ARL-TR-9623 ● JAN 2023

Human-Autonomy Teaming Trust Toolkit (HAT^3) Software Development Documentation and User Guide

by Catherine Neubauer, Sean M Fitzhugh, Anthony L Baker, 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-9623 ● JAN 2023

Human-Autonomy Teaming (HAT^3) Trust Toolkit Software Development Documentation and User Guide

Catherine Neubauer, Sean M Fitzhugh, Anthony L Baker, 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)* January 2023	**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) Software Development Documentation and User Guide

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Catherine Neubauer, Sean M Fitzhugh, Anthony L Baker, 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-RLH-FD
Aberdeen Proving Ground, MD 21005

8. PERFORMING ORGANIZATION REPORT NUMBER

ARL-TR-9623

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; Sean M Fitzhugh, 0000-0002-6283-2895; Anthony L Baker 0000-0001-7163-4439; 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 software document and user guide describe the purpose, scope, components, and system design of the Human-Autonomy Teaming Trust Toolkit. The objective of this technical report is to provide readers, developers, and potential first-time users of this system with background information, highlighting the impetus for the toolkit, its purpose, and a thorough understanding of its functionality. Lastly, we include an instructional guide for how to use the toolkit software.

15. SUBJECT TERMS

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

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	69	(954) 258-2287

Contents

List of Figures

List of Tables

1. Introduction and Background

Trust is understood to be critical for team functioning; trust facilitates various team processes such as information sharing, collaborative decision making, and overall team success (Grossman and Feitosa 2018). A robust literature exists with respect to trust in human teams; however, little is currently known about trust in human-autonomy teams that consist of one or more human teammates working interdependently with one or more autonomous systems or intelligent agents (IAs), to accomplish a task, goal, or a sequence thereof (Demir et al. 2019). In both human and human-autonomy teams, inappropriate levels of trust (e.g., too high or too low) can have serious implications for team functioning and outcomes. Proper calibration of team trust is needed, which forms as team members interact and work together over time (Fulmer and Gelfand 2012; Costa and Anderson 2017). However, trust is dynamic and fragile, so even though individuals or teams may have developed a level of trust, given the right circumstances (e.g., stress, vulnerability, risk) it can change, thereby impacting team performance. Therefore, being able to measure the dynamic and emergent nature of trust, as well as trust-based decisions and actions, will enable researchers to understand the factors that influence change, and when necessary, prescribe interventions that enable effective calibration of trust.

The specific focus of this technical report is to detail the capabilities of a software toolkit that was developed as a novel approach to measuring trust in human-autonomy teams. Traditional methods of trust assessment in human teams are typically accomplished through subjective measures. However, given the dynamic nature of military and civilian operations involving IAs and autonomy, it is necessary to expand trust measurement to incorporate methods that can capture trust flow and changes over an entire task or mission. This suggests a need for a multimodal measurement capability that can be linked to specific scenario events. For example, in a survey of the literature on trust in human teams, Krausman et al. (2022) identified several methods, in addition to subjective measures, which may be well-suited for inferring trust in teams that are composed of multiple humans and multiple autonomous systems. These measures include communication analyses that focus on communication flow and content, behavioral measures such as gestures and facial expressions, and physiological measures like heart rate and eye tracking. The advantage of using these novel measures is that they provide a rather unobtrusive, real-time measurement capability that allows researchers to capture data continuously, if desired, throughout the course of a mission.

With technology advancing at breakneck speeds, the desire to team humans and autonomous systems will only increase. To ensure productive interactions among team members, and successful team outcomes, trust must be a key consideration. This report describes the Human-Autonomy Teaming Trust Toolkit (HAT^3), a software platform to be used for assessing trust in human and human-autonomy teams using a multimodal approach that captures data in real- or near-real-time to provide a more comprehensive measure of trust. Following an iterative design approach, the initial Toolkit version 1.0 contains two modules: subjective and communications. The subjective module consists of 10 trust surveys, and the communications module provides 3 measurement methods that enable users to analyze crew communications, including communication flow, network analysis approaches, and semantic content with researchers as the initial target users. In the longer term, the toolkit will be expanded to include physiological and behavioral measures and will be fully customizable for different applications; thereby, increasing the target user base and opportunities for usage across the entire life cycle.

1.1 Purpose and Intended Use

The purpose and core functionality of the toolkit is to administer measurement tools (both established and nascent) that support making inferences about trust and trust-related behaviors within a heterogeneous human-agent team, using a multilevel approach (e.g., individual, dyads, sub-teams, and team-level). The tool is envisioned to achieve this goal through modular, multimodal measurements (e.g., self-report and similar subjective instruments, communication and performance monitoring, and physiological sensing) as available based on research questions, experiment needs, or test characteristics and instrumentation.

The HAT^3 system is a tool designed to enable the 1) collection, 2) storage, 3) analysis, 4) visualization, and 5) distribution of individual and team-level, multimodal trust data and related inferences. The system is a result of the efforts of members of the Human-Autonomy Teaming Essential Research Program (HAT ERP) as well as our software developers from DCS Corporation.

Finally, we see the toolkit as a part of a larger suite of assessment tools being developed under HAT ERP that include the following: the Dashboard and the Global After-Action Review Technology (GAART). Each tool provides the capability to assess Soldier-autonomy teams in complex operational environments.

1.2 Specific Benefits of the Toolkit

The toolkit provides a dynamic, near-real-time, capability for multimodal measurement and inference of trust-related phenomena in complex human-autonomy teams, which currently does not exist. Through a process of syncing measurement data with scenario events and/or experimental stimuli, users can identify instances when trust changes. In so doing, researchers can identify what led to the change—when it breaks down and/or when it strengthens, so that appropriate measures can be taken when necessary (e.g., do nothing or suggest intervention).

The toolkit also provides analysis capabilities and associated visualizations for each measurement type, which are customizable to help the user understand the dynamic nature of trust over time at different levels within the team (from individual to whole team). This capability allows the user to identify a specific point(s) in the measurement to further explore and understand which scenario may be associated with changes in trust, either as a cause or consequence.

Recognizing the need to satisfy a wide user base, the toolkit was designed to provide a modular capability, which allows the user to select what measures and modalities are most appropriate for their experimental or mission purposes, given the constraints of their situation (e.g., some users will not be interested in or able to capture physiological data modalities). Thus, the toolkit can receive and process multiple data streams, which enable multimodal trust measurement that is customizable by users across a variety of roles.

The toolkit can be used in virtually any setting with a wireless network connection (e.g., laptop/desktop, tablet, phone). While the toolkit has an initial design focus for use in Next-Generation Combat Vehicle contexts (e.g., simulation studies) and for specific customers such as those administering the Crew Optimization and Augmentation Technologies program at the Ground Vehicle Systems Center, the broader intention is for the software to ultimately be generalizable to other teaming and task domains (experimental; test, evaluation, verification, and validation [TEV&V]; training; etc.).

1.3 Intended Audience (Users)

The intention for this toolkit is to ultimately be useful throughout the life cycle/acquisition cycle, supporting basic and applied research, TEV&V, and training evaluation. That said, our intended toolkit users are anyone involved in the life cycle or activities of an experiment or mission; namely, researcher, test engineer, trainer, commander, and Soldier. However, in version 1.0, we are primarily focused on researchers and scientists who wish to collect individual and team-level data within human and HAT scenarios, either in a laboratory or real-world, operational setting. In future iterations of the toolkit, we anticipate additional functionality that will help further expand the user base.

2. HAT^3 System Overview

The remainder of this report briefly overviews the HAT^3 Software System, describes the system architecture and requirements, outlines the design of the data structures and storage, and describes the interface and interaction designs. An accompanying user guide is also included (see Appendix A).

2.1 Toolkit Capabilities

The following subsections detail the toolkit's capabilities and features. The toolkit can be used as a stand-alone tool and to expand the capabilities of other HAT tools (e.g., GAART and the Dashboard), all of which provide valuable information about the team and/or vehicle states, behaviors, and overall mission performance. The various components of the toolkit provide a visual, interactive capability along with a summary or output of the trust variables of interest. For example, version 1.0 analyzes and provides outputs for subjective measures and communication measures. The system also generates data logs of these variables for use in after-action reviews. HAT^3 returns data and visualizations as they happen, at a rate commensurate with the data type (e.g., subjective ratings may not change very often since they may rely on self-reported information), whereas sender–receiver information and related visualizations should be updated in real-time as the data becomes available to the system.

2.1.1 Dynamic, Online, and On-Demand Trust Assessments

Recognizing the complex nature of trust and how it emerges and propagates in the team, the toolkit provides a capability for multilevel assessments including individual-to-individual, individual-to-team, and overall team-level trust (aggregated across individuals). Teams and sub-teams can be flexibly defined (vehicle-based, role-based, task-based, etc.) and allow individuals to be simultaneously part of multiple sub-teams. In so doing, the on-demand trust assessments enable users to identify/visualize the current state of individuals, dyads, or teams and make judgments as to whether the level of trust is appropriately calibrated, too high, or too low. In addition to the temporal binning described, filtering the data by predefined role or team will enable users to observe specific sub-teams within the group, which allows cross-group comparisons (e.g., What does trust in group X look like in comparison to group Y?). Combining the temporal and group-level sub-setting previously described enables comparison of a group to itself. For instance, if Team X had high trust at time t_1 and low trust at time t_2, then comparison of attributes (i.e., physiological, psychological, communication) within the team can be made.

2.1.2 Online, On-Demand Administration and Assessment of Subjective Measures

Toolkit users can easily deploy subjective measures (i.e., questionnaires) to participants as needed by creating a study profile. In the profile, the user preselects the subjective measures that will be administered, identifies the participants to whom they will be administered, and the time at which they will be given (e.g., at specific time points or as specified before and/or after a mission). Depending on the form factor needed (tablet, mobile device, desktop computer, etc.), the participants will then be able to complete the subjective measures at either prescribed, as needed, or emergent time periods.

2.1.3 Graphic Visualization of Subjective Measures

Keeping with intuitive design principles for data visualization, the toolkit structures data so that the user, at first glance, can understand the overall state of the team and then drill down where needed. This helps reduce the amount of data displayed to the user at one time. Colors, which are tied to the level of trust (green indicates high trust, red indicates low trust, etc.), serve to direct user attention to instances where trust exceeds or falls below a threshold so they can further explore the situation and propose an intervention if necessary and any problem areas for further exploration. The toolkit has a GUI, or visual output to see information as it changes in real-time (e.g., multiscreens, multi-data inputs), or the user can use the sliding window to observe metrics within a particular time span (What does trust look like over the past 60 min? What did trust look like during time window *t*?) and within or across sub-teams (e.g., trust in sub-team A compared to trust across the entire team). This will also allow the system the opportunity to compare team state/network/configuration to what an ideal team (e.g., high performing) looks like.

2.1.4 Assessment of Team Communication and Associated Metadata

The mode of communication plays an important role in determining which types of communication metadata to capture. The toolkit provides a capability to collect and analyze the content of communication from email, chat, and other written communication in near-real-time. Additionally, typical communication logs of participants (e.g., chat or inbox history) accompanying those communications provide time stamps and information on communication sender and receiver, which are critical elements for dynamic, network-based analyses. Analyzing content from spoken communication requires some extra steps to include transcribing speech to text. Dragon Naturally Speaking (e.g., a commercial off-the-shelf speech-recognition software for diction and transcription) (Krausman et al. 2019) provides a real-time transcription capability; however, accuracy levels have remained low, making it difficult to analyze the content of the communication. Because communication content is a rich source of information about the team dynamic, we are exploring other means and models to increase transcription accuracy. Additionally, identifying critical metadata such as sender, receiver, time stamp, and duration will require software-based solutions (e.g., 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 as described in the following sections.

2.1.5 Graphic Output of Communication/Information from Senders and Receivers

In a scenario in which crew communication was captured, for example, if A speaks on the B network three times and C speaks to the B network two times, then a graphic would show A, B, and C with some larger, directional indicator, such as a line with an arrow at the end from AB and a slightly smaller one from CB (see Fig. 1). Communication logs that are clear, accurate, organized, and stored in a broadly readable format (or multiple formats—A to B at t_1, B to C at t_2, etc.) are essential for examining the flow of communication between team members. The tool is meant to support activities that require clear and accurate data; high-quality log-files are essential and should be incorporated/developed and tested from the outset. Full transcription of communication would be ideal—tagged to identify associations with crew station interactions via precise (sub-second resolution) time stamps as context for each communication "event." In a scenario with participants pushing information through a user interface (e.g., Warfighter Machine Interface [WMI] or a basic crew interface), if A sends some signal to B's interface and then B sends some signal to C's interface, then the graphic would show A-B and B-C. Ideally, the system will allow flexibility in how to display data as a function of the metadata described in Section 2.1.4. 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."

