ARL-TR-9622 ● Dec 2022

DEVCOM
ARMY RESEARCH
LABORATORY

Human-Autonomy Teaming Trust Toolkit (HAT^3) Executive Summary

by Catherine Neubauer, Anthony L Baker, Sean M Fitzhugh, Bret Kellihan, Justin Jagielski, and Andrea S Krausman

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-9622 ● Dec 2022

Human-Autonomy Teaming Trust Toolkit (HAT^3) Executive Summary

Catherine Neubauer, Anthony L Baker, Sean M Fitzhugh, and Andrea S Krausman
DEVCOM Army Research Laboratory

Bret Kellihan and Justin Jagielski
DCS Corporation

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

1. REPORT DATE *(DD-MM-YYYY)* December 2022	**2. REPORT TYPE** Technical Report	**3. DATES COVERED (From - To)** 1 October 2021–31 September 2022

4. TITLE AND SUBTITLE Human-Autonomy Teaming Trust Toolkit (HAT^3) Executive Summary	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Catherine Neubauer, Anthony L Baker, Sean M Fitzhugh, Bret Kellihan, Justin Jagielski, and Andrea S Krausman	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) DEVCOM Army Research Laboratory ATTN: FCDD-RLA-FA Aberdeen Proving Ground, MD 21005	**8. PERFORMING ORGANIZATION REPORT NUMBER** ARL-TR-9622
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES

ORCID IDs: Catherine Neubauer, 0000-0002-6686-3576; Anthony L Baker, 0000-0001-7163-4439; Sean M Fitzhugh, 0000-0002-6283-2895; Bret Kellihan, 0000-0002-7119-8013; Justin Jagielski, 0000-0003-2017-7383; Andrea S Krausman, 0000-0003-1955-8867

14. ABSTRACT

Advances in artificial intelligence capabilities in autonomy-enabled systems and robotics have pushed research to address the unique nature of human-autonomy team collaboration. The goals of these advanced technologies are to enable rapid decision making, enhance situation awareness, promote shared understanding, and improve team dynamics. Simultaneously, use of these technologies is expected to reduce risk to those who collaborate with these systems. Yet, for appropriate human-autonomy teaming to take place, especially as we move beyond dyadic partnerships, proper calibration of team trust is needed to effectively coordinate interactions during high-risk operations. To meet this end, critical measures of team trust for this new dynamic of human-autonomy teams are needed. This report provides an overview of the purpose, components, and functionality of the Human-Autonomy Teaming Trust Toolkit.

15. SUBJECT TERMS

team trust, human-autonomy teaming, Humans in Complex Systems, software development, trust measurement

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					Catherine Neubauer
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER (Include area code)**
Unclassified	Unclassified	Unclassified	UU	28	(954) 258-2287

Contents

List of Figures

List of Tables

1. Introduction

Advanced intelligent technologies will continually change the nature of the battlefield and the very nature of the tasks that Soldiers are required to perform. As such, there have been many discussions on the role of artificial intelligence (AI) in the battlefield, specifically focused on aspects of the mission where AI is most beneficial, the capability that Soldier–AI teams must provide to perform the mission effectively, as well as the necessary adaptation of both the human and the machine during the evolution of this mission. Here, systems must address adaptive, intelligent adversaries that attempt to take advantage of complex environments. Within this context, the concept of trust and trust measurement is crucial to understand. However, understanding the dynamic nature of trust and how to accurately measure and assess it is complex.

With more and more emphasis on integrating humans and autonomous systems within future combat operations, the US Army Combat Capabilities Development Command (DEVCOM) Army Research Laboratory (ARL) established the Human-Autonomy Teaming Essential Research Program (HAT ERP). The goal of the HAT ERP is to address the challenges associated with teaming humans and autonomous systems within a complex tactical environment to create synergistic teams that function effectively and adapt to the dynamic nature of combat. One specific area being addressed in Project 5 of the HAT ERP is how to effectively measure critical team processes such as trust and cohesion. Thus, the overall goal of HAT Project 5 is to develop novel, multimodal metrics of team trust and cohesion to effectively calibrate trust and improve human-autonomy team performance supporting the Next-Generation Combat Vehicle (NGCV). More specific goals of HAT Project 5 include 1) identifying unobtrusive, real-time/near-real-time measures of trust that capture the dynamic nature of team trust; and 2) informing appropriate trust interventions for appropriate calibration of individual and team trust.

Despite the known importance of measuring and evaluating trust in Solider-autonomy interactions, there are still some complexities and considerations for evaluation. The first centers on trust measurement. Trust is a complex construct that has traditionally been somewhat difficult to define and thus measure. For example, there is still work needed to understand the types of trust measurement and the appropriate metrics that should be utilized because not all trust measurement is created equal. While there are existing measures of trust, most of them use self-reporting questionnaires; these provide valuable information, but only at discrete time points. Measurement methods that align with the dynamic nature of trust and allow for more continuous measurement over a specific period of time are needed; thereby, providing more robust information about changes in trust and how it impacts team interactions and performance. Further, as demonstrated by research performed under Project 5 (Krausman et al. 2022), the assessment of human-autonomy team trust must consider the before, during, and after stages of team development and/or teamwork and must include multimodal metrics that go beyond performance (Schaefer et al. 2019; Brewer et al. 2022). See Fig. 1.

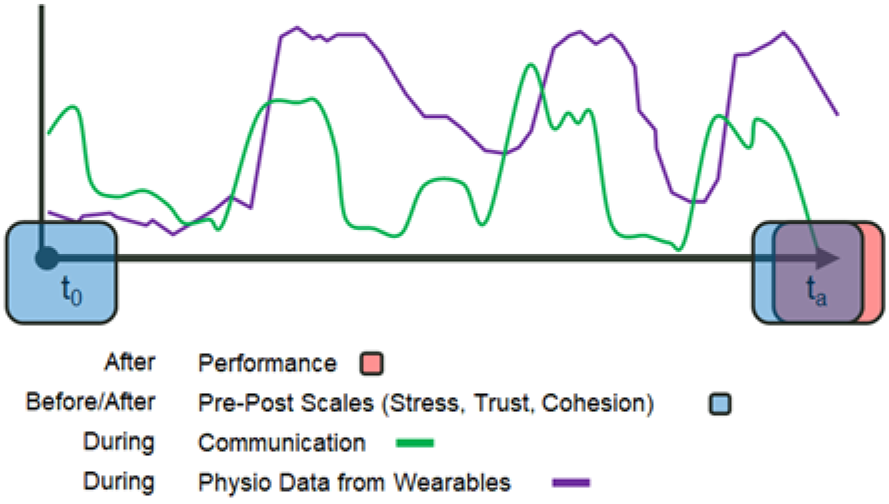


Fig. 1 Multimodal data representation of pre-post subjective state including stress, trust, and cohesion with data streams from communication metrics and physiological data

