Cohen RR, Burke MJ, McLendon CL. Cohesion and performance in groups: a meta-analytic clarification of construct relations. J Appl Psych. 2003;88(6):989.

Bethel CL, Murphy RR. Non-facial/non-verbal methods of affective expression as applied to robot-assisted victim assessment. Proceedings of the 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI2007); 2007; Arlington, VA. ACM Press. p. 287–294.

Biros DP, Daly M, Gunsch G. The influence of task load and automation trust on deception detection. Group Decis Negot. 2004;13(2):173–189. http://link.springer.com/content/pdf/10.1023%2FB%3AGRUP.0000021840.85686.57.pdf.

Bisk Y, Yuret D, Marcu D. Natural language communication with robots. Paper presented at the Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2016; San Diego, California.

Bitan Y, Meyer J. Self-initiated and respondent actions in a simulated control task. Ergonomics. 2007;50(5):763–788.

Brewer RW, Cerame E, Pursel ER, Schaefer KE. Human–autonomy teaming simulation: Wingman joint capability technology demonstration project closeout documentation. Aberdeen Proving Ground (MD); DEVCOM Army Research Laboratory (US); 2020 Mar. Report No.: ARL-TR-8920.

Chen Y, Keogh E, Hu B, Begum N, Bagnall A, Mueen A, Batista G. The UCR time series classification archive. 2015.

Chen JYC, Procci K, Boyce M, Wright J, Garcia A, Barnes M. Situation awareness-based agent transparency. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2014 Apr. Report No.: ARL-TR-6905.

Chollet M, Wörtwein T, Morency LP, Shapiro A, Scherer S. Exploring feedback strategies to improve public speaking: an interactive virtual audience framework. Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing; 2015 Nov. p. 1143–1154.

Cosenzo KA, Parasuraman R, Novak A, Barnes M. Implementation of automation for control of robotics systems. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2006 May. Report No.: ARL-TR-3808.

Dempster A, Petitjean F, Webb GI. ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels. Data Min Knowl Discov. 2020;34:1454–1495.

Deortentiis P, Summers J, Ammeter A, Douglas C, Ferris G. Cohesion and satisfaction as mediators of the team trust–team effectiveness relationship: an interdependence theory perspective. Career Dev Int. 2013;18:521–543.

DeVault D, Artstein R, Benn G, Dey T, Fast E, Gainer A, Georgila K, Gratch JM, Hartholt A, Lhommet M, et al. SimSensei kiosk: a virtual human interviewer for healthcare decision support. In Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems; 2014 May; Paris, France. p. 1061–1068.

Ekman P, Friesen WV. Manual for the facial action coding system. Consulting Psychologists Press; 1978.

Endsley MR. Toward a theory of situation awareness in dynamic systems. Hum Fact. 1995;37(1):32–64.

Endsley MR, Jones WM. Situation awareness, information dominance and information warfare. Brooks Air Force Base (TX): Air Force Armstrong Laboratory (US); 1997. Technical Report No.: 97-01

Fawaz HI, Forestier G, Weber J, Idoumghar L, Muller PA. Deep learning for time series classification: a review. Data Min Knowl Discov. 2019;33(4):917–963.

Fawaz HI, Lucas B, Forestier G, Pelletier C, Schmidt DF, Weber J, Webb GI, Idoumghar L, Muller PA, Petitjean F. Inceptiontime: finding AlexNet for time series classification. Data Min Knowl Discov. 2020;34:1936–1962.

Figner B, Murphy RO. Using skin conductance in judgment and decision making research. In: Schulte-Mecklenbeck M, Kühberger A, Ranyard R, editors. Society for judgment and decision making series. A handbook of process tracing methods for decision research: a critical review and user's guide. Psychology Press; 2011. p. 163–184.

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.

Gorman JC, Foltz PW, Kiekel PA, Martin MJ, Cooke NJ. Evaluation of latent semantic analysis-based measures of team communications content. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting; 2003.

Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: Hancock PA, Meshkati N, editors. Advances in psychology. North-Holland; 1988. p. 139–183. (Human mental workload; vol. 52). https://doi.org/10.1016/S0166-4115(08)62386-9.

Kozlowski SWJ, Ilgen DR. Enhancing the effectiveness of work groups and teams. Psychol Sci Public Interest. 2006;7(3):77–124. https://doi.org/10.1111/j.1529-1006.2006.00030.x.

Lee JD, See KA.Trust in automation: designing for appropriate reliance. Hum Fact. 2004;46(1):50–80.

Lubin B, Zuckerman M. Manual for the MAACL-R: multiple affect adjective check list–revised. San Diego (CA): Educational and Industrial Testing Service; 1999.

Mach M, Dolan S, Tzafrir SS. The differential effect of team members' trust on team performance: the mediation role of team cohesion. J Occup Organ Psychol. 2010;83:771–794.

Marathe A, Brewer R, Kellihan B, Schaefer KE. Leveraging wearable technologies to improve test & evaluation of human-agent teams, Theor Issues Egron Sci. 2020;21(4):397–417.

Marks MA, Mathieu JE, Zaccaro SJ. A temporally based framework and taxonomy of team processes. Acad Manage Rev. 2001;26(3):356–376.

Mathieu JE, Heffner TS, Goodwin GF, Salas E, Cannon-Bowers JA. The influence of shared mental models on team process and performance. J Appl Psychol. 2000;85(2):273–283.

Matthews R, McDonald NJ, Trejo LJ. Psycho-physiological sensor techniques: an overview. In: Schmorrow DD, editor. Foundations of augmented cognition. Mahwah (NJ): Erlbaum; 2005. p. 263–272.

Mehler B. Establishing the sensitivity of physiological measures of cognitive load in the driving environment. AutoUI Cognitive Load Workshop; 2009; Portsmouth, NH.

Mesmer-Magnus JR, DeChurch LA. Information sharing and team performance: a meta-analysis. J Appl Psych. 2009;94(2): 535–546. https://doi.org/10.1037/a0013773.

Milner A, Seong DH, Brewer RW, Baker AL, Krausman A, Chhan D, Thomson R, Rovira E, Schaefer KE. Identifying new team trust and team cohesion metrics that support future human-autonomy teams. In: Cassenti D, Scataglini S, Rajulu S, Wright J, editors. Advances in simulation and digital human modeling. AHFE 2020. Advances in Intelligent Systems and Computing, vol. 1206. Springer; 2021. p. 86–93. https://doi.org/10.1007/978-3-030-51064-0_12.

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

Montague E, Xu J, Chiou E. Shared experiences of technology and trust: an experimental study of physiological compliance between active and passive users in technology-mediated collaborative encounters. IEEE Trans Hum Mach Syst. 2014;44:614–624.

Mullen B, Cooper C. The relation between group cohesiveness and performance: an integration. Psych Bull. 1994:115:210–227. http://dx.doi.org/10.1037/0033-2909.115.2.210.

Neubauer C, Wooley J, Khooshabeh P, Scherer S. Getting to know you: a multimodal investigation of team behavior and resilience to stress. In Proceedings of the 18th ACM International Conference on Multimodal Interaction; 2016 Oct; New York, NY. Association for Computing Machinery. p. 193–200. https://doi.org/10.1145/2993148.2993195.

Neubauer C, Mozgai S, Chuang B, Woolley J, Scherer S. Manual and automatic measures confirm—intranasal oxytocin increases facial expressivity. Seventh International Conference on Affective Computing and Intelligent Interaction (ACII); 2017 Oct; San Antonio (TX). p. 229–235.

O'Malley T, Bursztein E, Long J, Chollet F, Jin H, Invernizzi L. Keras Tuner. 2019 [accessed July 2020]. https://github.com/keras-team/keras-tuner.

Parra F, Miljkovitch R, Persiaux G, Morales M, Scherer S. The multimodal assessment of adult attachment security: developing the biometric attachment test. J Medical Internet Res. 2017;19(4):e100.