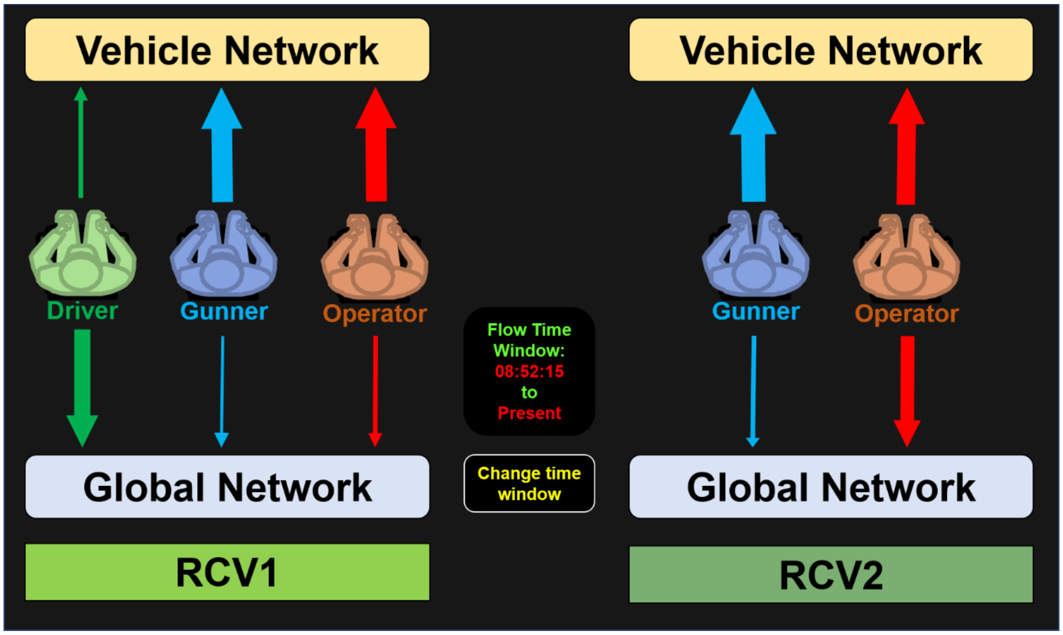


Fig. 1 An example of graphical output of communication flow

The example in Fig. 1 uses an Information for Mixed-Squads (INFORMS) laboratory simulation test bed paradigm where there is one driver of the crew's vehicle, along with two robotic combat vehicle (RCV) crews seated in the same vehicle. This graphical output demonstrates the amount of communication activity initiated by each crew member as well as the targeted destinations of those messages. Important features involve the display of each crew member and each possible destination for the communication. In this case, pressing the INFORMS button can easily tell us who is pressing a button and what network they are trying to speak on; therefore, these data can be visualized in real-time (or as quickly as they can be processed and displayed). Thicker arrows (see Fig. 1) indicate more communication. The user can also select the window of time from which the communication flow graph is generated; for example, selecting "Change Time Window" may allow the user to display the most recent 5 min, or the most recent full scenario, or all available communication data.

2.1.6 Scalability

As an initial goal, the system will be able to accommodate, at a minimum, the same number of participants equivalent to the capacity of the INFORMS laboratory (e.g., 14 individuals comprising differently sized teams), without experiencing significant degradation in response times, visualization update rate, or processing power. However, a long-term goal would be to allow this number to be scaled up to accommodate larger numbers of crew members (with the possibility for different sub-team structures as well).

2.1.7 Support for Different Form Factors

The intended multifunctionality of HAT^3 suggests that it will need to support multiple form factors. A workstation-type form factor will be most applicable and appropriate for a fully equipped laboratory-based study; whereas, mobile or tablet implementations of HAT^3 functions will allow for mobile visualization of data outputs supporting field-data collection (e.g., perhaps using self-report questionnaire measurements supported by HAT^3), and so on. Support for multiple form factors would improve the flexibility of HAT^3 and allow users a greater variety of technological access to utilize and benefit from HAT^3 capabilities.

2.1.8 Data Protection and Security

Recognizing the importance of following Institutional Review Board policies and maintaining confidentiality of participant data, HAT^3 provides secure features to protect all data. By operating on a closed network, the confidentiality of all data sources is maintained, and the storage and transfer of data files is secured. Further, we anticipate the need for maintaining historical data regarding both individuals and aggregated data at the team level, as well as transferring these data across performance instances to improve the quality and robustness of predictive algorithms that are meant to enhance team performance not only during individual missions, but also for continued improvement and optimization over time. Therefore, the toolkit protects stored data, including potential personally identifiable information, given that the system will need to recognize certain users to choose and apply necessary (stored) individualized calibrations/model parameters that have been determined through experimental and analytical methods.

3. System Features

3.1 Multimodal Approach

HAT^3 utilizes a multimodal approach for measuring trust in human-autonomy teams. This stems from the complexity of trust and recognizing that individual measures alone are not sufficient for understanding the team state and how it changes. As such, the toolkit is arranged in separate modules, each one addressing a specific data type and providing several measures from which to choose, depending on the research questions and study needs. Each module contains measures relevant for a specific data type.

The following sections will outline the different modules that the toolkit currently entails, which include subjective and communication modules. Each of these modules has specific metrics relating to trust measurement and will be discussed in more detail.

3.1.1 Subjective

The HAT^3 provides the ability to administer subjective measures to users (e.g., subjective trust measures) at predefined time points and provides near-real-time scoring, analysis, and visualization of results to give the researcher/user quantitative feedback on pre-, mid-, and post-task subjective differences related to trust. The system is also capable of allowing for readministration/analysis later so that users can gauge responses when deemed necessary.

3.1.1.1 Subjective Module Metrics

The benefits of subjective measurement are widespread and allow researchers to gain valuable insights into the subjective responses and changes of specific states over the course of time. Regarding trust measurement in human-autonomy teams, subjective scales can be differentiated into two main groups: 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, other trust surveys and subjective scales can be added if required by the research needs. In fact, while trust is the main construct of interest for the toolkit's platform, many users may require more than simply trust-based questionnaires; therefore, other widely used subjective surveys within the HAT literature have been incorporated (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

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

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

3.1.1.2 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. 2). Additionally, one of the goals of the HAT^3 software platform was to visualize data for a variety 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.

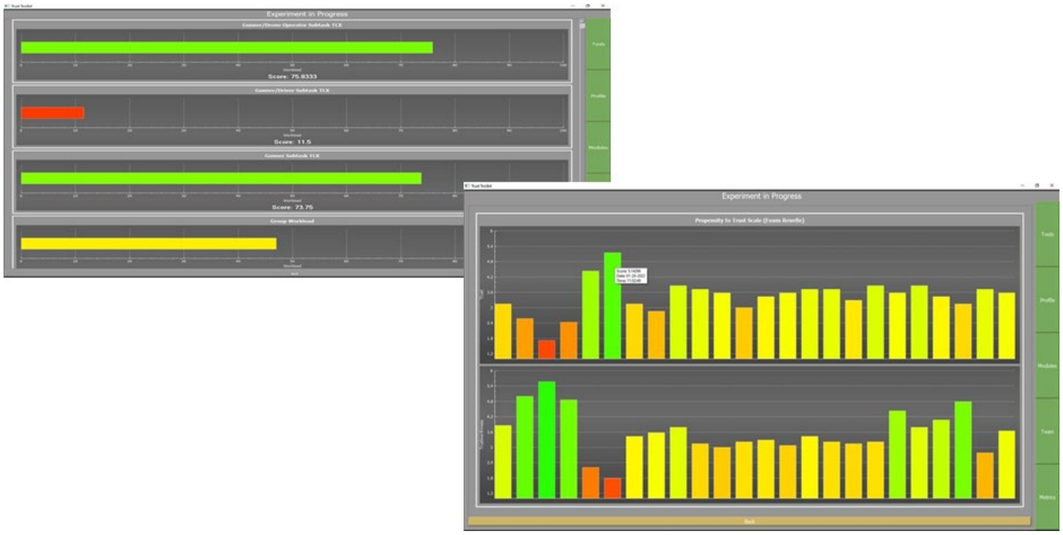


Fig. 2 Historical view of one team member's scores over time. Colors indicate a continuous spectrum of subjective trust state which ranges from high (green ranges), to medium (yellow–orange ranges), to low (red ranges).

3.1.2 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, goals, 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, 2022). By measuring aspects of when and how a team communicates, we can glean information that may relate to their trust and cohesion behaviors.

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, online, or 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 US Army Research Laboratory (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.

The next section 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 analyses.

NOTE: Data for the communication visualizations were collected from a vehicle crew of seven members during a simulation experiment. Each crew station is labeled as CS01 – CS07, and each crew member performed a specific role: Commander, Gunner, or Driver.

3.1.2.1 Communication Module Metrics

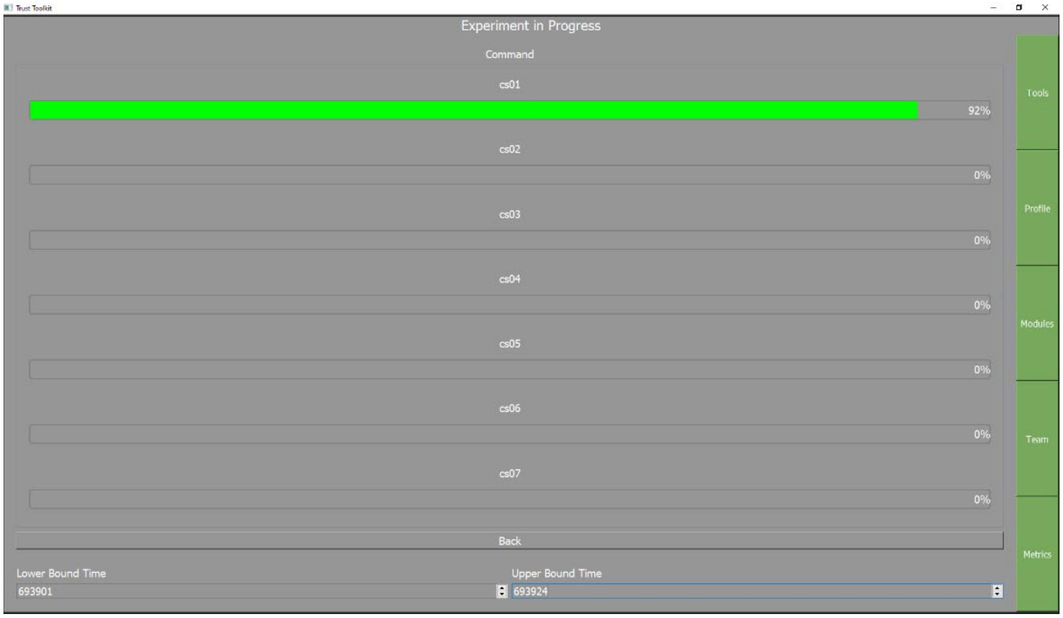
Communication Flow: Real-Time Event, Flow, and Coordination Tool (REFLECT)

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, can allow for a play-by-play understanding of how the crew interacted throughout a mission. REFLECT is aimed at supporting this concept; its purpose is to visually represent the flow of communication among a given team, regardless of the number of crew members or communication networks. Communication data, comprising information about the sender and recipient, is currently collected by the INFORMS environment (i.e., via Kafka) and is then imported into the HAT^3; however, in future iterations, the toolkit itself will collect the relevant data in real time and still be able to import data from INFORMS (or other similar) environments. This data is then represented visually in the GUI to show the user how the team's communication is structured. The tool could display, for example, communication networks during a particular time window or the communication network of a particular sender or receiver. Therefore, the aim of this tool is to provide a simplified but broad overview of the team's communication behaviors, allowing for a clear understanding of the aggregate communication events that occurred.