Given this requirement, and based on literature, laboratory, and field studies, Krausman et al. (2022) developed a conceptual toolkit that consists of novel measures of trust including the following: 1) subjective (i.e., interpersonal trust, technology trust); 2) communication (i.e., communication flow, network dynamics, semantic content analysis); 3) physiological (i.e., heart rate, heart rate variability, and respiration rate); 4) behavioral (eye tracking, interface interaction, etc.); and 5) affective (i.e., facial expression tracking). Recognizing the need for a platform for trust assessment, a multimodal software toolkit of trust measures evolved—the Human-Autonomy Teaming Trust Toolkit (HAT^3).

Section 2 will outline the HAT^3 software development and specific technologies included that are designed to measure trust. Further, each of the modules discussed will be further detailed in subsequent sections and will include an outline of the type of trust measurement, as well as specific metrics that are beneficial to the HAT ERP and the NGCV program.

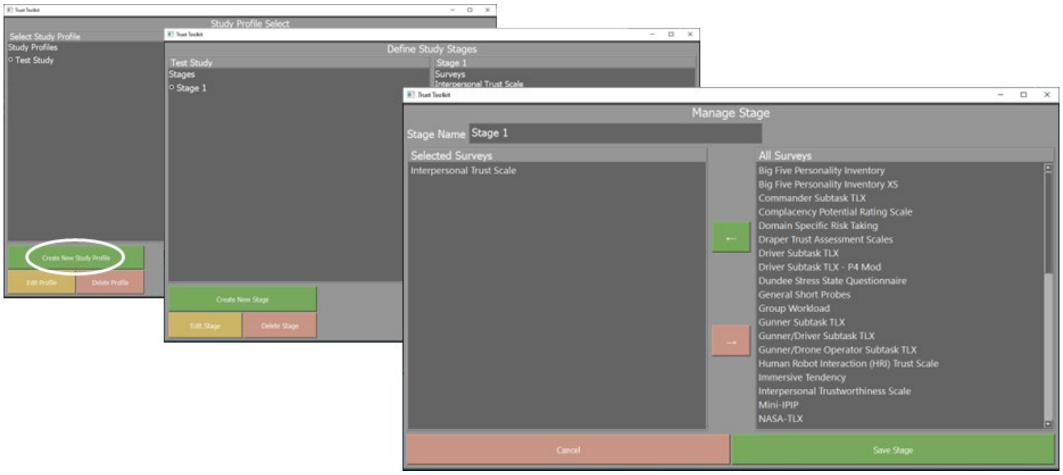
2. Human-Autonomy Teaming Trust Toolkit: Version 1.0

The HAT^3 is a modular software tool for multimodal inference of trust using subjective, communication, behavioral, and physiological indicators. The overall goal of HAT^3 is to provide a real-time, or near-real-time, data reporting, visualization, and prediction of trust-based decisions and actions. Specific capabilities of the toolkit include dynamic, real-time trust measurement from multiple modalities with modular capabilities that are customizable to the user's needs while being stand-alone and capable of mobile use. The target population for HAT^3 includes researchers, analysts, vehicle commanders, and commanding officers.

Section 2.1 will outline the different modules that the HAT^3 software platform currently entail, which include subjective as well as communication modules. Each of these modules has specific metrics relating to trust measurement and will be discussed in more detail. Screenshots of visualizations for each module are provided after the metrics discussion.

2.1 Subjective Module

Before continuing to the subjective module, users will first create a "study profile" in the first page of the toolkit. In the study profile page, users will select the types of measures desired and/or required for their experiment, as well as for each participant and subsequent stage of the experiment (see Fig. 2). This allows the users to predefine study specifics and methodology prior to the experiment. After the profile is created and reviewed, the user can save the profile and task, or the experiment can begin.



Study Profiles
○ Test Study

Test Study
Stages
○ Stage 1

Stage 1
Surveys
Interpersonal Trust Scale

Stage Name: Stage 1

Selected Surveys
Interpersonal Trust Scale

All Surveys
Big Five Personality Inventory
Big Five Personality Inventory XS
Commander Subtask TLX
Complacency Potential Rating Scale
Domain Specific Risk Taking
Draper Trust Assessment Scales
Driver Subtask TLX
Driver Subtask TLX - P4 Mod
Dundee Stress State Questionnaire
General Short Probes
Group Workload
Gunner Subtask TLX
Gunner/Driver Subtask TLX
Gunner/Drone Operator Subtask TLX
Human Robot Interaction (HRI) Trust Scale
Immersive Tendency
Interpersonal Trustworthiness Scale
Mini-IPIP
NASA-TLX

Fig. 2 Screenshot of the "Study Profile" page and subjective metric selection. Subjective module contains ~40 self-report options related to subjective trust. User may select the metrics that are most appropriate for their study needs.

2.1.1 Subjective Module Metrics

The benefits of subjective measurements are widespread and allow researchers to gain valuable insights into the subjective response and changes of specific states at specific points in time. Regarding trust measurement in human-autonomy teams, subjective scales can be differentiated into two main groups and include trust-propensity scales (i.e., an individual's predisposition to trust) and state-based trust scales (i.e., subjective trust report in response to interacting or working with autonomous systems). Table 1 outlines the trust-propensity and state-based trust scales that are currently integrated into the HAT^3 software platform. However, it should also be noted that while trust is the main construct of interest for the toolkit platform, many users may require more than trust-based questionnaires; therefore, other widely used subjective surveys within the HAT literature were incorporated as well (e.g., simulator sickness, subjective stress, and team cohesion).