Phillips E, Ososky S, Grove J, Jentsch F. From tools to teammates: toward the development of appropriate mental models for intelligent robots. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2011;55:1491–1495.

Porges SW. The polyvagal perspective. Biol Psychol. 2007;74(2):116–143.

Salas E, Rosen MA, Burke CS, Goodwin GF. The wisdom of collectives in organizations: an update of the teamwork competencies. In: Salas E, Goodwin GF, Burke CS, editors. Team effectiveness in complex organizations: cross-disciplinary perspectives and approaches. New York (NY): Routledge/Taylor & Francis Group; 2009. p. 39–79.

Salas E, Sims DE, Burke CS. Is there a "big five" in teamwork? Small Group Res. 2005;36(5):555–599. https://doi.org/10.1177/1046496405277134.

Sanders TL, Schaefer KE, Yordon RE, Billings DR, Hancock PA. Classification of robot form: factors predicting perceived trustworthiness. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2012;56(1):1548–1552.

Schaefer KE, Baker AL, Brewer RW, Patton D, Canady J, Metcalfe JS. Assessing multi-agent human-autonomy teams: US Army robotic Wingman gunnery operations. In: George T, Islam MS, editors. SPIE Defense + Commercial Sensing; 2019a; Baltimore (MD). International Society for Optics and Photonics. p. 109822B.

Schaefer KE, Brewer RW, Pursel ER, Zimmerman A, Cerame E, Briggs K. Outcomes form the first Wingman software-in-the-loop integration event. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2017b June. Report No.: ARL-TN-0830.

Schaefer KE, Brewer RW, Pursel ER, Zimmerman A, Cerame E. Advancements made to the Wingman software-in-the-loop (SIL) simulation: how to operate the SIL. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2017c Dec. Report No.: ARL-TR-8254.

Schaefer KE, Brewer RW, Pursel ER, Desormeaux M, Zimmerman A, Cerame E. US Army robotic Wingman simulation: integration workshop. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2018 Nov. Report No.: ARL-TR-8572.

Schaefer KE, Brewer RW, Baker AL, Pursel ER, Gipson B, Ratka S, Giacchi J, Cerame E, Pirozzo K. US Army Wingman Joint Capability Technology Demonstration (JCTD): initial Soldier and Marine feedback on manned-unmanned gunnery operations. Aberdeen Proving Ground (MD): DEVCOM Army Research Laboratory (US); 2019b Mar. Report No.: ARL-TR-8663.

Schaefer KE, Gremillion G, Krausman A. Human-autonomy teaming essential research program: team trust and team cohesion midpoint project updates and plans. Aberdeen Proving Ground (MD): DEVCOM Army Research Laboratory (US); 2020 June. Report No.: ARL-MR-1021.

Schaefer KE, Perelman BS, Gremillion GM, Marathe AR, Metcalfe JS. A roadmap for developing team trust for human-autonomy teams. In: Lyons J, Nam C, editors. Trust in human-robot interaction: research and applications. Elvesier. Forthcoming 2020.

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 Annual Meeting. 2012;56(1):1548–1552.

Schaefer KE, Scribner DR. Individual differences, trust, and vehicle autonomy: a pilot study. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2015;59(1):786–790. https://doi.org/10.1177/1541931215591242.

Schaefer KE, Straub ER. Will passengers trust driverless vehicles? Removing the steering wheel and pedals. 2016 IEEE International Multi-Disciplinary

Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA); 2016 Mar; San Diego, CA. p. 159–165.

Schaefer KE, Straub ER, Chen JYC, Putney J, Evans AW. Communicating intent to develop shared situation awareness and engender trust in human-agent teams. Cog Sys Res. 2017a;46:26–39.

Scherer S, Lucas GM, Gratch J, Rizzo AS, Morency LP. Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews. IEEE Trans Affect Comput. 2016;7(1):59–73.

Seppelt BD, Lee JD. Making the limits of adaptive cruise control visible. Int J Hum Comput Stud. 2007;65:192–205.