3.1.2.2 Communication Module Data Visualization

3.1.2.2.1 REFLECT Visualization

REFLECT 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 frequency of communication for each crew member, 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 network 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 provide an overview of the visualizations of the REFLECT module (Figs. 3–5).



cs01 92%

cs02 0%

cs03 0%

cs04 0%

cs05 0%

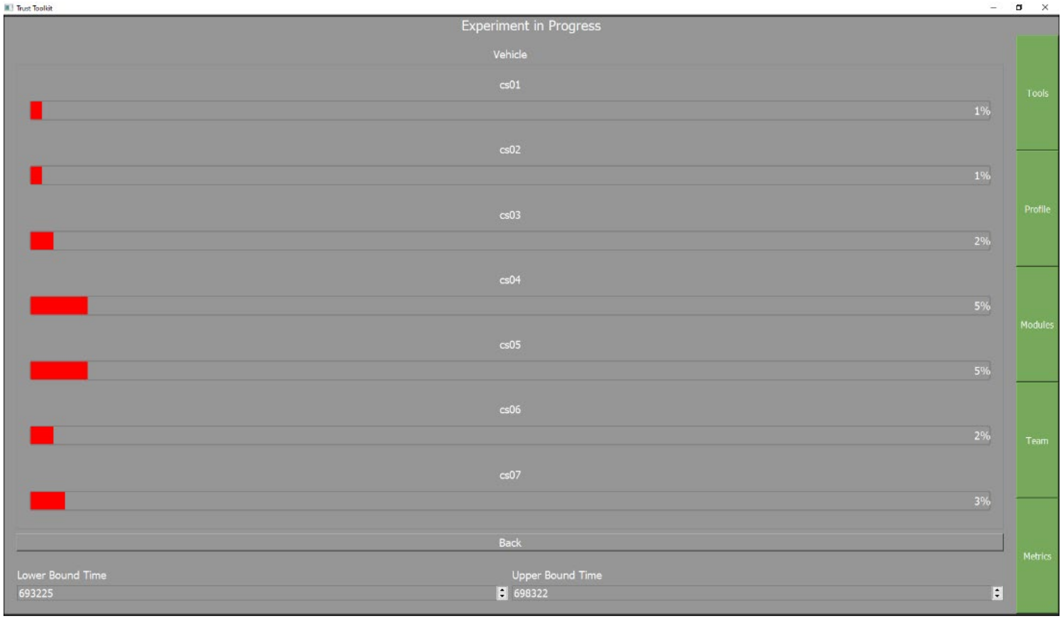
cs06 0%

cs07 0%

Lower Bound Time 693901

Upper Bound Time 693924

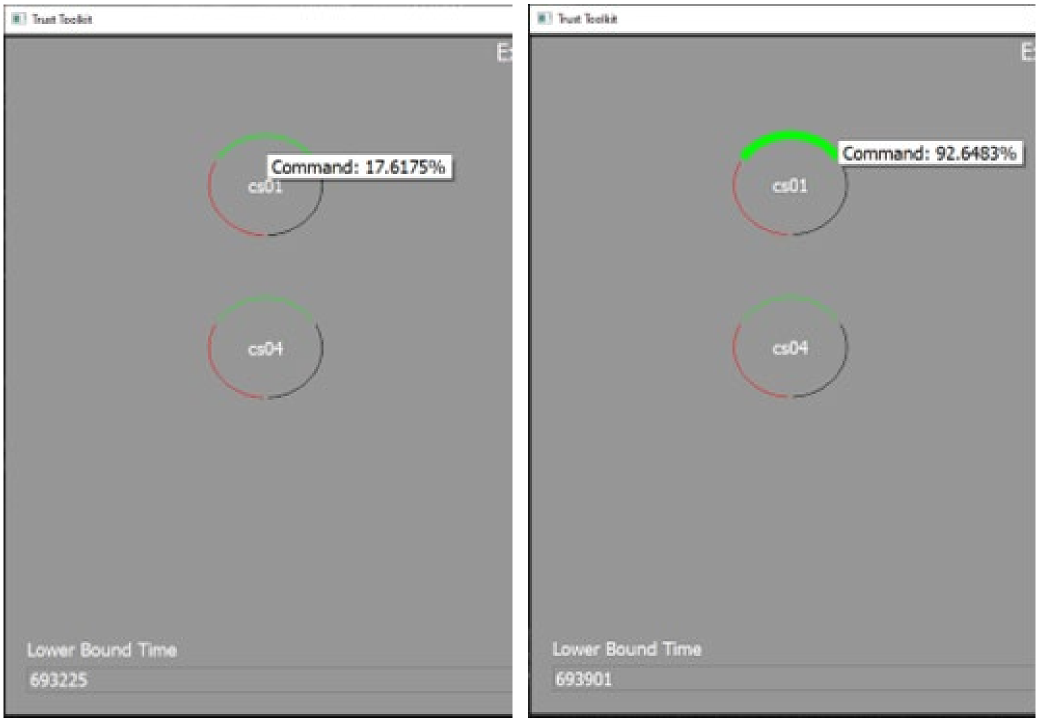
Fig. 3 Depiction of communication flow among a team, using simulated data. These represent communications on a Command network. In this screenshot, we note that between the time markers of 693901 and 693924 (a span of 23 s), the user cs01 was responsible for 92% of the team's communication.



Lower Bound Time 693225

Upper Bound Time 698322

Fig. 4 Crew communication flow 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.



Lower Bound Time 693225

Lower Bound Time 693901

Fig. 5 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 on 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.

3.1.2.2.2 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 measures within the team. For example, the graphic represented in Fig. 6 shows four color-coded teams with lines that indicate interactions within and between the team. The node size indicates the importance for routing messages through the network, measured by the "betweenness centrality" function. Because these larger nodes occupy positions where they are essential for routing information throughout the network, they need high levels of trust and to be trusted by team members. This type of visualization is a promising indicator for trust interventions as it clearly demonstrates that the interventions should target and prioritize the larger nodes.

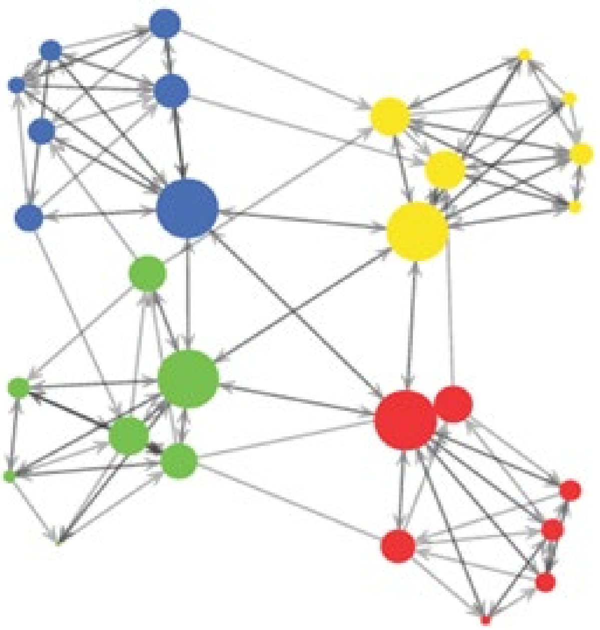


Fig. 6 An illustrative example featuring four color-coded teams. Circles represent individuals, and lines between them represent communication ties. The size of each node reflects its betweenness centrality, or the tendency to occupy positions on shortest paths between nodes.

While the visualization approach has considerable overlap with REFLECT—in that it still captures the sender, receiver, and time stamp information—the network analysis tool differs in a few key ways. These differences principally arise from 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 provides 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. Additionally, these individual-level statistics can be useful for identifying where to deploy a trust-enhancing intervention. In the following we highlight several key statistics we aim to capture and represent in the network module.

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 decomposed into 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 many 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 quickly 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).

   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.

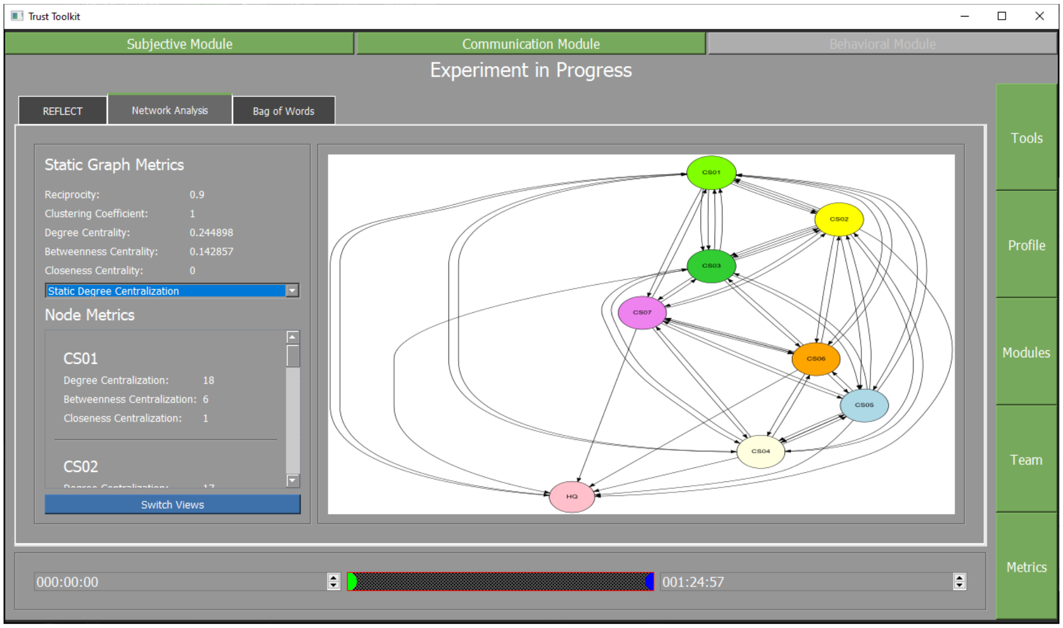
d) Betweenness centralization: Betweenness centralization measures the concentration of betweenness centrality in the network. Higher scores indicate that a small number of nodes occupy essential positions for routing messages through the network while lower scores indicate that messages can more easily 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 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, operating independently of other systems, or as a system that can be integrated into other HAT systems (e.g., P8 Dashboard, WMI). The capability for both 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.

3.1.2.2.3 Network Analysis Visualization

The Network Analysis tool has a GUI, or visual output to see information as it changes in real-time (e.g., multi-screens, multi-data inputs; see Fig. 7). 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?). Attributes of individual nodes, such as size or color, can be modified in real time to reflect centrality scores. For example, nodes most important for routing information through the network could 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
Degree Centralization: 17

000:00:00
001:24:57

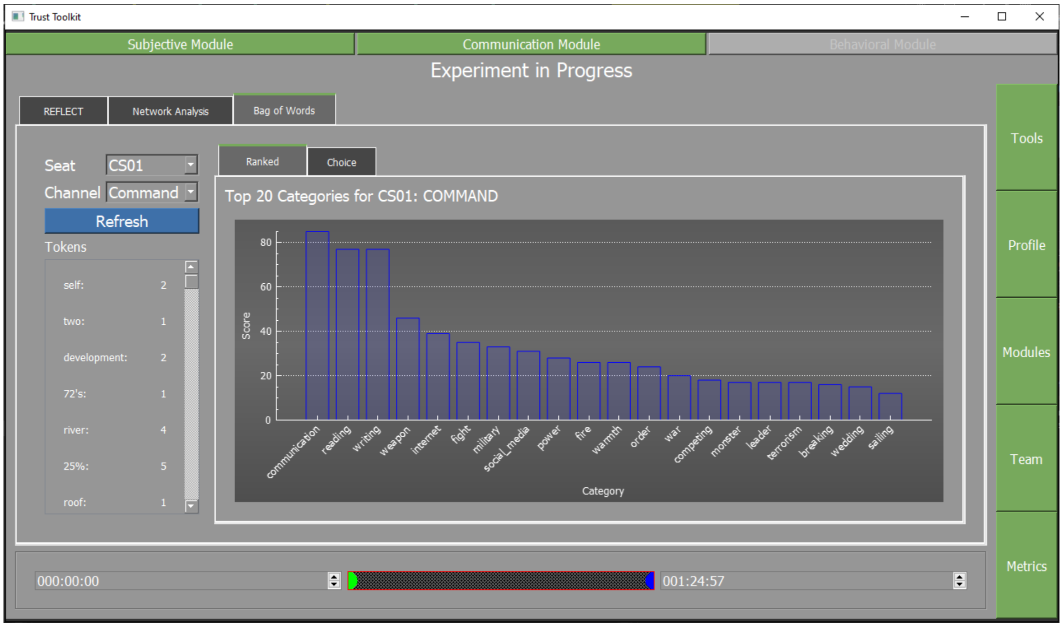
Fig. 7 Example network analysis data visualization. Node size and color can be used to highlight features such as individual attributes (role, team membership) and network attributes (centrality scores, etc.). Additional details on the left provide individual- and network-level statistics.