Table 1 List of trust-propensity and state-based trust scales used in the subjective module of the HAT^3 software platform

	Trust-propensity scales	
Scale name	**Description**	**Citation**
Interpersonal trust scale	25-item scale measuring general propensity to trust people	Rotter 1967
Propensity to trust survey	21-item scale measuring general propensity to trust others and propensity for trustworthiness	Evans and Revelle 2008
Complacency – potential rating scale	20-item propensity to trust automation scale	Singh et al. 1993
Propensity to trust technology	6-item scale measuring an individual's trust in technology	Schneider et al. 2017
	State-based trust scales	
Scale name	**Description**	**Citation**
Integrated model of trust	12-item system trustworthiness scale	Muir and Moray 1996
Checklist of trust between people and automation	12-item system trustworthiness scale	Jian et al. 2000
Human-Robot Interaction (HRI) trust scale	32-item scale measuring trust perception	Yagoda and Gillan 2012
System trustworthiness scale	5-item scale measuring perceived trustworthiness for robotic systems	Schaefer et al. 2012
Trust perception scale-HRI	40-item general trust scale for use with intelligent system	Schaefer 2016
Draper trust assessment scales	7 scales assess visibility of system behavior, probable system behavior, system capabilities/limitations, accessibility of system rationale, awareness of latency and delays, and transparency of failure	Jackson et al. 2016

2.1.2 Subjective Module Data Visualization

One of the benefits of the HAT^3 software platform is real- or near-real-time visualization of the data being collected. The following screenshots show two graphical representations of "dummy" data (Fig. 3). Additionally, one of the goals of the HAT^3 software platform was to visualize data for many different types of users. We acknowledge that some users may not be familiar with these metrics; therefore, we attempted to create visualizations that were intuitive and easily understandable from many perspectives. We decided to represent the subjective trust data via color coding where the colors green, yellow-orange, and red indicate high, medium, and low subjective trust states, respectively. This provides the user with a "quick-look" at their study participants or Soldiers in the field as they work in teams and operate complex technology. However, a future module of HAT^3 will include a trust-based interventions piece, which will contain recommendations for when to intervene if trust is too high or too low. In these cases, the color-coding scheme represented here may need to be adapted to represent these miscalibrated states (i.e., red may indicate trust states that are too high or too low and green may indicate appropriate, calibrated trust states).

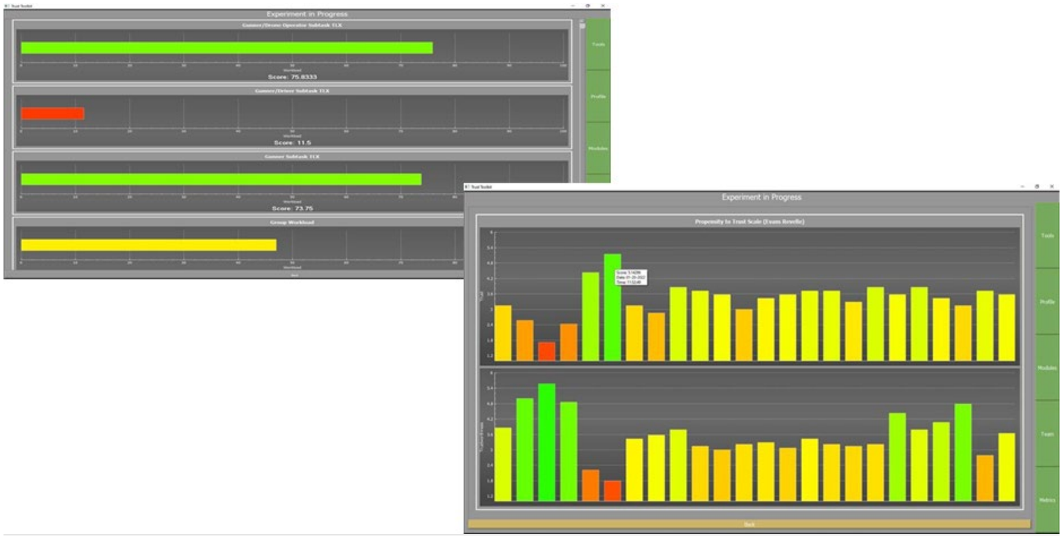


Fig. 3 Historical view of one team member's scores over time. Colors indicate a continuous spectrum of subjective trust states, which ranges from high (green ranges), to medium (yellow–orange ranges), to low (red ranges).

2.2 Communication Module

The second module in the HAT^3 software platform includes streams of data focused on communication metrics. Good communication is the basis for effective teamwork and plays a key role in the success or failure of teams (Salas et al. 2008; Mesmer-Magnus and DeChurch 2009). It enables the core functions of teams such as task coordination, information dissemination, goal and strategy development, and more. By analyzing team communication, we can understand factors such as crew intent (e.g., developing shared situation awareness) or task-related adaptations (e.g., patterns of communication changing to overcome loss of a team member, user display, or autonomy connectivity). Importantly, research in our lab has found that it is possible to infer human-autonomy team dynamics of trust and cohesion from metrics of individual and team communication (Schaefer et al. 2019; Baker et al. 2020, 2021, 2022b). By measuring aspects of when and how a team communicates, we can glean information that may relate to their trust and cohesion behaviors.

Further, the mode of communication plays an important role in determining which types of communication metadata to capture. Communication media such as chat, email, and other written communication will make communication content trivial to collect—enabling near-real-time, or online, content-based analyses. Additionally, typical communication logs (e.g., inbox history) accompanying those communications will provide time stamps and information on communication sender and receiver, which are critical elements for dynamic, network-based analyses. In verbal media such as telephone, face-to-face, or computer-mediated channels (e.g., Teamspeak or Mumble), communication content will only be available through on-the-fly transcription functions such as Dragon Naturally Speaking (i.e., a commercial off-the-shelf speech-recognition software for diction and transcription) (Krausman et al. 2019). The low tested accuracy in prior ARL experimental settings pose challenges, however. Additionally, identifying critical metadata such as sender, receiver, time stamp, and duration will require software-based solutions (i.e., the push-to-talk functionality in TeamSpeak). Regardless of medium, once the relevant communication-based data are captured by the system and stored, they may then be processed and provided to the user for review and analysis. For the communication module, specific plans include utilizing several different types of in-house communication measures and visualizations to indicate trust.

Section 2.2.1 will outline the different communication analysis capabilities that have been integrated into the HAT^3 system. Each section will cover the same types of information (overview and data visualization, etc.) for each of the different communication capabilities. Note that data for the communication visualizations were collected from a vehicle crew of seven members during a simulation experiment. Each crew station is labeled (cs01–cs07), and each crew member performed a specific role (Commander, Gunner, or Driver).