Shi Y, Ruiz N, Taib R, Choi E, Chen F. Galvanic skin response (GSR) as an index of cognitive load. Extended Abstracts–Proceedings of the Conference on Human Factors in Computing Systems, 2007 Apr 28–May 3; San Jose, CA.

Spain RD, Bliss JP. The effect of sonification display pulse rate and reliability on operator trust and perceived workload during a simulated patient monitoring task. Ergonomics. 2008;51(9):1320–1337. https://doi.org/10.1080/00140130802120234.

Stanton NA, Young MS, Walker GH. The psychology of driving automation: a discussion with Professor Don Norman. Int J Veh Des. 2007;45(3):289–306.

Appendix A. Performance Scores

Table A-1 Performance scores by step

Crew	Step 1 EX1A/2B 61	Step 2 EX1A/2B 67	Step 3 EX1A/2B 65	Step 4 EX1A/2B 64	Step 5 EX1A/2B 69	Step 6 EX1B/2A 60	Step 7 EX1B/2A 62	Step 8 EX1B/2A 66	Step 9 EX1B/2A 68	Step 10 EX1B/2A 63	EX1A/2B Total Points	EX1A/2B Qual Eng	EX1B/2A Total Points	EX1B/2A Qual Eng	Table Final Score	Total Qual Eng	Crew Rating
101-102	75	50	90	78	90	96	100	93	0	0	383	4	289	3	672	7	U
201-202	15	88	50	100	87	75	100	97	0	0	340	3	272	3	612	6	U
301-302	30	50	91	0	94	96	100	88	94	100	265	2	478	5	743	7	Q
401-402	83	0	50	93	88	0	100	70	39	100	314	3	309	2	623	5	U
501-502	35	82	90	79	98	86	100	91	0	0	384	3	277	3	661	6	U
601-602	~~45~~	~~50~~	~~49~~	~~40~~	~~7~~	70	100	75	48	100	191	0	393	3	584	3	U
701-702	0	57	94	79	86	93	100	50	77	100	316	3	420	3	736	6	U
801-802	81	90	93	10	50	0	0	70	30	100	324	3	200	1	524	4	U
901-902	30	44	37	30	95	89	0	50	45	100	236	1	284	2	520	3	U
1001-1002	0	0	0	0	0	89	0	50	20	100	0	0	259	2	259	2	U
1101-1102	0	50	77	80	98	75	100	77	43	100	305	2	385	3	700	5	U
1201-1202	55	65	47	77	94	60	100	86	52	100	338	2	398	3	736	5	U
1301-1302	55	42	89	70	91	93	0	50	38	100	347	3	281	2	628	5	U
1401-1402	0	40	50	93	50	70	100	75	17	0	262	2	233	1	495	3	U
Averages	36.0	50.6	64.8	59.2	73.4	70.9	71.4	73.0	35.9	71.4	286.1	2	319.9	3	606.6	5	
Avg 1A/1B	32.14	53.57	81.14	59.43	93.14	52.00	71.43	74.71	29.43	71.43	319.43	2.57	294.86	2.14			
Avg 2A/2B	39.86	47.57	48.43	59.00	53.71	89.71	71.43	71.29	42.43	71.43	252.71	1.86	344.86	3.00			

1A/1B	1st set of targets - 100 seconds
2A/2B	2nd set of targets - 50 seconds
	Single target engagements
	Multiple target engagements
	Defense
	Offense

Note: Crew 101–102 did not complete Step 9 and Step 10 of Exercise 2 because of time constraints. Crew 1001–1002 were not able to complete any steps in Exercise 2 because of technical issues with the simulation. The zero scores for these incomplete steps are noted by a black line through them. The crew rating is determined by score with 900–1000 as Distinguished (D), 800–899 as Superior (S), 700–799 as Qualified (Q), and 0–699 as Unqualified (U).

