Identifying New Team Trust and Team Cohesion Metrics that Support Future Human-Autonomy Teams

Alyssa Milner[1], Dae Han Seong[1], Ralph W. Brewer[2], Anthony L. Baker[2], Andrea Krausman[2], David Chhan[2], Robert Thomson[1], Ericka Rovira[1], Kristin E. Schaefer[2]

[1] United States Military Academy, West Point, NY USA
{alyssa.milner, daehan.seong, robert.thomson, ericka.rovira}@westpoint.edu,
[2]US Army Combat Capabilities Development Command Army Research Laboratory, Aberdeen Proving Ground, MD USA
{ralph.w.brewer,civ, anthony.l.baker61.ctr, andrea.s.krausman.civ, david.chhan.civ, kristin.e.schaefer-lay.civ}@mail.mil.

Abstract. A driving simulation study of a manned-unmanned vehicle gunnery team was conducted to assess potential metrics of team trust and cohesion for evaluating future human-autonomy teams. Cadet dyads worked with a veteran commander within driving simulation to direct a weaponized robotic ground vehicle from a command and control vehicle and identify and engage targets on a gunnery range. Subjective, behavioral, performance, communication, and physiological data were collected to identify possible team trust and team cohesion metrics. Findings suggest that performance, behavior, and physiological data may provide useful windows into the trust and cohesion exhibited by crew members in human-autonomy teams.

Keywords: Human-autonomy teaming · Simulation · Wingman · Metrics · Trust · Cohesion

1 Introduction

The US Army seeks to identify emerging technologies and technology-enabled concepts that could provide significant military advantage during operations in complex, contested, or congested environments. As robotic technologies advance from teleoperation to advanced autonomy, it is essential to develop appropriate interdependent collaboration between the human and autonomy-enabled team members [1]. Effective teaming and appropriate use of the technology depends on the human's understanding of the system, its behaviors, and the reasoning behind those behaviors [2]. If human expectations do not match system behaviors, users will question the accuracy and effectiveness of the system's action, which can lead to degraded trust, and potentially, misuse or disuse of the system, even if it is operating effectively [3, 4].

Driving simulation research has led to major advances in both the development of autonomy and the understanding of how trust can be measured and develops. For example, recognizing the limitations of subjective assessments of trust [5], driving simulation research has investigated other promising methods such as behavioral measures

(e.g., eye gaze) and physiological indicators of changes in trust-based behaviors by measuring electrodermal activity and changes in facial response [6, 7]. For human-autonomy teams, driving simulation has provided viable insights into system needs and interactions in early development, teaming with remote operation of unmanned robotic vehicles, and multi-method approaches for assessing human-autonomy interactions [8].

The Combat Capabilities Development Command Army Research Laboratory's Human Autonomy Teaming Essential Research Program has recently made headway in determining and developing metrics for assessing human-autonomy teams through the use of driving simulation and field research. This research suggests the importance of team dynamics, communication, and a multi-method approach for developing team trust metrics [9-11], and seeks to expand trust and cohesion metrics to larger human-autonomy teams. Therefore, the goal of this research is to explore these metrics related to performance, behavior, communication, and physiological indicators for human-autonomy teams during simulated manned-unmanned team gunnery exercises.

2 Methodology

2.1 Simulation

The Wingman simulation testbed is a software-in-the-loop simulation environment. This means that it integrates all the real-world Wingman vehicle mobility and lethality autonomy software into a lab-based virtual setting [12-14]. It was designed to support a 5-man crew station on a command and control vehicle, where the roles (commander, LRAS3 operator, vehicle driver, robotic operator and robotic gunner) could be manned or simulated. For this experiment, the vehicle driver was a simulated role that allowed the manned vehicle to move through the virtual gunnery range. The LRAS3 operator role was simulated by targets being identified and communicated through the Warfighter Machine Interface (WMI) user display. The commander role was filled by a confederate to communicate tasks and team instructions, and the participants filled the roles of the robot mobility operator and robot lethality operator. The robotic mobility operator was responsible for maneuvering the vehicle via teleoperation or initiating the vehicle autonomy, and helping the lethality operator identify target and trajectory locations. The robotic lethality operator was tasked with detection, identification, and engagement of the target using teleoperation or engaging weapon system autonomy.

