Human Factors and Simulation, Vol. 30, 2022, 52–59
https://doi.org/10.54941/ahfe1001491

Evaluation of Human-Autonomy Team Trust for Weaponized Robotic Combat Vehicles

Ralph Brewer, Anthony Baker, Andrea Krausman, Catherine Neubauer, Daniel Forster, Angelique Scharine, Samantha Berg, Kristi Davis, and Kristin E. Schaefer

US DEVCOM Army Research Laboratory APG, MD 21005, USA

ABSTRACT

Phase I of the Soldier Operational Experiment was held at Fort Carson, Colorado in 2020, to assess the current capability of a manned vehicle and unmanned weaponized vehicle collaborative team capabilities during live fire gunnery operations and situational training exercises. Here we discuss the performance of the crews during these exercises, and the implementation of team trust metrics to evaluate crew dynamics in these human-autonomy lethality teams. The gunnery exercise performance scores demonstrated that teams were often able to achieve qualifying scores on the relevant gunnery standards. However, subjective measures showed relatively low to moderate levels of trust across crew members. Through further analysis we found that Soldiers opted to perform many tasks manually and were slow to adapt to and use the technologies, even with substantial training on the systems. One possible reason for this response to the technology was due to the technology being early in a development cycle and completely new to the users. Linguistic analyses were conducted on the crew communication in order to provide a more fine-grained analysis of the team dynamic. Results indicated that higher performing crews used more formal communication with words associated with perception (e.g., seeing, hearing, etc.). In line with previous field studies through the Wingman Joint Capabilities Technology Demonstration, this study further validated a multi-method approach to understanding performance, trust, and cohesion in human-autonomy teams.

Keywords: Human-Autonomy team, Trust, Gunnery, After-Action review, Robotic combat vehicle, Next generation combat vehicle

INTRODUCTION

As part of modernization efforts, the US Army identifies emerging technologies and technology-enabled concepts that enhance Soldier survivability and lethality in complex and contested environments. One such concept, involves teams of optionally manned combat vehicles and multiple autonomy-enabled, weaponized robotic combat vehicles (RCVs), equipped with advanced sensors and optics. These advanced technologies enhance Soldier situation awareness (SA) and potentially reduce overall decision-cycle time. Future autonomous agents may eventually become interdependent team members instead of being perceived as "tools" used by Soldiers to accomplish their mission (Phillips et al., 2011). Understanding how these inherent differences change team dynamics and interactions is critical; otherwise, technology can have the reverse effect - making tasks require more resources, training, and expertise (Johnson & Vera, 2019).

Although Soldiers already utilize advanced technologies, integrating autonomous systems as full-fledged team members requires a paradigm shift as autonomy reasons, acts and communicates differently than humans which may alter team dynamics and trust. It is clear from the human team literature that trust and cohesion facilitate team processes such as information sharing, decision making, and ultimately team effectiveness (Mesmer-Magnus & DeChurch, 2009). In a team with trust, members look out for the interests of others rather than self-interests, which in turn minimizes the sense of vulnerability and uncertainty, allowing team members to take risks that promote cooperation and team effectiveness (Colquitt, Scott, & LePine, 2007). Where trust is lacking, however, team members work to protect their self-interest at the expense of the interests and goals of the team, creating a breakdown in teamwork (Joshi, Lazarova, & Liao, 2008). Soldiers who distrust teammates will not rely on them. Similarly, technology perceived to be unreliable will be "left in the box" (unused) if perceived as unsafe. This lack of trust can have severe, even life-threatening consequences for teams operating in adverse environments.

An accurate understanding of the system and its behavior is critical for effective teaming (Chen & Barnes, 2014) and is, therefore, vital for human–autonomy teaming (HAT). Measuring trust within human-autonomy teams requires novel approaches that capture changes in trust over time and enabling intervention when needed. This research on measuring trust builds on previous research from the Wingman Joint Capabilities Technology Demonstration (JCTD) and measures team dynamics in manned-unmanned gunnery crews using a multi-method approach (Schaefer et al., 2019; Baker et al., 2020; Milner et al., 2020; Schaefer et al., 2021).

SOLDIER OPERATIONAL EXPERIMENT

The Soldier Operational Experiment (SOE) was designed to obtain soldier feedback on new technologies integrated into two types of vehicles and associated technologies: the Mission Enabling Technology Demonstrator (MET-D) and RCV vehicles (Hatch, 2020). The MET-D vehicle is a modified military armored fighting vehicle which has enhanced technologies, with the Soldiers commanding the RCV vehicles from inside. Information derived from the SOE helps quantify the potential benefit of several new technologies and paradigms such as crew sizes for effective operation of vehicles. Simulated operational exercises and gunnery scenarios at the SOE allow for evaluation of how the platoon deploys the RCVs during realistic exercises, which helps in forming doctrine.