Appendix B. Inter-Beat-Interval (IBI) and Heart-Rate (HR) Measures Incorrectly Derived by the Empatica

Figures B-1 and B-2 showed comparisons of IBI and HR time-series data using IBI.csv and HR.csv that were output by the Empatica software package and data using our own processing pipelines. Compared with our derived estimates, the IBI data output by the Empatica showed big gaps of missing values, suggesting the algorithms used by the Empatica to derive its IBI values from the blood-volume-pulse (BVP) signals were not robust and reliable. Our beat-to-beat HR signals' processing pipelines outperformed the Empatica's, producing IBI time-series data that were the basis for calculating HR and heart-rate-variability metrics. Similarly, not only did the HR data from the Empatica not show the appropriate inversed relationship with the IBI data, it also appeared to be too smooth to be true (see Fig. B-2).

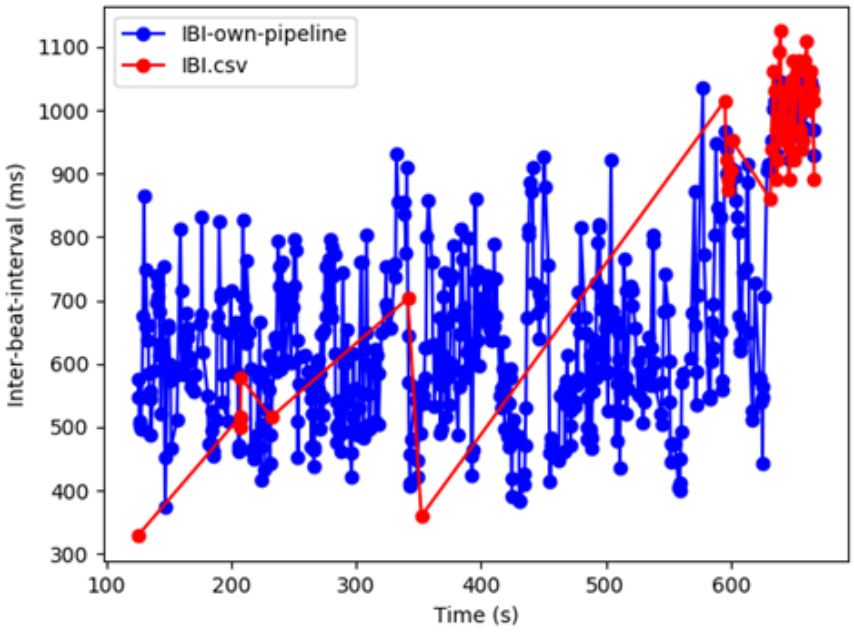

Fig. B-1 Comparison of IBI time series using IBI.csv output by Empatica and our own processing pipelines derived from BVP signals

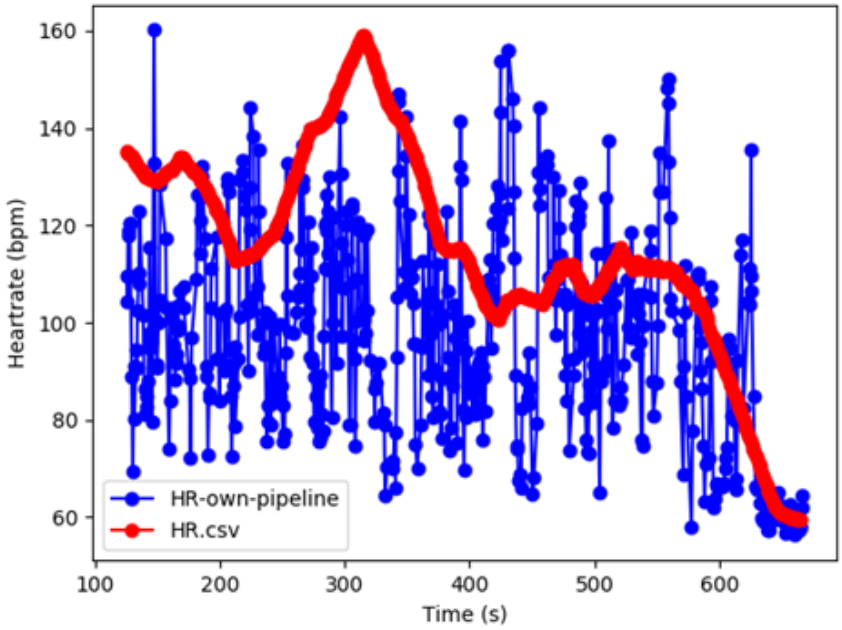

Fig. B-2 Comparison of HR time series using HR.csv output by Empatica and our own processing pipelines derived from BVP signals