3.1.2.2.4 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 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.

3.1.2.2.5 Bag of Words 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. 8). 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. An alternative view presents a fixed set of categories that are related to an area of interest, for example Trust or Affect.



Seat CS01

Channel Command

Top 20 Categories for CS01: COMMAND

Tokens

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

000:00:00

001:24:57

Fig. 8 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. Path Forward

The development of the HAT^3 software platform is ongoing with plans to incorporate several more modules and elements into the capabilities of the system. First, the "VocStr" emotion detection model will be added to the communication module. Specifically, the VocStr module of HAT^3 will detect 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 will be used as a criterion to deploy/suggest interventions.

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. 2022).

The toolkit can also be integrated into other ARL platforms such as the 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. 2022;65(1):1157–1161.

Baker AL, Fitzhugh SM, Huang L, Forster DE, Scharine A, Neubauer C, Lematta G, Bhatti S, Johnson CJ, Krausman A, Holder E, Schaefer KE, Cooke NJ. Approaches for assessing communication in human-autonomy teams. Human-Intelligent Systems Integration. 2021;3(2):99–128. doi: 10.1007/s42454-021-00026-2.

Costa AC, Anderson N. Team trust. The Wiley Blackwell handbook of the psychology of team working and collaborative processes. Hoboken: Wiley-Blackwell; 2017. p. 393–416.

Demir M, Likens AD, Cooke NJ, Amazeen PG, McNeese NJ. Team coordination and effectiveness in human-autonomy teaming. IEEE Transactions on Human-Machine Systems. 2019; 49(2):150–159. doi: 10.1109/THMS.2018.2877482.

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.

Fulmer AC, Gelfand MJ. At what level (and in whom) we trust: trust across multiple organizational levels. J Management. 2012:38(4):1167–1230.

Grossman R, Feitosa J. Team trust over time: modeling reciprocal and contextual influences in action teams. Hum Resour Manag Rev. 2018;28:395–410.

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, 2020. 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, Schaefer, Oiknine, Thurman, Files, Gordon, Bradford, Spangler, Gremillion. 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 and 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. https://www.army.mil/article/222153/army_futures_leveraging_mission_command_for_effective_soldier_robot_teams.

Appendix A. Human-Autonomy Teaming Trust Toolkit (HAT^3) User Guide

A.1 Introduction

The US Army Combat Capabilities Development Command (DEVCOM) Army Research Laboratory (ARL) HAT^3 is a collection of tools for accessing subject trust. It currently comprises a subjective module that prompts participants to respond to subjective questionnaires and a communications module that provides various analyses of communication data collected during an exercise or experiment. This user's guide will cover the basic setup and use of these tools in an experimental setting. It assumes initial installation of the toolkit and that associated dependencies have already been performed on all necessary computers.

A.2 Usage Overview

The ARL HAT^3 comprises three software components: the SurveyWrapper, the TrustDelegate, and the TrustToolkit. It also depends on MongoDB to store data it collects. These components perform the following functions:

- TrustTookit – The interface an experimenter uses to setup groups of subjective surveys and send requests to subjects, view the results of subjective surveys, and view communication analyses.
- SurveyWrapper – The software that presents subjective surveys to users and collects their responses in a local file.
- TrustDelegate – The software that transmits responses collected by the SurveyWrapper to MongoDB for storage. It can also broadcast the responses on the local network through Kafka for centralized collection in environments such as the Information for Mixed-Squads (INFORMS) laboratory.

While all components can be installed on a single computer system for test or demonstration, the typical use case will involve a single installation of MongoDB for the TrustToolkit software on the experimenter's computer. The SurveyWrapper and TrustDelegate are installed on all computers that subjects will use to respond to subjective surveys.

A.2.1 Start-Up

On each subject's computer perform the following actions:

1) Double click on the TrustDelegate shortcut on the Desktop.
2) Double click on the SurveyWrapper shortcut on the Desktop.

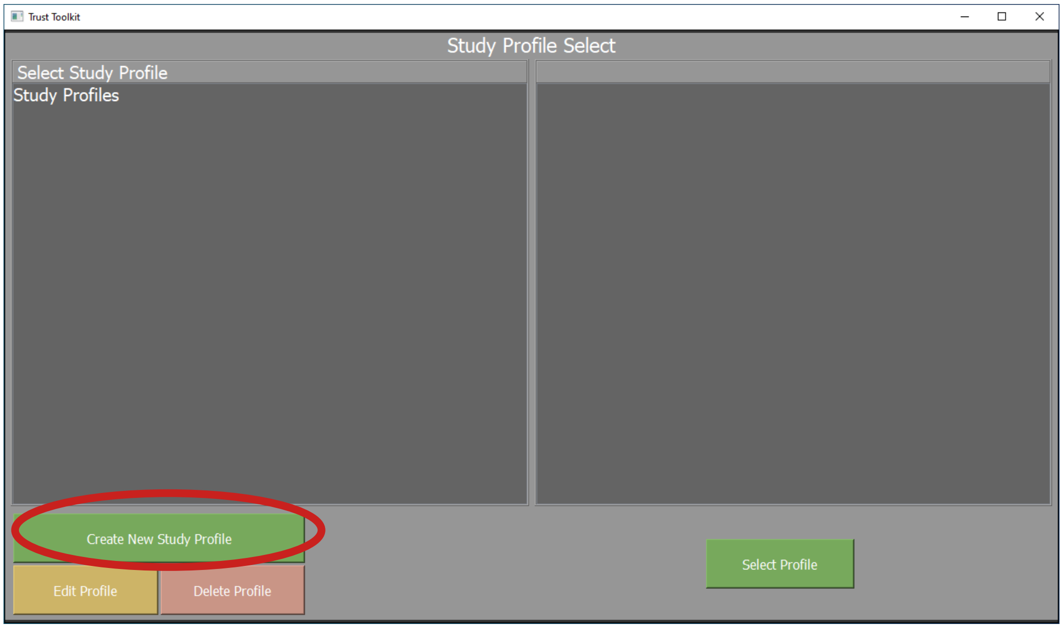
On the experimenter's computer perform the following action:

1) Double click on the TrustToolkit shortcut on the Desktop.

A.2.2 Setting Up a Study Profile

Once the Toolkit is up and running you may setup a study/experimental profile.

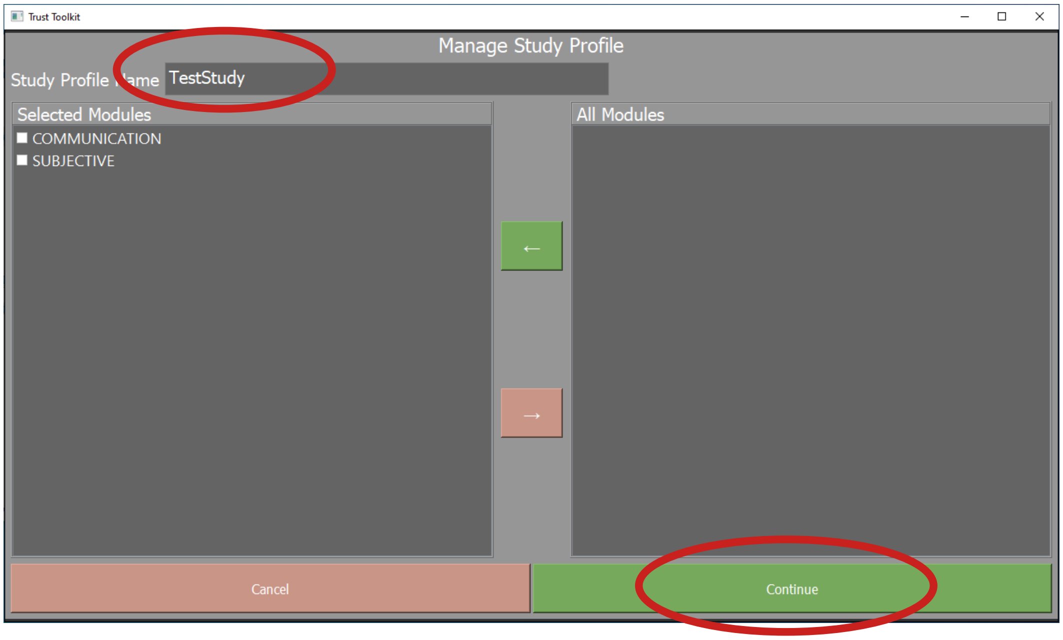
Step 1: Create a new study profile by clicking the "Create New Study Profile" button on the toolkit home screen.



Step 2: Create a name for the study and select the modules required from the list on the right side of the screen labeled “All Modules”.

Once selected, click the left (green) arrow to move the module(s) to the “Selected Modules” column on the left side of the screen.

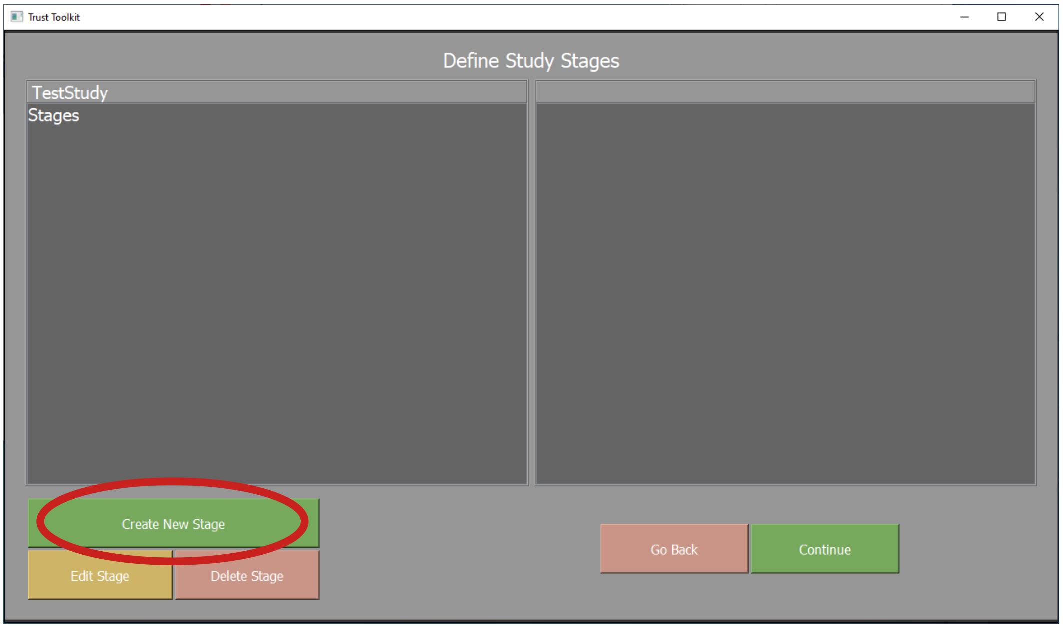
When this step is complete, click “Continue”. If a module is incorrectly moved to “Selected Modules” and not intended for use, select it and click the red arrow to move it back under All Modules.