2.2 Participants

A total of 36 United States Military Academy Cadets, enrolled in introductory psychology courses were recruited through the SONA participant pool. Analyses were conducted on 12 dyads following removal of incomplete data on 2 dyads, and 8 no show participants.

2.3 Design

The Wingman task required participants to work as a team with an unmanned weaponized robotic combat vehicle to identify and engage multiple targets on a gunnery range. Teams completed two sets of gunnery exercises, each consisting of five target engagements, or sets of targets. These engagements included two offensive and three defensive operations, or postures, on stationary targets. The study was a within-subjects design where Exercise 1 had a target exposure time of 100s and Exercise 2 had a target exposure time of 50s (in line with Army requirements). To avoid training effects, two different courses were used. Course order was counterbalanced with the order for the two gunnery exercises being fixed. The dependent variables were selected to help identify metrics of team trust and cohesion. Performance was measured using the Army's standard for remote weapon station gunnery [15]. Subjective scales provided insight into team trust and cohesion [9], stress using the Multiple Affect Adjective Checklist-Revised (MAACL-R) [16], and workload using the NASA-Task Load Index (NASA-TLX) [17]. Behavioral data was captured via video of the participants' facial expressions and interactions with the system [18]. Psychophysiological changes were measured by electrodermal activity (EDA), heart rate (HR), and heart rate variability (HRV), to associate with a change in trust or the onset of a trust-based decision. Audio recordings of spoken communication provide semantic content. Only a subset of these findings are reported here.

2.4 Equipment

The simulation testbed was set up in an isolated 8x10 laboratory. The room was set up with three touch screen monitors each running the WMI software. The WMI provides an interactive customizable display for the mobility operator, lethality operator, and vehicle commander to provide supervisory control over the associated autonomous systems on the weaponized vehicle [18]. A script running on each WMI computer recorded each interaction between the user and the screen. A Logitech web camera was placed on top of each WMI to capture the facial expressions of the Cadets. A boom microphone was connected to the evaluator computer to capture verbal communication from the crew. Each participant had an Empatica E4 wristband sensor to read and record physiological data during the study.

2.5 Procedure

Following informed consent, participants were randomly assigned either the lethality or mobility operator role, and fitted with an Empatica sensor. If both participants agreed to be visually and audio recorded, Logitech cameras and audio recorders were initiated. Participants were trained on their respective WMI, the fundamental controls, and their specific role. They then completed two exercises. After each exercise, participants completed the trust in the robotic combat vehicle, team readiness, stress, and workload questionnaires. The entire study took 55 minutes to complete.

3 Results and Discussion

3.1 Performance

Gunnery performance was measured as a crew and reported in Table 1. Even though the length of the exercises was different, the standard for qualification remained the same.

Table 1. Qualification Scores

Crew	Exercise 1			Exercise 2		
	# qualified	Total Score	Avg Score	# qualified	Total Score	Avg Score
2	3	272	54.4	3	340	68
3	2	265	53	5	478	95.6
4	2	309	61.8	3	314	62.8
5	3	384	76.8	3	277	55.4
6	3	393	78.6	0	191	38.2
7	3	316	63.2	3	420	84
8	1	200	40	3	324	64.8
9	1	236	47.2	2	284	56.8
11	2	305	61	3	385	79
12	3	398	79.6	2	338	67.6
13	3	347	69.4	2	281	56.2
14	1	233	52.4	2	262	46.6

Note. Each engagement had a possible maximum score of 100 points whereby 70 points is considered a qualifying score.

A two-way ANOVA was conducted on the performance scores for exercise and posture. 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=.613. There was a significant difference in posture, $F(1, 116)=10.70$, p=.001 where by 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=.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 six 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. The same is true here with only two exercises totaling 55 minutes.

3.2 Subjective Response

Paired samples t-tests were conducted to assess differences in subjective responses on trust, cohesion, stress (T-scores), and workload between Exercise 1 and 2. The only significant difference was in workload, $t(23)=2.30$, p=.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).