2.2.1 Communication Module Metrics

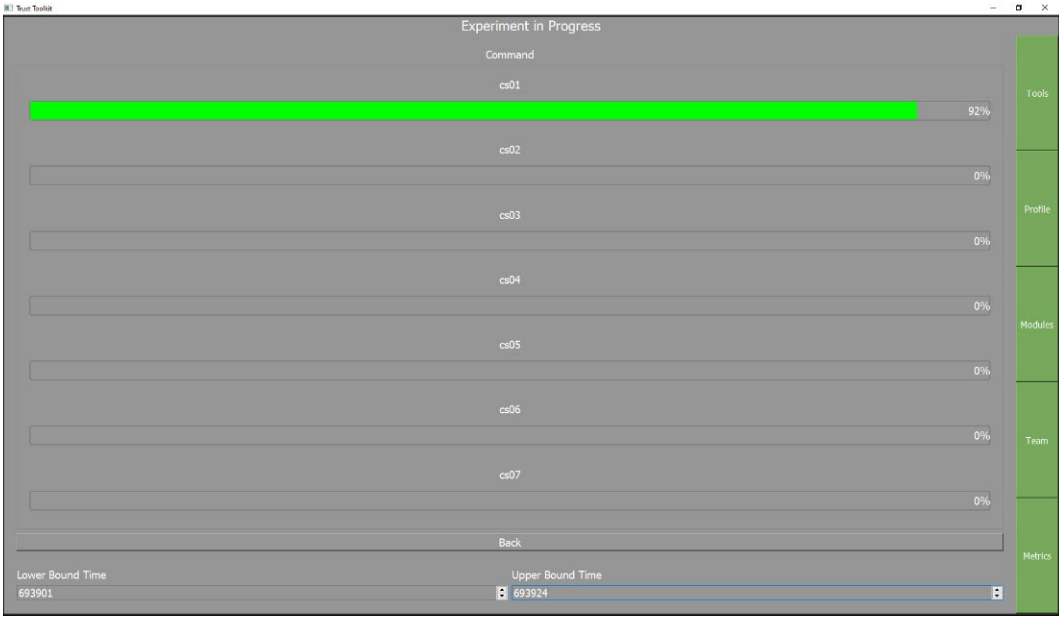
The following subsections outline specific tools used to capture communication data.

2.2.2 Communication Flow: Real-Time Event, Flow, and Coordination Tool (REFLECT)

It is not possible to understate the importance of a communication log that is accurate, organized, and stored in a broadly readable format (or multiple formats). Knowing that A spoke to B at time t_1, and B spoke to C at time t_2, and so on, allows for a play-by-play understanding of how the crew interacted throughout a mission. REFLECT is aimed at supporting this concept. Ideally the tool will allow flexibility with how to display communication data as a function of the metadata described in Section 2.2.1. The system could display, for example, communication networks during a particular time window, the communication network of a particular sender or receiver, or a communication network composed of messages containing the word "agent." In any case, the goal of REFLECT is to visually represent the flow of communication among a given team, regardless of the number of crew members or communication networks.

2.2.3 REFLECT Visualization

The REFLECT tool consists of an information display to represent the various team members' amount of communication in real-time. The interface shows the number of messages initiated by each crew member, and a different page of the interface visualizes how many messages each communication network is receiving. For example, the user will be able to determine the relative proportion of time each team member was speaking, as well as the distribution of those messages onto various networks (e.g., a communication net just for crew members in each vehicle, or a net spanning several vehicles). This allows for a clear, overall understanding of how the crew is communicating in general, and if there are any issues or irregularities in the expected patterns (e.g., if a Commander is sending very few messages to their in-vehicle network during a high-stress event). To this end, the following screenshots of the latest REFLECT prototype provide an overview of the visualizations of this tool (Figs. 4–6).



Command

cs01 92%

cs02 0%

cs03 0%

cs04 0%

cs05 0%

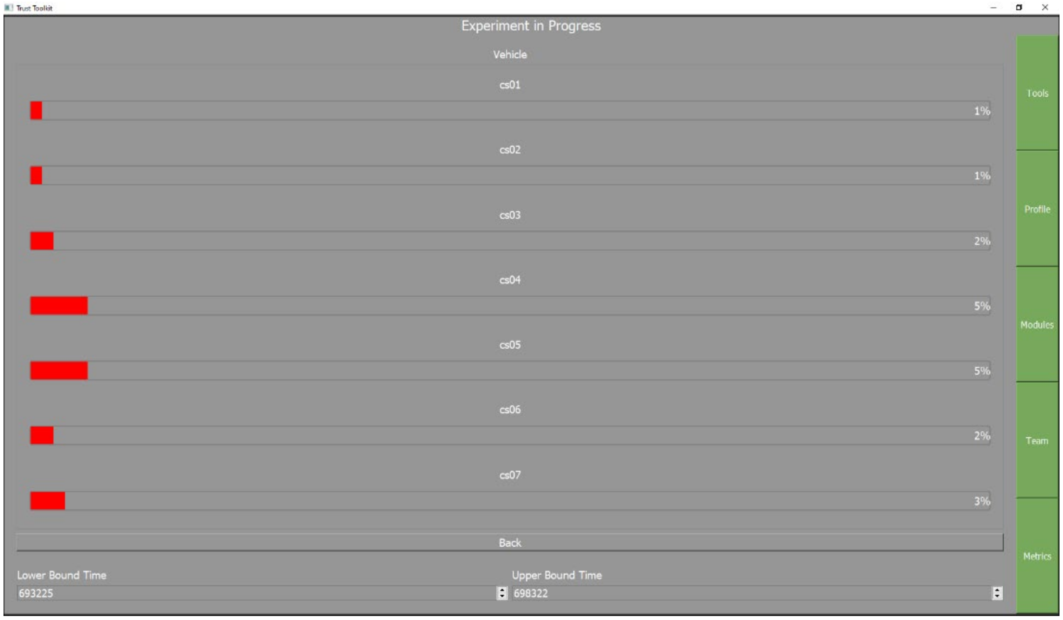
cs06 0%

cs07 0%

Lower Bound Time 693901

Upper Bound Time 693924

Fig. 4 Depiction of individual crew member communication events, using simulated data. These represent communications on a Command network. In this screenshot, we note that between the time markers of 693901 and 693924 (a span of 23 s), the user labeled "cs01" was responsible for 92% of the team's communication on the Command network.