Study Profile Name TestStudy

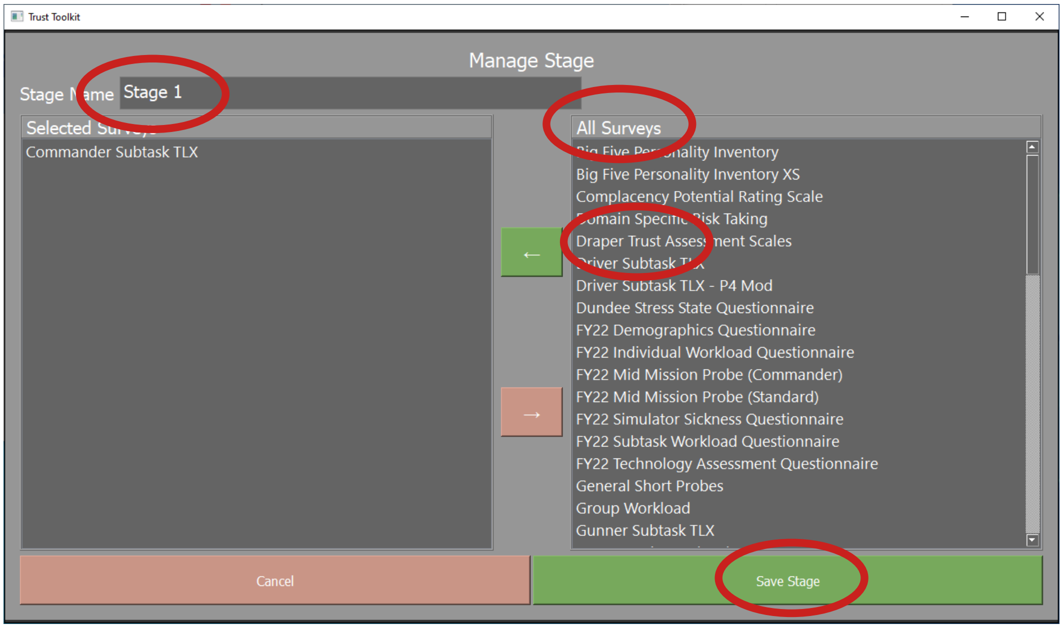
□ COMMUNICATION
□ SUBJECTIVE

Step 3: Next, the study stages (if applicable) will be defined by clicking "Create New Stage". Stages are the different time points in a study (e.g., pre-, mid-, and post-study). This will determine when the data sources or streams (e.g., subjective, communication) will be collected.



Step 4: In the “Manage Stage” step, the measures of interest will be defined. Enter a name in “Stage Name” for the stage. From the “All Surveys” column, select the desired surveys and click the green arrow to move them under the “Selected Surveys” column. To undo this selection, select the survey from Selected Surveys and click the red arrow to move it back to “All Surveys”. Once the surveys are selected, this step is complete.

Click “Save Stage”.



Trust Toolkit

Manage Stage

Stage Name Stage 1

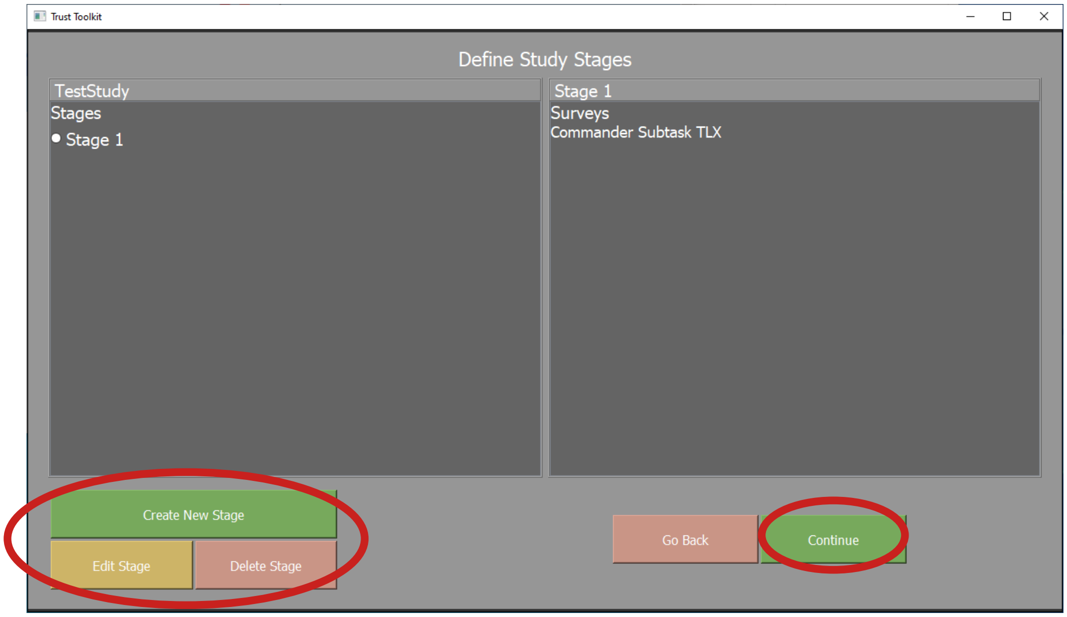
Selected Surveys

Commander Subtask TLX

All Surveys

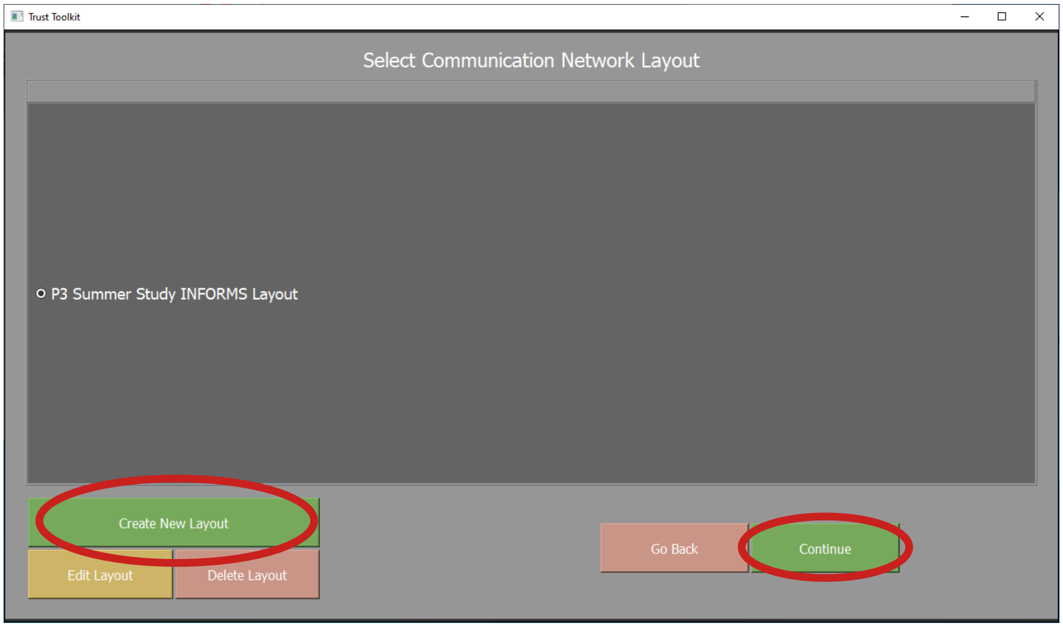
Big Five Personality Inventory
Big Five Personality Inventory XS
Complacency Potential Rating Scale
Domain Specific Risk Taking
Draper Trust Assessment Scales
Driver Subtask TLX
Driver Subtask TLX - P4 Mod
Dundee Stress State Questionnaire
FY22 Demographics Questionnaire
FY22 Individual Workload Questionnaire
FY22 Mid Mission Probe (Commander)
FY22 Mid Mission Probe (Standard)
FY22 Simulator Sickness Questionnaire
FY22 Subtask Workload Questionnaire
FY22 Technology Assessment Questionnaire
General Short Probes
Group Workload
Gunner Subtask TLX

Step 5: To add additional stages, click “Create New Stage”. To edit a stage, select the stage by clicking the radio button next to it and click “Edit Stage”. To delete a stage, select the stage by clicking the radio button next to it and click “Delete Stage”. Once entries are finalized, click “Continue”.

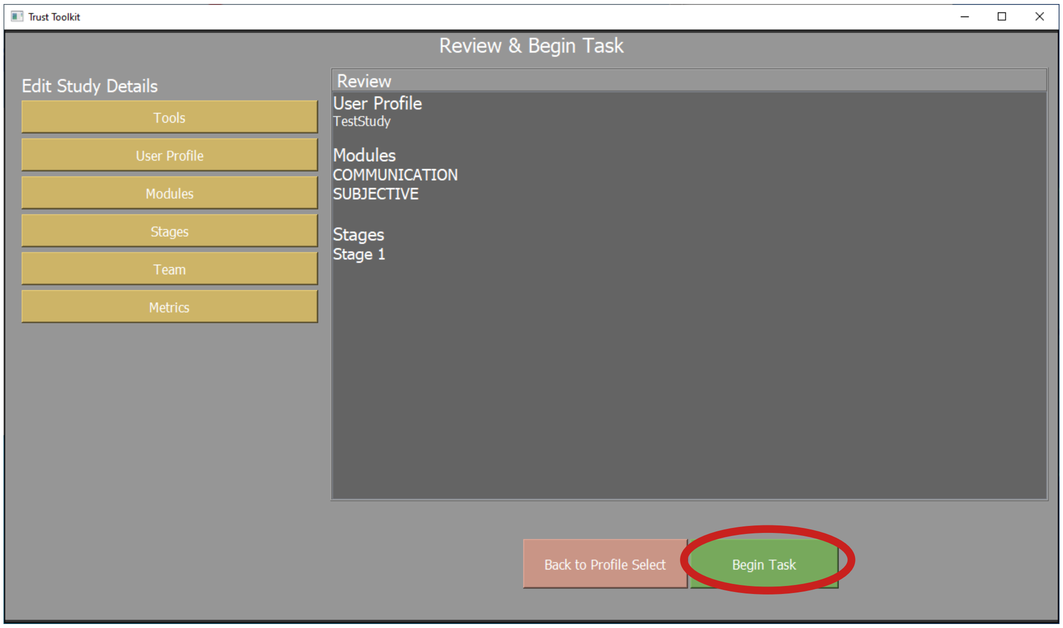


Step 6: If the Communications module has been selected, the Select Communication Network Layout screen will appear next. If the Communications module is not selected, proceed to Step 7.

On this screen, select the “Network Layout” for the environment the Trust Toolkit will be run in and click “Continue”. This is necessary to properly assign communications to the correct users during analysis.



Step 7: In the "Review and Begin" task step, verify all entries. To make additions or edits, click "Go Back"; otherwise, click "Begin Task".