The participants of this experiment were from a platoon of Soldiers from a unit at Ft. Carson, CO. Four Soldiers were trained as MET-D operators

Table 1. Performance metrics.

Engagement ID\ Posture	1 DEF	2 OFF	3 DEF	4 DEF	5 DEF	6 DEF	7 DEF	Final Score
7	0	5	94	0	100	100	0	299
3	100	89	100	100	100	100	100	684
4	0	100	100	100	100	100	100	600
6	100	85	100	100	100	100	100	685
AVG	50	69.8	98.5	75	100	100	75	568.3

(2 MET-D crews) and eight Soldiers as RCV operators/gunners. Prior to participating in the SOE, vehicle crews were given time to get familiar with the vehicles and the systems being evaluated using a combination of simulation followed by field training on the actual vehicles. A multi-method data analysis approach was devised to gain a broad understanding of the Soldiers' interactions with the autonomous systems, the crews' performance, and their trust dynamics throughout the SOE. The analyses will draw from performance scores, self-report questionnaire data, in-vehicle audio recordings, and after-action reviews (AARs).

PERFORMANCE

The US Army provides a set of standards for training and evaluating live fire gunnery operations for manned crews of direct fire ground systems with Training Circular (TC) 3-20.31 titled Training and Qualification, Crew ([HQDA], 2015). During the evaluation an experienced and trained team of vehicle crew evaluators (VCEs) work in concert to assess the proficiency of each crew on their ability to negotiate the course during each of the ten engagements. During the engagements the VCEs record both video and audio of the run. The VCEs enter the timing of the engagement of targets into the common crew score sheet to determine a base score of the engagement.

A total of four RCV crews took part in the live fire evaluations and their performance is scores are listed in Table 1. A qualifying score for each engagement is 70 to 100 points. Crew 7 was the most novice of all the crews and their scores reflect their green status. Their problem stemmed from not rapidly identifying the targets coupled with taking too long to engage and destroy the targets. The performance metrics show where crews had problems and by conducting an AAR the issues were identified and addressed for future improvement.

FEEDBACK

Feedback from the Soldiers was elicited during AARs. Both verbal responses as well as questionnaires were administered during the experiment. This section highlights results from the AARs and three questionnaires: Situational Awareness Systems Trust (adapted from Muir & Moray, 1996), Robotic

Table 2. Mean (sd.) SA systems ratings.

		Competence	Dependability	Responsibility	Trust
MET-D	Operator	52.6 (30.1)	51.3 (29.4)	51.4 (29.6)	52.3 (31.6)
	Gunner	43.6 (27.7)	45.1 (28.0)	39.6 (32.7)	42.9 (33.2)
RCV	Operator	26.1 (30.8)	26.9 (31.9)	26.9 (31.9)	25.9 (32.6)
	Gunner	46.9 (36.8)	47.6 (35.9)	43.0 (36.8)	50.6 (36.8)

Vehicle Trust (adapted from Schaefer et al., 2012), and the Gunnery questionnaire. These questionnaires were used to assess trust in the systems and obtain feedback as they interacted with the technologies.

The Soldiers' responses during the AAR process provide explanations for why they sometimes had difficulty trusting their systems. Safety and reliability concerns played a key role; the novelty of the hardware and software systems, combined with a possible lack of transparency when systems exhibited issues or failures, influenced some of the Soldiers to have less than optimal trust in the systems. This may also be a factor of the state of the systems during the SOE; many features and capabilities of the systems are still under development and need further testing and evaluation before successful implementation. Regardless, the general tone of the AAR responses was that the Soldiers were cautious toward the systems, and unsure of their reliability at the time of testing.

The SA Systems Trust questionnaire measured the competence, dependability, responsibility, and overall trust in the SA Systems (e.g., 360-degree cameras) on a scale from 0 (no trust) to 100 (high trust). Means were calculated for each scale item (Table 2). Overall, RCV Operator scores were lower than the other crew members across all four scale items suggesting that the RCV Operators were more reluctant to trust the SA Systems. Trust scores for the other crew positions although still relatively low, are promising since the SA systems were new to the Soldiers and to use them effectively, they had to adapt to new methods to accomplish their tasks. Further, technical issues with vehicle displays negatively impacted SA levels and likely created a hesitancy to trust the systems and contributed to the large variability in scores. These findings support previous research (Schaefer et al. 2017) highlight the importance of system transparency and user interface design for developing effective trust in human-autonomy teams, perhaps even more so in real-world environments.