Vehicle

cs01 1%

cs02 1%

cs03 2%

cs04 5%

cs05 5%

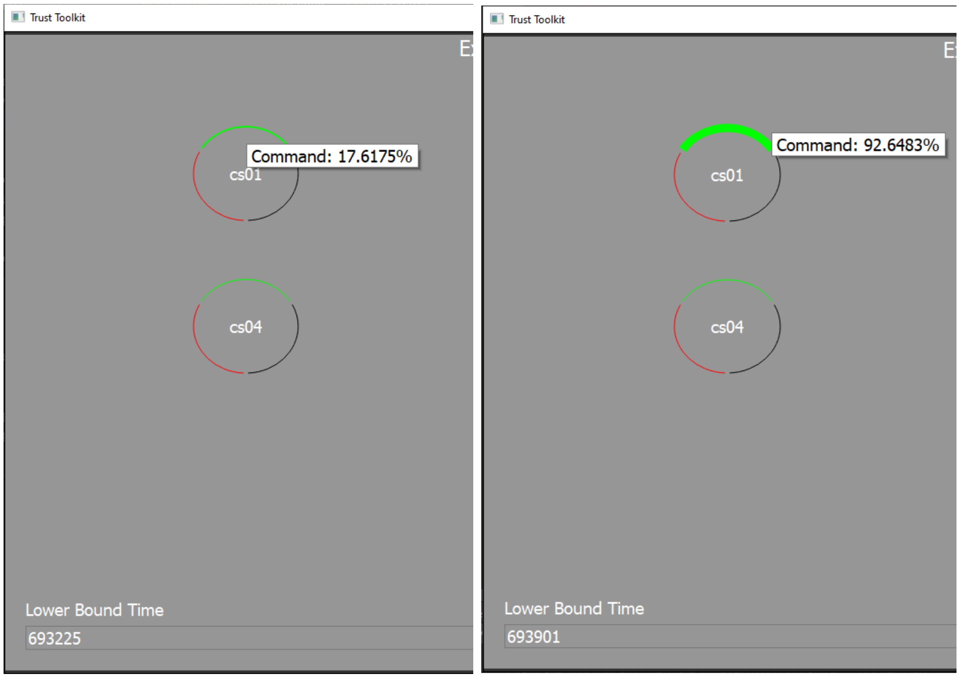
cs06 2%

cs07 3%

Lower Bound Time 693225

Upper Bound Time 698322

Fig. 5 Crew member communication events using a different time window in simulated data. These represent communications on a Vehicle network. Between time markers 693225 and 698322 (a span of 84 min), crew communication rates are somewhat balanced; however, cs04 and cs05 account for relatively more communication events than their peers. These crew members are likely more closely associated with the team's core tasks (the gunners in a security or area defense task, the mobility operators in a navigation task, etc.).



Lower Bound Time
693225

Lower Bound Time
693901

Fig 6 Two visual overviews of communication rates for a selection of crew members. The thickness of a given line indicates the amount of communication on a specific network. Compare the thickness of the top green line (representing communications on the Command network) for cs01 in the left figure with the same line for a different time horizon in the right figure, along with the percentage of communication events represented by each.

3. Network Analysis

Trust has been linked to information flow in teams (Hung and Gatica-Perez 2010; Tiferes et al. 2016; Baker et al. 2020). Regarding information flow, the Network Analysis tool from the communication module of HAT[3] represents team interactions as a network, visualizes that network, and provides descriptive statistics of the network. This provides quantitative measures of the network that may be related to trust within the team. For example, the graphic represented in Fig. 7 shows four color-coded teams with lines that indicate interactions between individuals. Each individual speaker is represented by a node, whose size reflects its betweenness-centrality score. This measure of centrality captures the tendency for nodes to occupy positions in shortest paths between other nodes. In practice, higher betweenness centrality reflects a node's potential importance for routing messages through the network. Owing to their higher betweenness-centrality scores, larger nodes (i.e., more important for information flow) need high levels of trust and need to be trusted by team members if information is to flow effectively within and across teams. This type of visualization is a promising indicator for trust interventions as it clearly demonstrates that interventions should target and prioritize the larger nodes.

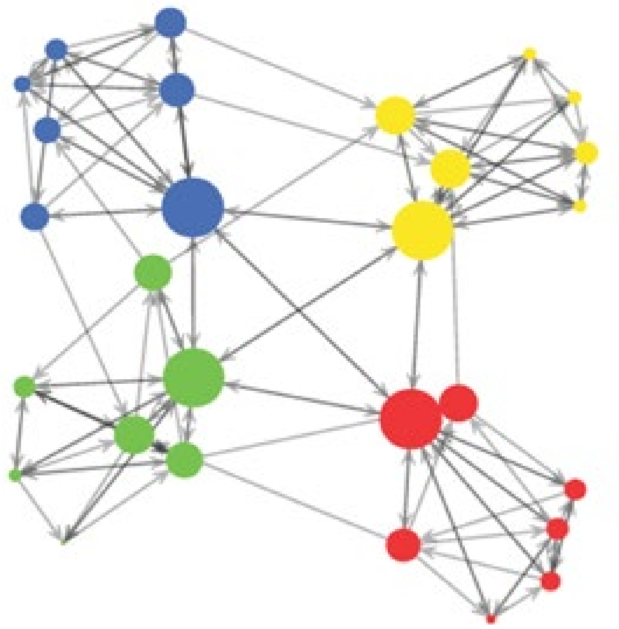


Fig. 7 An illustrative example featuring four color-coded teams. Circles represent individuals and lines between them represent communication ties. The size of each node reflects its betweenness centrality, or the tendency to occupy positions on shortest paths between nodes.

While the visualization approach has considerable overlap with REFLECT—in that we aim to capture the sender, receiver, and time stamp for each communication event—this tool also differs in a few key ways. These differences principally arise due to the focus on statistical measures of the communication network. While REFLECT is aimed at providing a clear but simple overview of the crew's communication, the Network Analysis tool will provide more in-depth analyses and useful statistics relating to crew communication behaviors. These statistics can inform the visualization of the network such that the size of individual nodes may reflect their centrality scores, for example. In the following we highlight several key statistics we aim to capture and represent in the Network Analysis tool.

1) Individual-level measures

   a. Degree centrality: Degree centrality captures the total number of ties an individual has in the network. This can be separately represented as in-degree centrality (the total number of incoming ties), and out-degree centrality (the total number of outgoing ties). The degree centrality measure captures the volume of interactions a given node has within the network. Individuals with higher degree centrality can directly reach more individuals in the network.

b. Betweenness centrality: Betweenness centrality captures the extent to which a given node sits on a large number of shortest paths between other nodes in the network. Higher betweenness-centrality scores indicate that a node can be a critical hub for routing messages between other nodes in the graph. High-betweenness individuals can play a crucial role in routing information through the network.

c. Closeness centrality: Closeness centrality measures the extent to which a node can easily reach other nodes in the graph. Higher closeness scores demonstrate that a message from a given node needs to pass through fewer intermediaries to read any other node in the network. Such individuals are best positioned to reach many individuals in the network quickly.