User Profile
TestStudy

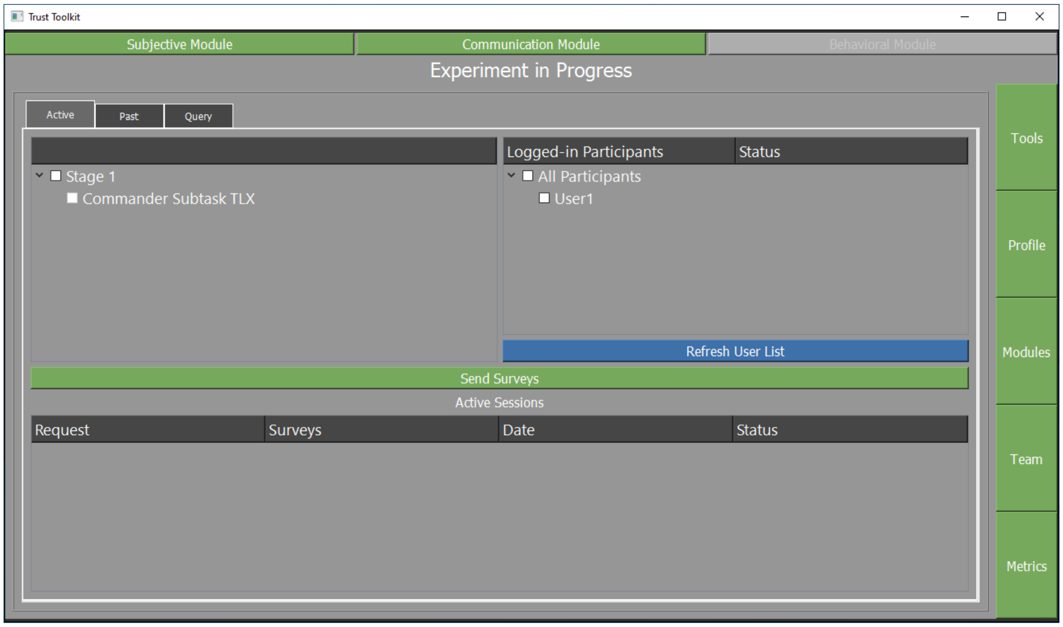
Modules
COMMUNICATION
SUBJECTIVE

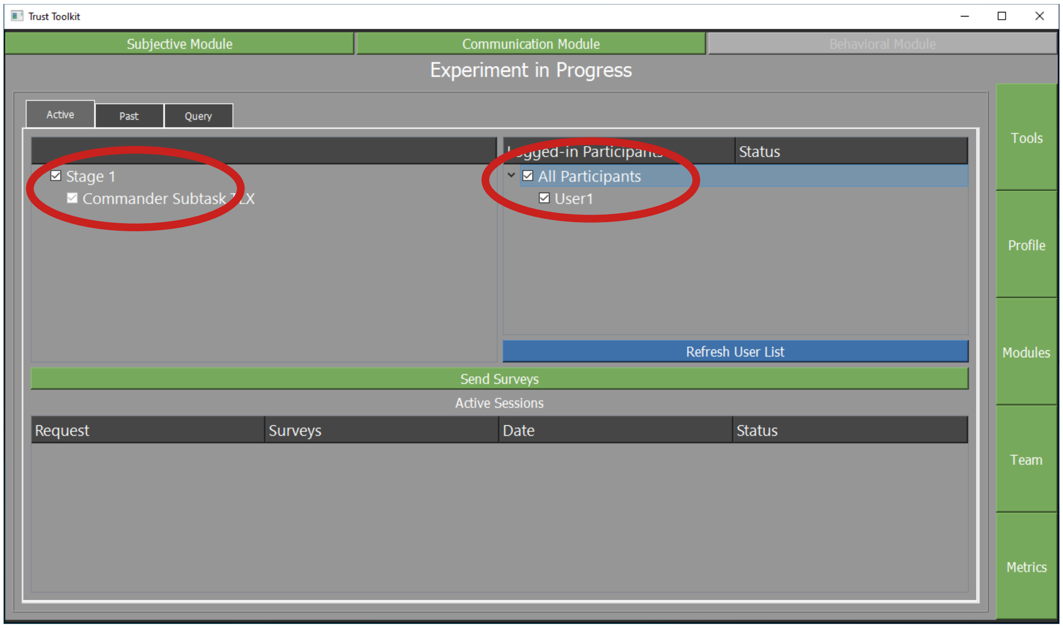
Stages
Stage 1

A.3 Subjective Module

Step 1: Once the experiment begins, the view will change to the "Experiment in Progress" screen. The left side lists stages that were setup in the previous steps. The stages can be expanded to see the subjective surveys that are part of the stage. The right-hand side lists Logged-in Participants. These are subjects who have SurveyWrapper and TrustDelegate running on their computer. The Active Session view is at the bottom; it shows the active survey requests that are in progress.

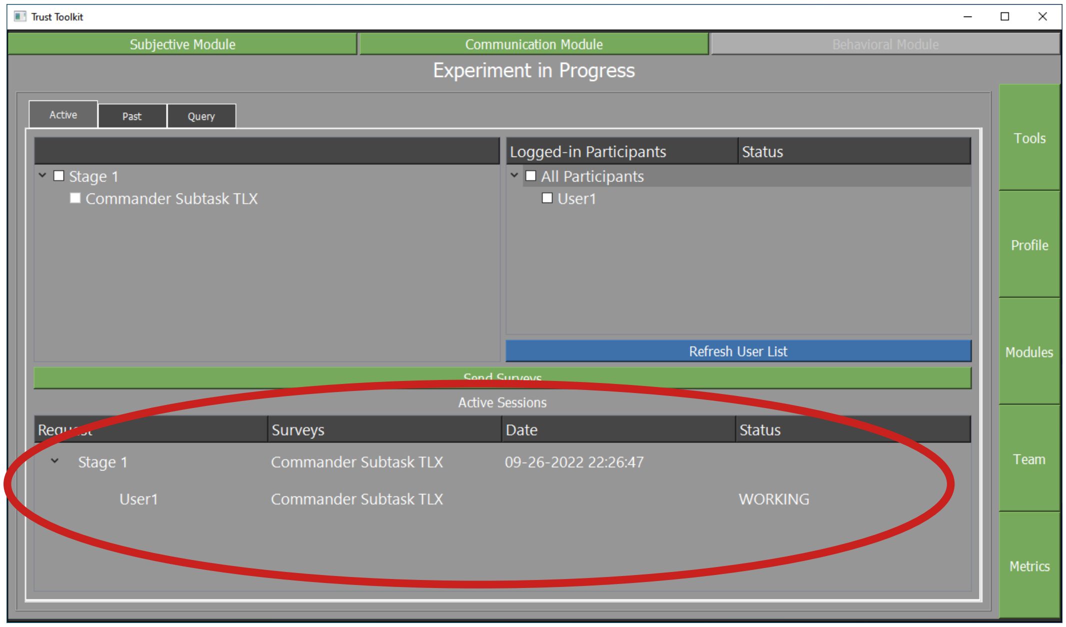
To send a subjective survey request, select one or more Stages from the left side, select one or more Logged-in Participants from the right side, and then click "Send Surveys". Subjects will receive a notification on their computer that there are surveys to complete, and their progress will be displayed in the "Active Session" area at the bottom.





- ☑ Stage 1
  - ☑ Commander Subtask TLX
- ☑ All Participants
  - ☑ User1

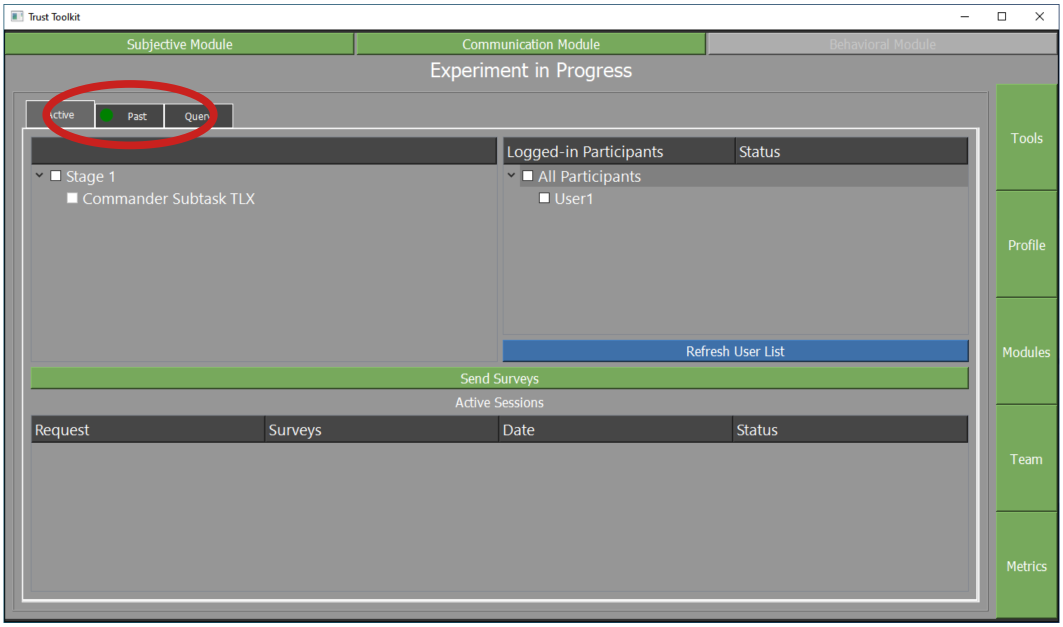
Request	Surveys	Date	Status



- □ Stage 1
  - □ Commander Subtask TLX
- □ All Participants
  - □ User1

Request	Surveys	Date	Status
Stage 1	Commander Subtask TLX	09-26-2022 22:26:47	
User1	Commander Subtask TLX		WORKING

Step 2: When subjects complete a survey request it is removed from the Active Sessions area and moved to the "Past" tab. A green dot on the Past tab indicates new results are available.



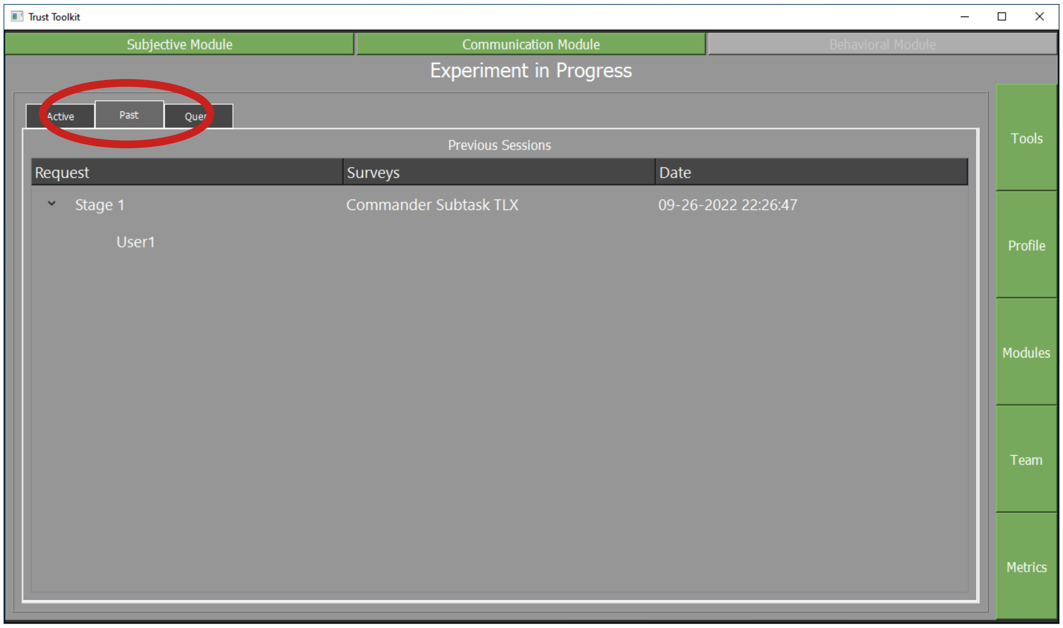
☐ Stage 1
☐ Commander Subtask TLX

Logged-in Participants	Status
☐ All Participants	
☐ User1	

Active Sessions

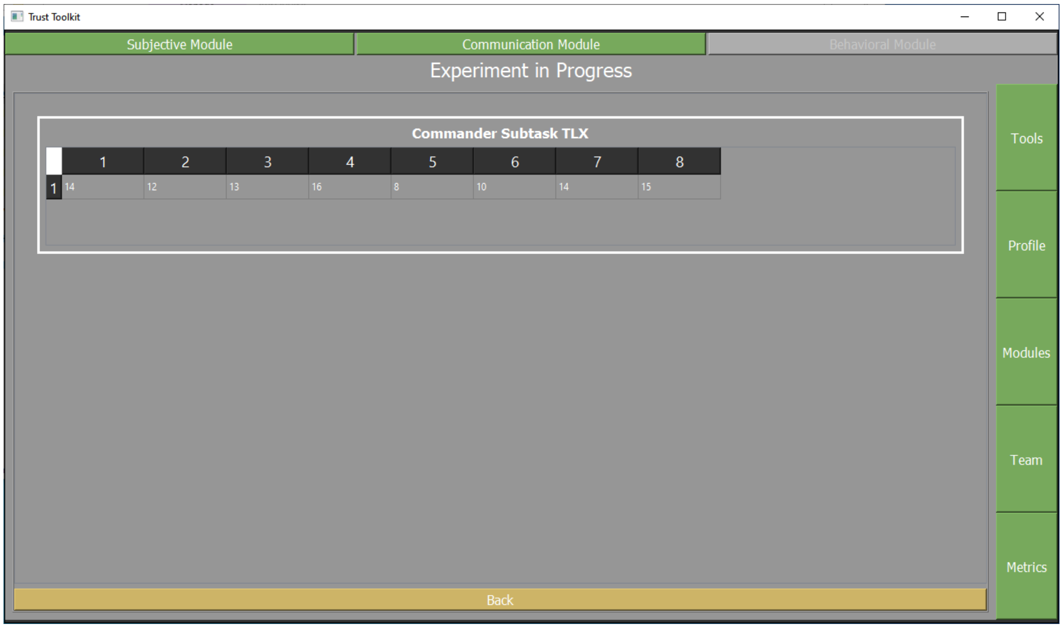
Request	Surveys	Date	Status

Step 3: Click on the “Past” tab to view individual results. Double clicking on a user will show their individual results either as response codes or graphically (if a graphical method is defined for the survey).