Additional analysis was conducted to assess the impact of role on subjective ratings. When analyzing stress, we found that positive affect for the mobility operator was significantly higher, $t(39.79)=-2.40$, $p=.021$, whereas the lethality operator scored significantly higher on sensation seeking), $t(44.39)=2.08$, $p=.044$. With respect to the dysphoria scale, a composite of anxiety, depression, and hostility scales, lethality operators scored significantly higher, $t(34.82) = 2.35$, $p = .025$, which may suggest the presence of emotional distress. This matches with findings from the workload assessment which showed a significant main effect of role on total workload , $F(1,44)=19.75$, $p<.001$, mental demand, $F(1,44)=6.61$, $p=.014$, temporal demand, $F(1,44)=14.46$, $p<.001$, frustration, $F(1,44)=5.51$, $p=0.24$, and effort, $F(1,44)=15.38$, $p<.001$, whereby the lethality operator had higher workload than the mobility operator. Yet no significant differences were found in trust in the robotic vehicle, trust in the weapon system autonomy, mobility autonomy, or team. However interestingly, trends show that the mobility operator had higher trust in the weapon system autonomy, and the lethality operator had higher trust in the mobility autonomy, and overall the mobility operator had higher trust in the team as a whole.

3.3 Physiological Data

Photoplethysmogram (PPG) is a noninvasive technique used for heart rate monitoring. It uses a light source and a photodetector to measure changes in blood flow. While PPG based wrist-worn wearable sensors such as the Empatica used in this study provide useful information related to the cardiovascular system during task performance, research has shown that the data is less reliable and susceptible to noise when the wrist is moving. Therefore, although Figure 2 indicates that participants reduced their movements during experiment phases (colored windows), these insights should be considered exploratory in nature.

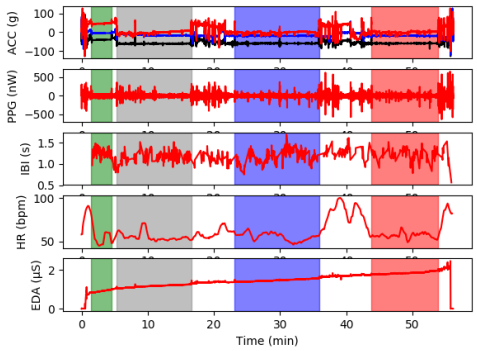

Fig 2. An example of raw Empatica data from one participant during the whole session. Shown are time series of the 3-axis acceleration (ACC), the photoplethysmogram (PPG) that was the basis of inter-beat-interval (IBI) and heart rate (HR) measures, and the electrodermal activity (EDA). Periods of participant resting, training, Exercise 1 and Exercise 2 are demarcated respectively with shaded green, grey, blue and red.

To understand how physiological measures could be used to predict team trust and cohesion over time and how they relate to team performance, we computed IBI synchrony, the cross correlation of the inter-beat interval time series for each dyad during each period. It is suggested that higher IBI synchrony indicates greater strength of team coordination and interaction that may explain variability in team performance. Though we did not find any significant correlations between IBI synchrony and team performance using this approach, we continue to explore different analytical techniques to quantify team physiology (i.e. heart rate variability synchrony) that might shed light on using physiological coupling to predict team performance. It is possible that, with more data to draw from and further testing, these real-time assessments may allow us to understand fluctuations in team trust and cohesion in real-time. Additional analyses are being conducted to look at the relationship between performance, subjective response and physiological indicators that occur during posture, between roles, and at specific points throughout each engagement.

4 Conclusion

Study results highlight the benefit of simulation research by providing the ability to understand system concepts and team constructs, such as trust and cohesion, during early system development. Overall, results showed a reasonable level of performance and trust in the robotic vehicle and team given minimal training and experience with the robotic system, team, and task. Strikingly, participants tended to trust the autonomy more than human teammates. This suggests that a robotic vehicle can be viewed as a teammate, although additional analysis on behavioral interaction with the autonomy and more in depth analysis of physiological response will provide additional insights into the team relationship. A key finding specific to team trust metrics specified the importance of quantifying workload and stress associated with individual roles. Future work will include how participants judge their trust in autonomy: do they perceive it to have a lower workload, or overall better performance than a human operator? Follow on studies are planned for field operations with the real-world vehicles with increased training and practice.

Acknowledgments. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the CCDC Army Research Laboratory or the US Government. The US Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. We would like to thank Scott Kerick, Jonroy Canady, Cathy Neubauer, Sean Fitzhugh, and Debbie Patton for their support in the study development and analysis.