2) Network-level measures

a. Reciprocity: This statistic captures the extent of symmetry among all possible pairs of nodes in the network. If every observed interaction is reciprocated (i.e., if every tie from *i* to *j* is reciprocated by a tie from *j* to *i)*, the network will have a maximal reciprocity score. Likewise, networks with much lower levels of reciprocal ties will have low-reciprocity scores.

b. Clustering coefficient: The global clustering coefficient reflects the proportion of closed triads in the network (e.g., A has a tie to B, B has a tie to C, and a tie between A and C completes the triad). This measure reflects the amount of clustering in the graph, or the extent to which individuals form tight groups.

c. Degree centralization: Unlike degree centrality, an individual measure capturing one’s number of connections, degree centralization is a network-level measure capturing the distribution of individual-level centrality. Higher centralization scores indicate that centrality is more concentrated (i.e., centrality is concentrated within a handful of nodes while the network is populated by several less-central nodes). Lower centralization scores indicate that centrality is more evenly dispersed (i.e., most nodes in the network have similar centrality scores). Degree centralization scores specifically demonstrate whether connectivity is focused on a handful of individuals or whether it is more equally spread across the network.

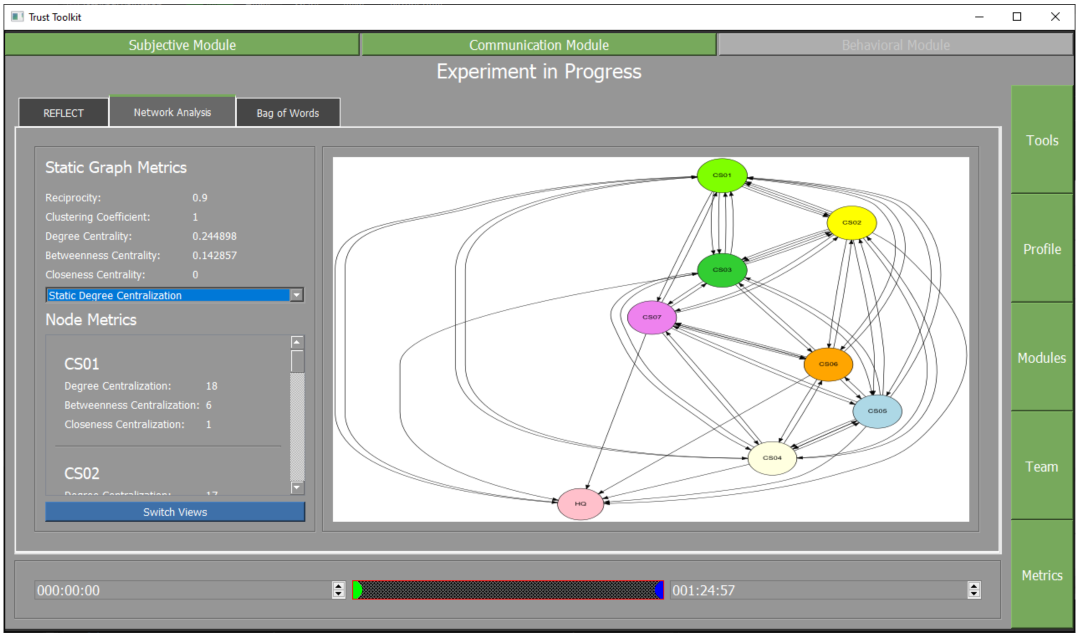
d. Betweenness centralization: Betweenness centralization measures the concentration of betweenness centrality in the network. Higher scores indicate that a small number of nodes occupy essential positions for routing messages through the network while lower scores indicate that messages can more easily traverse the network without relying on a small number of central individuals.

e. Closeness centralization: Closeness centralization captures the concentration of closeness centrality on a small number of nodes in the network. Networks higher in closeness centralization have a relatively small number of individuals with close access to many nodes in the network while most others are more distantly connected (i.e., traveling from one node all others may require many "hops"). Networks with lower closeness centralization have a more equitable distribution of closeness centrality such that most nodes are relatively equidistant from each other (i.e., the distance from one node to any other node in the network is similar).

Like REFLECT, the Network Analysis tool will ultimately be capable of being deployed either as a stand-alone platform capable of operating independently of other systems or, rather, as a system that can be integrated into other HAT systems. The capability for both of these options is preferred to allow fairly broad usage across a number of scenarios and use-cases; however, the implications of the data access discussion for operation of a stand-alone system will need to be considered. We see many options (e.g., a separate database function could be developed for selective application, or the stand-alone version may need to ingest configuration files that provide the history-based calibration information). This type of modularity is important as the capability to allow users to customize the trust-inference method, models, and inputs based on their experimental needs and/or scientific preferences regarding trust metrics and their applications for different purposes (subjective and/or communication modalities, etc.) that are most applicable to their individual use-cases.

3.1 Network Analysis Data Visualization

The Network Analysis tool has a GUI, or visual output to display information as it changes in real-time (e.g., multi-screens, multi-data inputs; see Fig. 8). This GUI visualizes the team communication network with an adjustable sliding window to show how the network is changing over time (e.g., What did the network look like over the last 15 min? Over the last hour?). Visual attributes of individual nodes, such as size or color, can be modified in real time to reflect centrality scores. For example, nodes with higher betweenness-centrality scores can be enlarged relative to nodes with lower betweenness scores.