Request	Surveys	Date
Stage 1	Commander Subtask TLX	09-26-2022 22:26:47
User1		

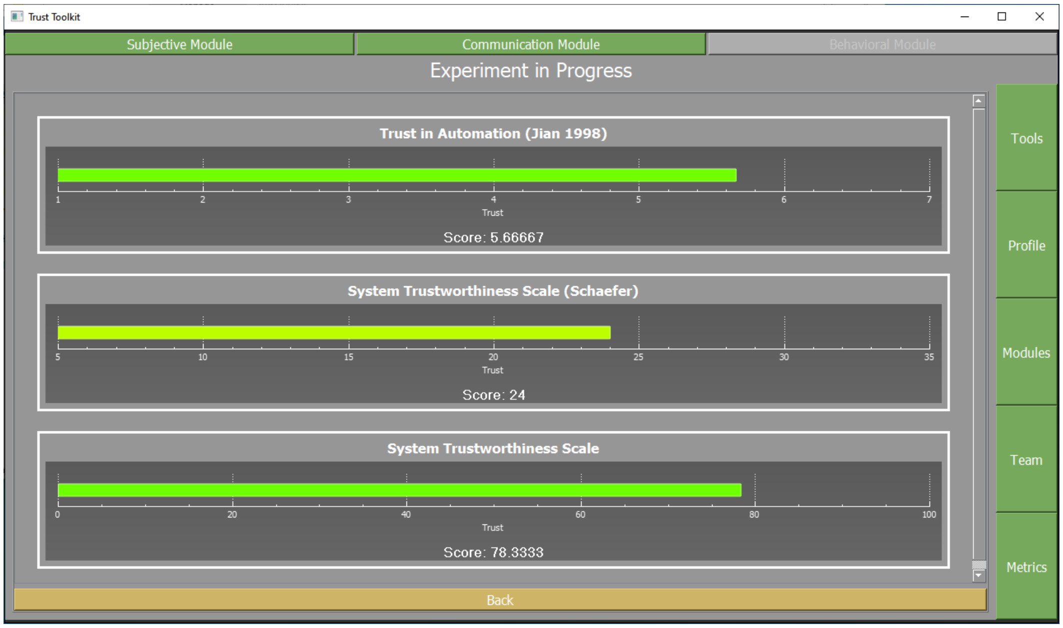
Example of non-scored response code display.



Commander Subtask TLX

	1	2	3	4	5	6	7	8
1	14	12	13	16	8	10	14	15

Example of scored graphical visualization of subjective responses.



Experiment in Progress

Trust in Automation (Jian 1998)

Score: 5.66667

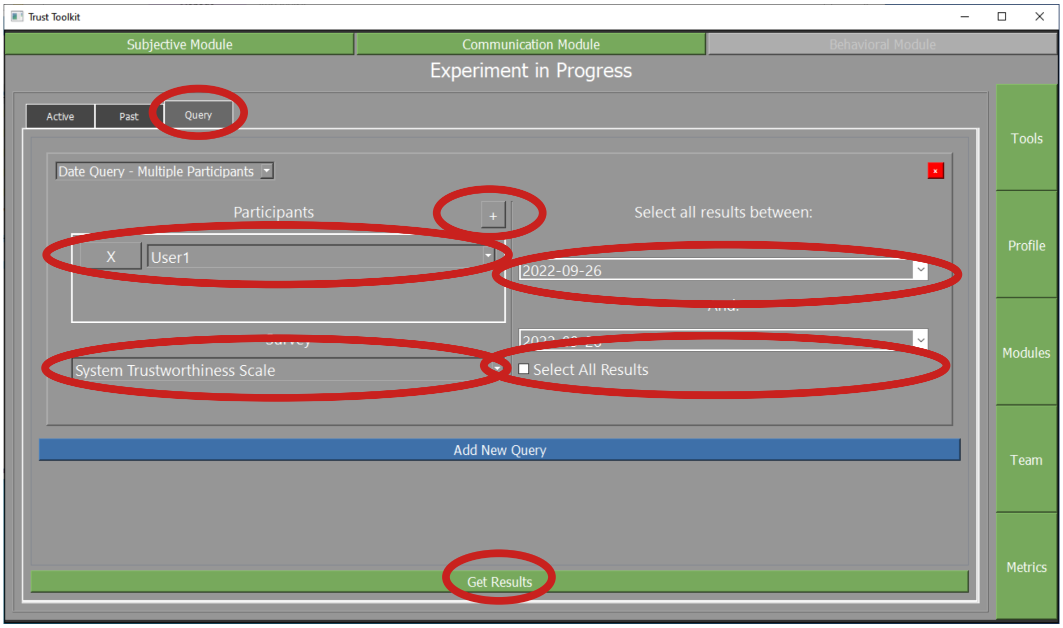
System Trustworthiness Scale (Schaefer)

Score: 24

System Trustworthiness Scale

Score: 78.3333

Step 4: View multiple user responses by clicking on the “Query” tab. In the “Select Query” drop-down menu, select “Date Query – Multiple Participants”. In the “Participants” area, click “+” and then select a participant. Select a survey in the “Survey” selection drop-down menu. Select the desired date range on the right. Click “Get Results” to run the query.



Date Query - Multiple Participants

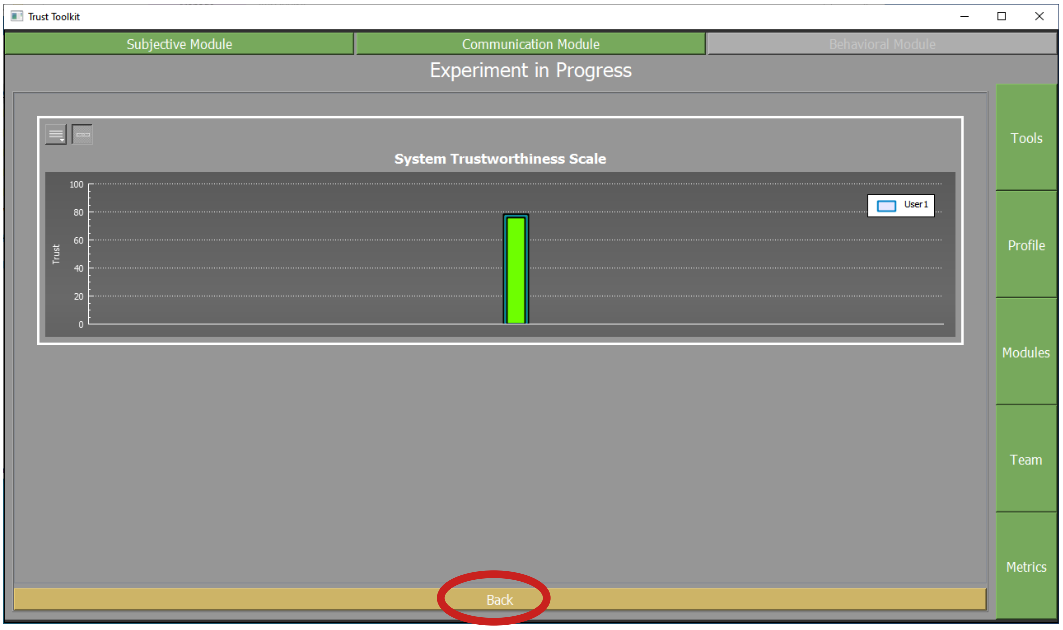
User1

2022-09-26

System Trustworthiness Scale

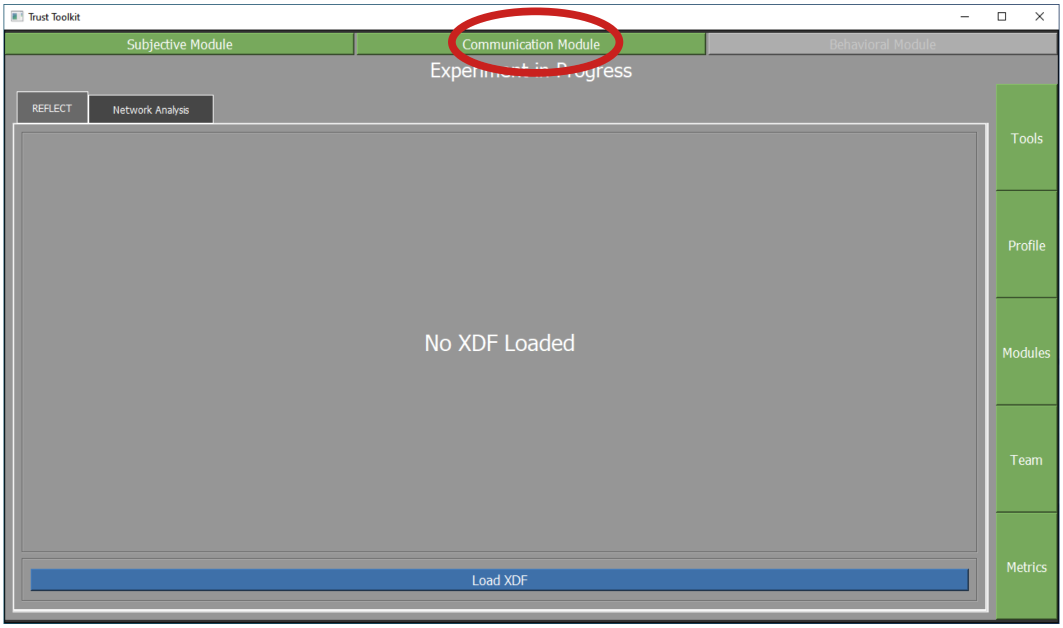
In this example there is only one user; however, multiple users would be displayed as additional vertical bars. Each bar has two colors; the inner color varies from red, to yellow, to green based on the higher score (green in this example). The outline color of the bar is a color assigned to the user (blue in this example).

Click "Back" to return to the "Query" screen.