List of Symbols, Abbreviations, and Acronyms

ANOVA	analysis of variance
ARL	Army Research Laboratory
AU	action unit
BVP	blood volume pulse
CI	confidence interval
DA	Department of the Army
DEVCOM	US Army Combat Capabilities Development Command
DIDEA	detect, identify, decide, engage, assess
DOD	US Department of Defense
DYS	dysphoria
EDA	electrodermal activity
FACS	Facial Action Coding System
HR	heart rate
HRV	heart rate variability
IBI	inter-beat-interval
IR	infrared
JCTD	Joint Capabilities Technology Demonstration
LRAS3	Long-Range Advanced Scout Surveillance System
LSA	latent semantic analysis
LSM	language style matching
MAACL-R	Multiple Affect Adjective Checklist–Revised
NASA	National Aeronautics and Space Administration
NASA-TLX	NASA Task Load Index
PASS	Positive Affect and Sensation Seeking
PPG	photoplethysmography
RCV	Robotic Combat Vehicle

RMSSD	Root Mean Square of Successive Differences
ROCKET	**R**and**o**m **C**onvolutional **Ke**rnel **T**ransform
SA	situation awareness
SD	standard deviation
TRQ	Team Readiness questionnaire
TSC	time series classification
USMA	US Military Academy
WMI	Warfighter Machine Interface

1 (PDF) DEFENSE TECHNICAL INFORMATION CTR DTIC OCA

1 (PDF) DEVCOM ARL FCDD RLD DCI TECH LIB

1 (PDF) DEVCOM ARL FCDD RLH B T DAVIS BLDG 5400 RM C242 REDSTONE ARSENAL AL 35898-7290

1 (PDF) DEVCOM ARL FCDD HSI J THOMAS 6662 GUNNER CIRCLE ABERDEEN PROVING GROUND MD 21005-5201

1 (PDF) USAF 711 HPW 711 HPW/RH K GEISS 2698 G ST BLDG 190 WRIGHT PATTERSON AFB OH 45433-7604

1 (PDF) USN ONR ONR CODE 341 J TANGNEY 875 N RANDOLPH STREET BLDG 87 ARLINGTON VA 22203-1986

1 (PDF) USA NSRDEC RDNS D D TAMILIO 10 GENERAL GREENE AVE NATICK MA 01760-2642

1 (PDF) OSD OUSD ATL HPT&B B PETRO 4800 MARK CENTER DRIVE SUITE 17E08 ALEXANDRIA VA 22350

3 (PDF) USCENTCOM A LESTER G TIGHE A VAID

2 (PDF) US ARMY MILITARY ACADEMY E ROVIRA R THOMSON

5 (PDF) GROUND VEHICLES SYSTEM CENTER RTI GVR T MICHALIK K BRIGGS E CERAME K PIROZZO J PALOMINO

1 (PDF) NGCV CFT C WALLACE

2 (PDF) NAVAL SURFACE WARFARE CENTER R BEALE E PURSEL

ABERDEEN PROVING GROUND

18 (PDF) DEVCOM ARL FCDD RLH J LANE Y CHEN P FRANASZCZUK K MCDOWELL K OIE A MARATHE FCDD RLH F J GASTON FCDD RLH FA A DECONSTANZA E CARTER J CANADY C NEUBAUER D CHHAN FCDD RLH FB D BOOTHE FCDD RLH FC K COX FCDD RLH FD A FOOTS A KRAUSMAN A BAKER FCDD RLH FE D HEADLEY FCDD RLH P A EVANS FCCD RLV J RIDDICK B PIEKARSKI FCDD RLV A K SCHAEFER-LAY R W BREWER