Static Graph Metrics

Reciprocity: 0.9
Clustering Coefficient: 1
Degree Centrality: 0.244898
Betweenness Centrality: 0.142857
Closeness Centrality: 0

Static Degree Centralization

Node Metrics

CS01
Degree Centralization: 18
Betweenness Centralization: 6
Closeness Centralization: 1

CS02

000:00:00
001:24:57

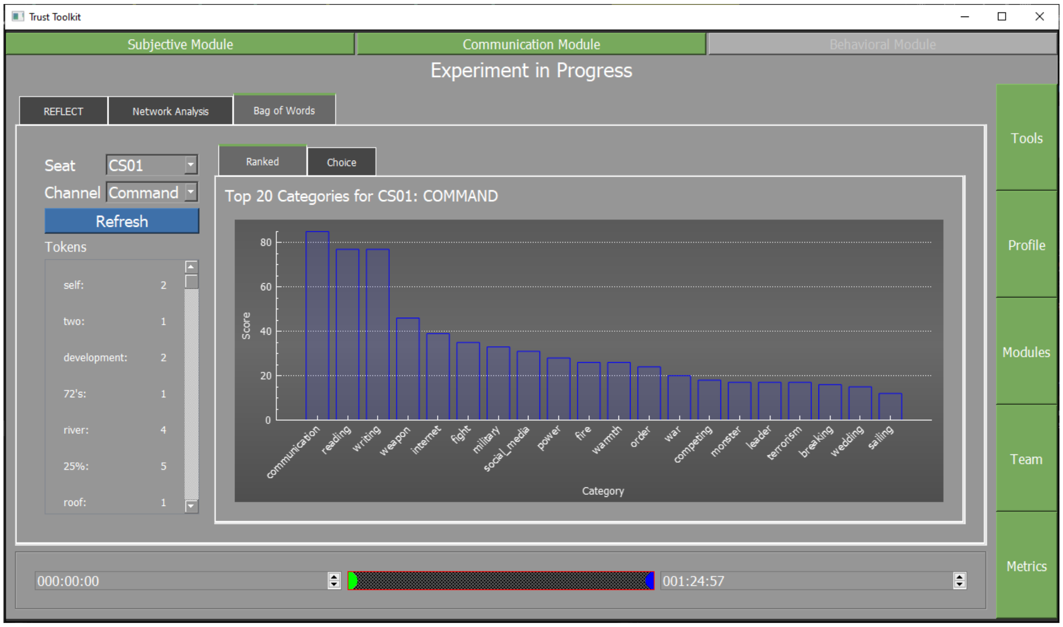
Fig. 8 Example network analysis data visualization. Node size and color can be used to highlight features such as individual attributes (role, team membership) and network attributes (centrality scores, etc.). Additional details on the left provide individual- and network-level statistics.

Semantic Content: Bag of Words

Finally, a communication tool known as “Bag of Words”, which analyzes semantic content, has been added. This will allow the user to analyze all text in a data set, then evaluate the number of words associated with different semantic word categories (positive emotions, causation, perception/cognition, past tense, etc.). Unusual values in categories could suggest trust or cohesion issues, such as increases in anger or swear word categories.

3.2 Bag of Words Data Visualization

The Bag of Words communication tool has a visual display that presents specific word categories with which the speech used between and among participants was categorized (Fig. 9). The communication and speech used can be automatically transcribed with specific words falling into various meaningful word categories. There are many possible categories that may or may not be relevant to the specific task at hand; therefore, only the top categories for speech content are presented to the user to determine what is being said among participants.



Seat CS01

Channel Command

Tokens

self:	2
two:	1
development:	2
72's:	1
river:	4
25%:	5
roof:	1

000:00:00

001:24:57

Fig. 9 Example Bag of Words data visualization. The top number of word categories that were categorized from speech used during communication between participants are presented to the user.

4. Future Plans

The development of the HAT^3 software platform is ongoing with plans to incorporate several more modules and elements, thereby extending the capabilities of the system. First, the "VocStr" emotion detection model will be added to the communication module. Specifically, VocStr detects the degree of stress or cognitive load using acoustical features of speech. This is represented as a value between 0 and 1 and is visualized by the color of the status frame associated with that speech signal. This measure of load provides an indicator of increased task stress and potential opportunities for autonomous assistance to help alleviate the load. VocStr works in noisy backgrounds, can be used with a single talker or multiple speakers, and requires no additional sensors beyond the communications headset (Scharine 2021). Further, the VocStr value may also indicate times when interventions are beneficial.

A third module for the Trust Toolkit is also being developed. It will contain physiological indicators of trust including heart rate variability, pupil dilation, and electrodermal activity (Neubauer et al. 2020b). Additional modules relating to behavioral data (e.g., eye tracking, interface interactions) and affective cues (e.g., facial expressions) (Neubauer et al. 2020a) are also planned.

One of the key goals of the project and software development is to be able to further validate these measures and synthesize them into metrics via algorithms or data fusion. Specifically, communication flow, communication rates, physiological measures, entrainment, and facial-analysis methods all show promise as direct or proxy measures of trust, but they have yet to be fully validated. Next, we intend to build on this measurement research to quantify appropriate metrics of team trust. While the measures provide values for specific constructs, metrics will provide a more standardized assessment strategy for evaluating team trust in the HAT context. In addition, it is necessary to understand the causal structure that underlies the relationships between the construct(s) of team trust and their measures, which will inform effective strategies for trust-based interventions for appropriate trust calibration (Baker et al. 2022b).

The toolkit can also be integrated into other ARL platforms such as the Human Research and Engineering Directorate's Information for Mixed-Squads (INFORMS) laboratory and is capable of data synchronization utilizing stream-processing platforms such as Lab Streaming Layer or Apache Kafka. This will provide on-the-fly and/or real-time data collection, storage, and analysis with visualization capabilities to monitor, predict, and suggest trust-based interventions for human-agent teams.

Finally, the development team plans to expand the toolkit's capabilities beyond trust to also include team cohesion measures as well as system-performance data (i.e., autonomy) and other types of human-performance measures and latent states of interest (situation awareness, workload, dynamic resource allocation, stress, etc.) to arrive at a more comprehensive understanding of the team and to identify areas where interventions may enhance team effectiveness.

5. Conclusions

Overall, measurement of trust in human-autonomy teams remains a complex problem and, as a result, there is no one-size-fits-all approach. Rather, researchers must continually assess which types of measures are best suited to the context, recognizing that different measures will have a stronger impact on appropriately quantifying trust at a given time. Therefore, for many applications, a multi-method, multimodal approach is warranted (see also Schaefer et al. 2019, 2021; Milner et al. 2020). Continued research building on this toolkit will support the development of more appropriate metrics of team trust, which will help us understand how human-autonomy teams perform—especially as autonomous capabilities increase—and identify when interventions are needed. Specifically, these interventions can be directed toward several possible changes in the team operations: training recommendations, changes in autonomy behavior, implementation of algorithmic assurances in the autonomy, improving communication and transparency elements, or even supporting after-action reviews. To that end, the HAT^3 will help us assess team interactions in near-real-time, and through algorithm creation and visualization techniques will be able to observe changes in trust over time and identify areas where an intervention is warranted. Although still early in the development process, this technology, coupled with the research presented here, will enable researchers to develop more precise, valid measures of trust, and deploy effective, trust-enhancing interventions in practical settings.