Using the Robotic Vehicle Trust questionnaire, RCV crews rated their trust in the robotic vehicles by indicating their agreement on statements pertaining to vehicle intelligence, safety, trustworthiness, and autonomous capability using a scale of 1 (very low agreement) to 7 (very high agreement). With respect to crew members, across the four scale items, RCV Gunner scores were higher than those for RCV Operators (Table 3). All scores remained below the neutral score of 4 suggesting that the RCV crew members did not perceive the robotic vehicle at this early stage of development to be especially intelligent, safe, or trustworthy, and were reluctant to use the autonomous capabilities, which is potentially due to technical issues and limitations on the

Table 3. Mean (sd.) robotic vehicle trust ratings.

	Intelligent	Safe	Trustworthy	Autonomous
RCV Operator	2.14 (2.04)	1.86 (1.86)	1.57 (1.13)	1.43 (1.13)
RCV Gunner	3.50 (1.38)	3.67 (1.51)	3.50 (1.38)	2.00 (1.67)

Table 4. Mean (sd.) Gunnery survey ratings by vehicle and role.

	MET-D	RCV
Operator	2.15 (1.20)	2.97 (1.47)
Gunner	2.66 (1.76)	3.34 (0.94)

speed the RCV could travel in autonomous mode (e.g., for safety reasons) and technical issues. Improvements in system reliability, safety, and autonomous capabilities, coupled with training and hands-on experiences will likely result in more positive perceptions of the systems and increased trust.

For the Gunnery questionnaire participants rated their level of agreement on thirteen questions related to operating the RCV and associated technologies during the live fire exercise using a scale from 1 (strongly disagree) to 5 (strongly agree). Overall mean ratings were calculated by averaging scores from the thirteen scale items (Table 4). Mean ratings were moderate-to-low across participants for each vehicle type and role. Specifically, MET-D crew ratings were lower than for the RCV crews, with slightly higher variability which suggests crews were cautious with respect to using the vehicle capabilities, which was supported by AAR Feedback. These findings are likely a function of the reduced role of the MET-D vehicle during Gunnery - meaning MET-D crew members had little interaction with the RCV mobility and weapon system. This speaks to the need to understand how the attitudes of one individual influence the other crew members, especially since the RCV and MET-D crews are intended to work together to accomplish a mission. Further defining the role of the MET-D crew during Gunnery is another area to be addressed – perhaps the MET-D crew members could support the RCV crew either through weapon handoffs or providing security during a mission.

COMMUNICATION ANALYSIS

Transcripts of four of the gunnery runs were created, and these were used to conduct exploratory linguistic analyses of the dialogue between crew members. The software Linguistic Inquiry and Word Count (LIWC; Pennebaker, Booth, & Francis, 2007) was used to analyze the linguistic properties of the words and phrases used by crew members while conducting the gunnery runs. LIWC analyzes text using pre-defined dictionaries, or categories, to determine the percentage of a given interaction associated with each word category. Categories are associated with function words (e.g., verbs, pronouns), psychological constructs (e.g., positive or negative emotions), or other concepts (e.g., achievement, future-oriented language).

In this dataset, perception-related words (such as 'look', 'heard', and 'saw') accounted for about 6-8% of the language during a given run. Notably, linguistic patterns in "normal" communication media such as blog posts, everyday conversations, novels, etc. demonstrate an average score for perception of about 2.7% (Pennebaker, Boyd, Jordan, & Blackburn, 2015), implying that LIWC was able to capture the considerably perceptual nature of conducting this gunnery task. Indeed, the run with the lowest performance score (run 3), which also had the lowest levels of perception-related words, still showed approximately 4% of its language associated with perceptual processes, above the standard 2.7% baseline mark. These results suggest that perception-related word usage may be a useful indicator of team functioning, especially given the highly perceptual nature of the gunnery task. If these word categories are meaningful indicators of team functioning, they may also improve upon existing measures of trust in human-autonomy teams.

CONCLUSION

Linguistic analyses, performance data, subjective scores, and AAR themes all provide helpful insights for exploring crew interactions with the technologies during the SOE as well as their trust perceptions, especially during early development of new technologies. Overall, performance alone does not tell a full story of the effectiveness and team dynamics for human-autonomy teams. However, subjective feedback, behavioral responses, and even linguistic analysis of crew communication can provide critical details as to supporting and advancing technological capabilities, developing appropriate metrics for evaluation, and team operations. For example, an interesting aspect of the subjective data is that for the Robotic Trust Survey, the RCV Gunner perceived the mobility autonomy to be problematic, even though they did not use the mobility autonomy to control the RCV. It is possible that the Operator influenced the Gunner's perspective about the autonomous mobility of the platform. This is an important area for future research as we refine the trust measures and explore how trust flows within a team (e.g., how one member's trust influences other crew members). Finally, novel metrics are needed to fully understand team dynamics. For example, the better performing crews used more perceptual terms related to target location, aiming, vehicle orientation, which aligns with the types of information being exchanged during the gunnery task and as a result may serve as indicators of crew effectiveness and trust.