References

1. Phillips, E., Ososky, S., Grove, J., Jentsch, F.: From tools to teammates: Toward the development of appropriate mental models for intelligent robots. In: Proc. of the Human Factors and Ergonomics Society Annual Meeting, pp. 1491--1495. SAGE Publications, CA (2011)

2. Chen, J.Y., Procci, K., Boyce, M., Wright, J., Garcia, A., Barnes, M.: Situation awareness-based agent transparency (ARL-TR-6905). US Army Research Laboratory: APG, MD (2014)
3. Lee, J.D., See, K.A.: Trust in automation: Designing for appropriate reliance. Human Factors 46, 50-80 (2004)
4. Schaefer, K.E., Straub, E.R.: Will passengers trust driverless vehicles? Removing the steering wheel and pedals. In: 2016 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support, pp. 159--165. IEEE (2016)
5. Walker, F., Verwey, W., Martens, M.: Gaze behavior as a measure of trust in automated vehicles. In: Proc. of the 6th Humanist Conference, pp. 1--6. (2018).
6. Hergerth, S., Lorenz, L., Vilimek, R., Krems, J.F.: Keep your scanners peeled: Gaze behavior as ameasure of automation trust during highly automated driving. Human Factors, 58(3), 509--519. (2016).
7. Shahrdar, S., Park, C., Norjoumain, M.: Human trust measurement using an immersive virtual reality autonomous vehicle simulator. In: Proc. Conference on AI, Ethics, and Society. pp. 515--520, AAAI/ACM, (2019).
8. Schaefer, K.E., Brewer, R.W., Perelman, B.S., Ray Pursel, E., Cerame, E., Drnec, K., Paul, V., Haynes, B., Donavanik, D., Gremillion, G., Metcalfe, J.S.: Challenges with Developing Driving Simulation Systems for Robotic Vehicles. In: Cassenti D. (ed) Advances in Human Factors in Simulation and Modeling, pp. 139--150. Springer (2019)
9. Schaefer, K.E., Baker, A.L., Brewer, R.W., Patton, D., Canady, J., Metcalfe, J.S.: Assessing multi-agent human-autonomy teams: US Army Robotic Wingman gunnery operations. In: Micro-and Nanotechnology Sensors, Systems, and Applications XI, pp. 109822B. International Society for Optics and Photonics, (2019)
10. Baker, A.L., Schaefer, K.E., Hill, S.G.: Teamwork and Communication Methods and Metrics for Human-Autonomy Teaming (ARL-TR-8844). US Army Research Laboratory: APG, MD (2019)
11. Huang, L., Cooke, N., Gutzwiller, R., Berman, S., Chiou, E., Demir, M., Zhang, W.: Distributed Dynamic Team Trust in Human, Artificial Intelligence, and Robot Teaming. In: Nam C. and Lyons J. (eds.) Trust in Human-Robot Interaction: Research and Applications. Elsevier (in press)
12. Schaefer, K.E., Brewer, R., Pursel, E., Zimmermann, A., Cerame, E., Briggs, K.: Outcomes from the first Wingman software-in-the-loop integration event: January 2017 (ARL-TR-0830). US Army Research Laboratory: APG, MD (2017)
13. Schaefer, K.E., Brewer, R.W., Pursel, E.R., Zimmermann, A., Cerame, E.: Advancements Made to the Wingman Software-in-the-Loop (SIL) Simulation: How to Operate the SIL (ARL-TR-8254). US Army Research Laboratory: APG, MD (2017)
14. Schaefer, K.E., Brewer, R.W., Pursel, E.R., Desormeaux, M., Zimmermann, A., Cerame, E.: US Army Robotic Wingman Simulation: June 2018 Integration Workshop (ARL-TR-8572). US Army Research Laboratory: APG, MD (2018)
15. US Army Trainng and Doctrine Command: Training and qualification, crew. Training Circular No.: TC 3-20.31. Department of the Army (US), Washington (DC), 17 March 2015
16. Lubin, B., Zuckerman, M.: Manual for the MAACL-R: Multiple Affect Adjective Check List-Revised. Educational and Industrial Testing Service: San Diego, CA, (1999)
17. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. Advances in Psychology, 52, 139--183 (1988)
18. Brewer, R.W., Cerame, E., Pursel, E.R., Zimmerman, A., Schaefer, K.E.: Manned-unmanned teaming: US Army Robotic Wingman Vehicles. In: Cassenti D. (ed.) Advances in Human Factors in Simuation and Modeling, pp. 89--100. Springer (2019)