6. References

Baker AL, Brewer RW, Schaefer KE. Development and usability assessment of the realtime event, flow, and coordination tool (REFLECT). CCDC Army Research Laboratory (US); 2020. Report No.: ARL-TR-9012.

Baker AL, Fitzhugh SM, Forster DE, Schaefer KE. Communication metrics for human-autonomy teaming: lessons learned from US Army gunnery field experiments. Proc Hum Factors Ergon Soc Annu Meet. 2022a;65(1):1157–1161.

Baker AL, Fitzhugh SM, Huang L, Forster DE, Scharine A, Neubauer C, Lematta G, Bhatti S, Johnson CJ, Krausman A, et al. Approaches for assessing communication in human-autonomy teams. Hum Int Syst Integrat. 2021;3(2):99–128. doi: 10.1007/s42454-021-00026-2.

Baker AL, Forster DE, Reichnberg RE, Neubauer CE, Fitzhugh SM, Krausman A. Toward a causal modeling approach for trust-based interventions in human-autonomy teams. AAAI Springer. Forthcoming 2022b.

Brewer R, Baker A, Neubauer C, Krausman A, Forster D, Scharine A, Berg S, Davis K, Schaefer K. Evaluation of human-autonomy team trust for weaponized robotic combat vehicles. In: Wright J, Barber D, editors. Human Factors and Simulation. AHFE (2022) International Conference; 2022. AHFE Open Access. vol 30. AHFE International, USA. doi: 10.54941/ahfe1001491.

Evans AM, Revelle W. Survey and behavioral measurements of interpersonal trust. J Res Pers. 2008;42(6);1585–1593. doi: 10.1016/j.jrp.2008.07.011.

Hung H, Gatica-Perez D. Estimating cohesion in small groups using audio-visual nonverbal behavior. IEEE Trans Multimed. 2010;12(6):563–575.

Jackson KF, Prasov Z, Vincent EC, Jones EM. Draper trust assessment framework - trust assessment scales. Draper; 2016.

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

Krausman A, Kelley T, McGhee S, Schaefer KE, Fitzhugh S. Using Dragon for speech-to-text transcription in support of human-autonomy teaming research. CCDC Army Research Laboratory (US); 2019. Report No.: ARL-TN-0978.

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

Mesmer-Magnus JR, Dechurch LA. Information sharing and team performance: a meta-analysis. J Appl Psychol. 2009;94:535–546. doi: 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; 2021. vol 1206. Springer, Cham. doi: 10.1007/978-3-030-51064-0_12.

Muir BM, Moray N. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. Ergonomics. 1996;39(3):429–460. doi: 10.1080/00140139608964474.

Neubauer C, Gremillion G, Perelman B, La Fleur C, Metcalfe J, Schaefer KE. Analysis of facial expressions: explaining affective state and trust-based decisions during interaction with automation. CCDC Army Research Laboratory (US); 2020a. Report No.: ARL-TR-8945.

Neubauer C, Schaefer KE, Oiknine A, Thurman S, Files B, Gordon S, Bradford C, Spangler D, Gremillion F. Multimodal physiological and behavioral measures to estimate human states and decisions for improved human autonomy teaming. CCDC Army Research Laboratory (US); 2020b. Report No.: ARL-TR-9070.

Rotter JB. A new scale for the measurement of interpersonal trust. J Pers. 1967;35(4):651–665.

Salas E, Wilson KA, Murphy CE, King H, Salisbury M. Communicating, coordinating, and cooperating when lives depend on it: tips for teamwork. JT Comm J Qual Patient Saf. 2008;34:333–341. doi: 10.1016/S1553-7250(08)34042-2.

Schaefer KE, Baker AL, Brewer RW, Patton D, Canady J, Metcalfe JS. Assessing multi-agent human-autonomy teams: US Army robotic wingman gunnery operations. Proceedings of the SPIE Defense + Commercial Sensing, Micro- and Nanotechnology Sensors, Systems, and Applications XI Conference; 2019 May; Baltimore, MD.

Schaefer KE, Chen JYC, Szalma JL, Hancock PA. A meta-analysis of factors influencing the development of trust in automation: implications for understanding autonomy in future systems. Human Factors.2016;58:377–400.

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. 2021. Chapter 12, Elsevier; p. 261–300. doi: 10.1016/B978-0-12-819472-0.00012-5.

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; 1548–1552.

Scharine A. Development of a neural network algorithm to detect Soldier load from environmental speech. In: Wright JL, Barber D, Scataglini S, Rajulu SL, editors. Advances in Simulation and Digital Human Modeling. AHFE 2021; 2021. Lecture Notes in Networks and Systems. vol 264. Springer, Cham. doi: 10.1007/978-3-030-79763-8_7.

Schneider TR, Jessup SA, Stokes C, Rivers S, Lohani M, McCoy M. The influence of trust propensity on behavioral trust. Proceedings of the Meeting of Association for Psychological Society; 2017; Boston, MA.

Singh IL, Molloy R, Parasuraman R. Automation-induced "complacency": development of the complacency-potential rating scale. Int J Aviat Psychol. 1993;3:111–122.

Tiferes J, Hussein AA, Bisantz A, Kozlowski JD, Sharif MA, Winder NM, Ahmad N, Allers J, Cavuoto L, Guru KA. The loud surgeon behind the console: understanding team activities during robot-assisted surgery. J Surg Educ. 2016;73(3):504–512.

Yagoda R, Gillan DJ. You want me to trust a ROBOT? The development of a human–robot interaction trust scale. Int J Soc Robot. 2012;4(3):235–248.

List of Symbols, Abbreviations, and Acronyms

AI	artificial intelligence
ARL	Army Research Laboratory
DEVCOM	US Army Combat Capabilities Development Command
GUI	graphical user interface
HAT	Human-Autonomy Teaming
HAT^3	Human-Autonomy Teaming Trust Toolkit
HAT ERP	Human-Autonomy Teaming Essential Research Program
HRI	Human-Robot Interaction
INFORMS	Information for Mixed-Squads
NGCV	Next-Generation Combat Vehicle
REFLECT	Real-Time Event, Flow, and Coordination Tool

1 DEFENSE TECHNICAL
(PDF) INFORMATION CTR
DTIC OCA

1 DEVCOM ARL
(PDF) FCDD RLB CI
TECH LIB

4 DEVCOM ARL
(PDF) FCDD RLA FA
C NEUBAUER
S FITZHUGH
FCDD RLA FD
A BAKER
A KRAUSMAN