To more fully understand human-autonomy team dynamics, integration of the robotic vehicles and team metrics will need to be integrated into the AAR process. If the crew has a shared mental model of the task, roles and interaction then they will be better able to anticipate the needs of one another (Smith-Jentsch et al. 1998). Having an autonomous unmanned asset as part of the crew adds a layer of complexity to the discussion. Therefore, future SOEs and studies will include the Global After-Action Review Technology (Taberski et al. 2021), which was developed specifically to support manned–unmanned gunnery team operations.

REFERENCES

Baker, A. L., Fitzhugh, S. M., Forster, D. E., Brewer, R. W., Krausman, A., Scharine, A., & Schaefer, K. E. (2020). *Team Trust in Human-Autonomy Teams: Analysis of Crew Communication during Manned-Unmanned Gunnery Operations*. (No. ARL-TR-8969).

Chen JYC, Barnes MJ. (2014). Human–agent teaming for multirobot control: a review of human factors issues. IEEE Trans Hum-Mach Sys; 44(1):13–29

Colquitt, J. A., Scott, B. A., & LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. Journal of Applied Psychology, 92(4), 909–927.

Hatch, L. (2020). Troops test emerging Army robotic vehicle technology. Retrieved 18 Sep 2021from https://www.fortcarsonmountaineer.com/2020/08/troops-test-emerging-army-robotic-vehicle-technology/.

[HQDA] Headquarters, Department of the Army. Training and qualification, crew. Washington (DC): HQDA; 2015 Mar 17. Training Circular No.: TC 3-20.31.

Johnson, M., & Vera, A. (2019). No AI Is an Island: The Case for Teaming Intelligence. AI Magazine, 40(1). doi:10.1609/aimag.v40i1.2842

Joshi, A., Lazarova, M. B., & Liao, H. (2008). Getting Everyone on Board: The Role of Inspirational Leadership in Geographically Dispersed Teams. Organization Science, 20(1), 240–252. doi:10.1287/orsc.1080.0383

Mesmer-Magnus, J. R., & Dechurch, L. A. (2009). Information sharing and team performance: a meta-analysis. The Journal of Applied Psychology, 94(2), 535–546. doi:10.1037/a0013773

Milner, A., Seong, D. H., Brewer, R. W., Baker, A. L., Krausman, A., Chhan, D., . . . Schaefer, K. E. (2021). Identifying New Team Trust and Team Cohesion Metrics that Support Future Human-Autonomy Teams, Cham.

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

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic Inquiry & Word Count (LIWC): A computerized text analysis program*. Austin, TX: LIWC.net

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015.

Phillips, E., Ososky, S., Grove, J., & Jentsch, F. (2011). From Tools to Teammates: Toward the Development of Appropriate Mental Models for Intelligent Robots. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 55(1), 1491–1495.

Schaefer, K. E., Baker, A. L., Brewer, R. W., Patton, D., Canady, J., & Metcalfe, J. S. (2019). Assessing multi-agent human-autonomy teams: US Army Robotic Wingman gunnery operations. *Proceedings of SPIE: Micro- and Nanotechnology Sensors, Systems, and Applications XI, 10982,* 10982B.

Schaefer, K. E., Brewer, R. W., Baker, A. L., Krausman, A., Neubauer, C., Fitzhugh, S., Forster, D. E., Chhan, D., Milner, A., Seong, D. H., Thomson, R., & Rovira, E. (2021). *Wingman Joint Capabilities Technology Demonstration: Trust Metrics for Manned–Unmanned Lethality Teams*. (No. ARL-TR-9182).

Schaefer, K. E., Sanders, T. L., Yordon, R. E., Billings, D. R., & Hancock, P. A. (2012, September). Classification of robot form: Factors predicting perceived trustworthiness. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 56, No. 1, pp. 1548-1552). Sage CA: Los Angeles, CA: SAGE Publications.

Schaefer, K. E., Straub, E. R., Chen, J. Y., Putney, J., & Evans III, A. W. (2017). Communicating intent to develop shared situation awareness and engender trust in human-agent teams. *Cognitive Systems Research*, *46*, 26–39.

Smith-Jentsch KA, Cannon-Bowers JA, Tannenbaum SI, Salas E. Guided team self-correction: impacts on team mental models, processes, and effectiveness. Small Group Res. 2008;39:303–327.

Taberski M, Davis K, Schaefer KE, Brewer R. Visualizing human–autonomy team dynamics through the development of a global after-action review technology. International Conference on Applied Human Factors and Ergonomics; 2021. pp. 46–53.