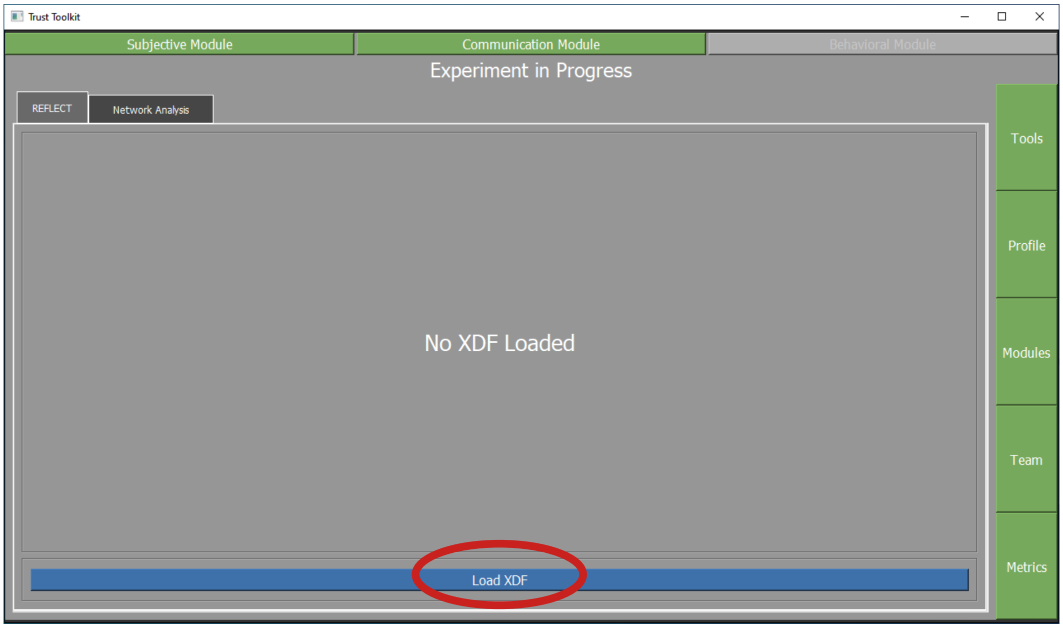


A.4 Communications Module

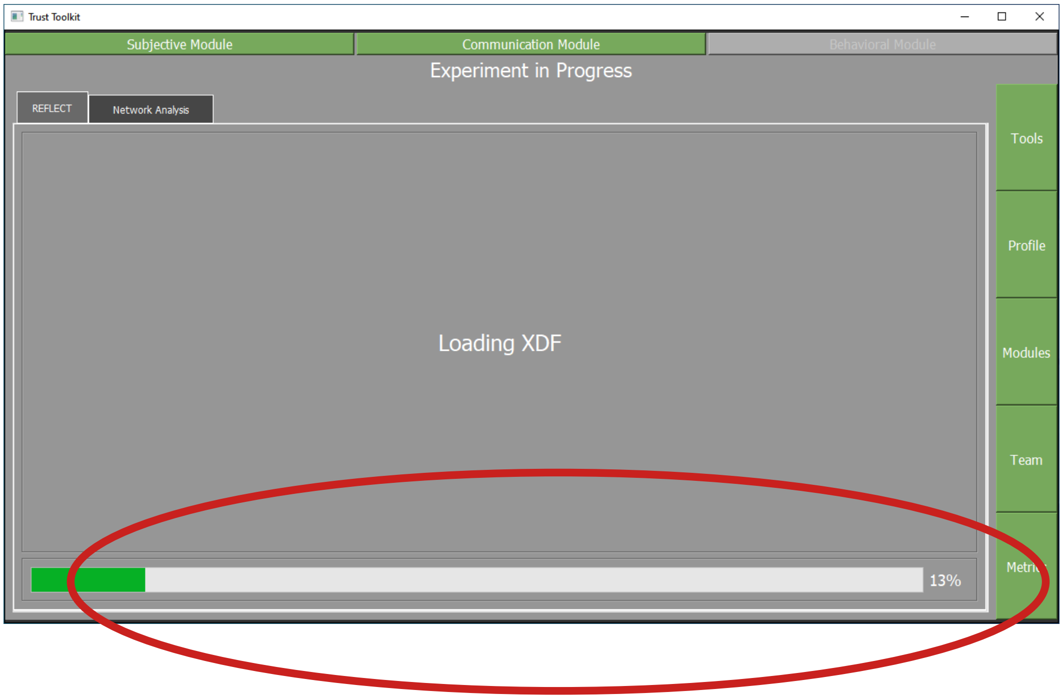
Step 1: From the “Experiment in Progress” screen (reached in Step 7 of the Setting Up a Profile section), click on “Communications Module”.



Step 2: Click “Load XDF” and select the XDF file collected during an experiment containing “Push to Talk” events and audio transcriptions.



Step 3: Wait while the XDF file is parsed and analyzed.



Experiment in Progress

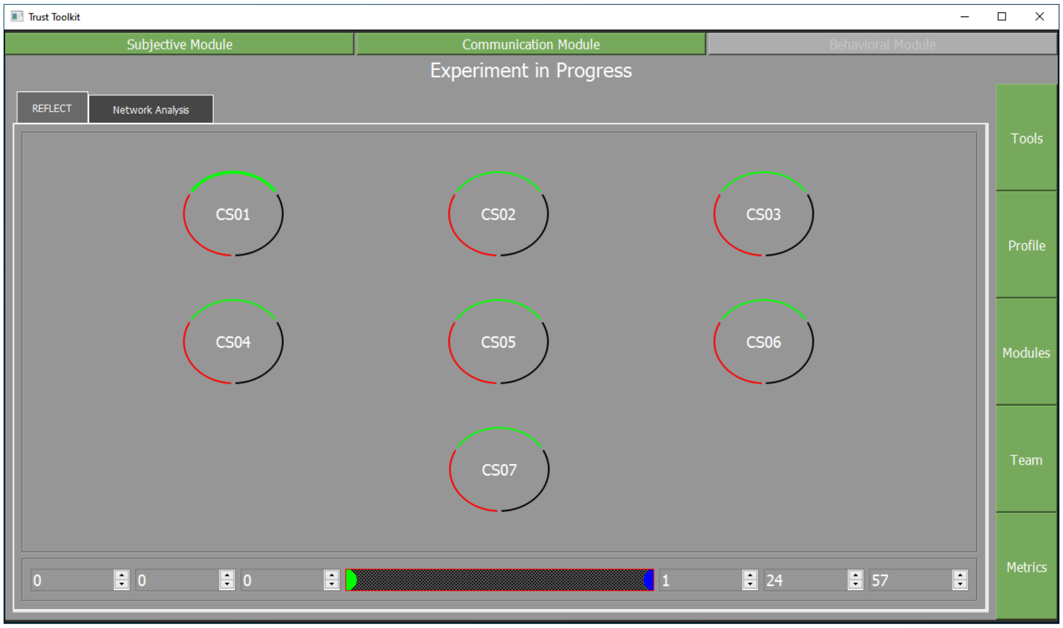
Loading XDF

13%

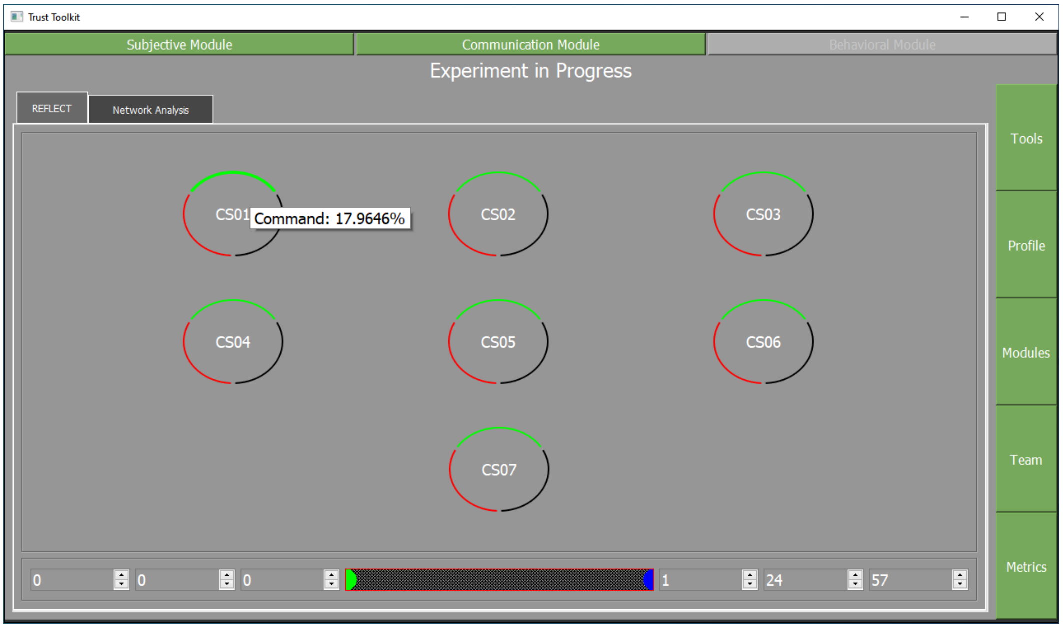
Step 4: When the XDF file is parsed, the Real-Time Event, Flow, and Coordination Tool (REFLECT) will show an overview of the percentage of time each participant spoke on each network. Each user is represented by an identifier enclosed in a multicolored circle. Here, the identifiers represent users at individual Crew Stations (CS01) in the INFORMS laboratory. Three networks are present: Command (Green), Section (Black), and Vehicle (Red). The thickness of the segment indicates proportionally how much time that person spoke on each network. For example, note that CS01's green bar is thicker than the others indicating they spent more time speaking on the Command network.

NOTE: In the following screenshots for Steps 4–8, the numbers along the bottom of the screenshot represent the specific window of time being visualized from a scenario, in the time format of HH:MM:S. Here, the time is from 00:00:00 to 01:24:57; therefore, from the start of the study to the 1 h, 24 min, and 57 s mark.

Also, the nomenclature CS01–CS07 represents individual crew stations within a simulated vehicle.

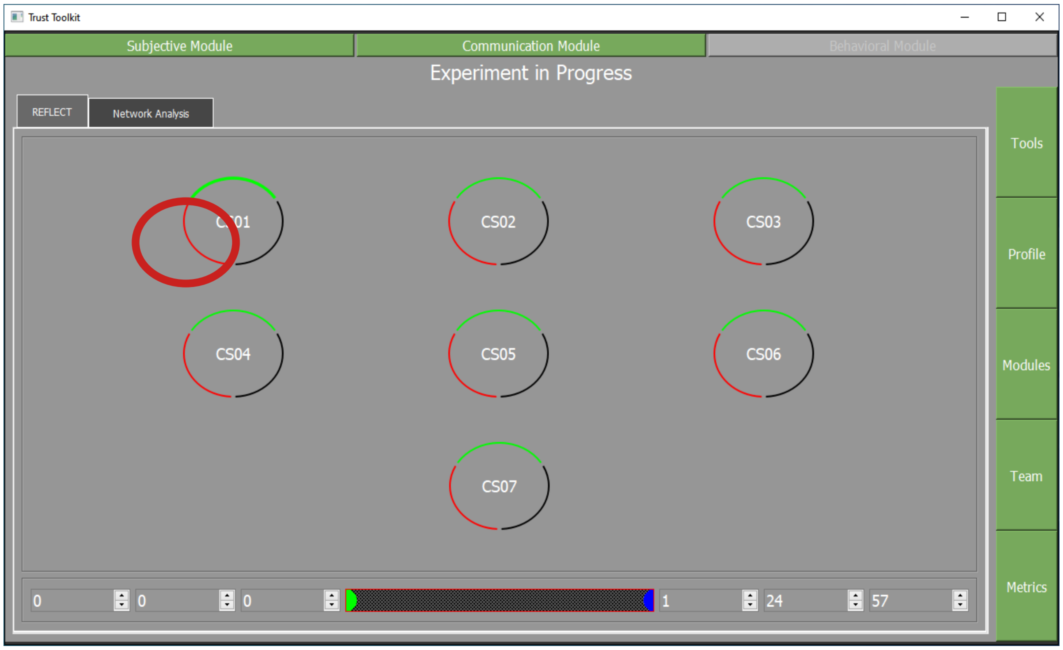


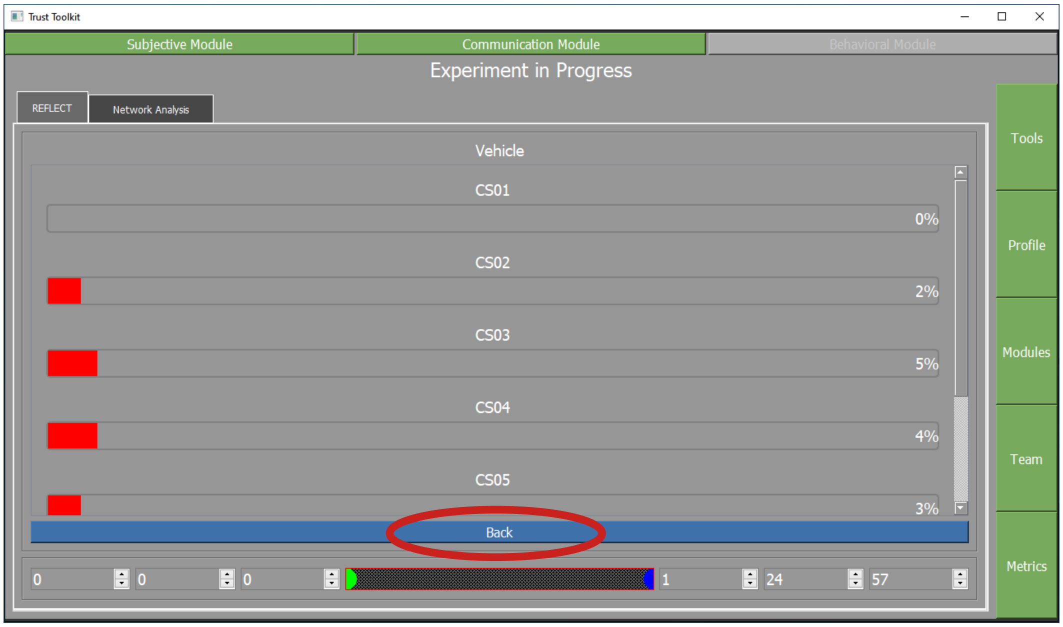
Step 5: Hover over a segment to view the percentage of time a subject spent talking on a particular network.



Command: 17.9646%

Step 6: Click on one of the colored segments for any user to switch to a view showing a comparison of each subject's time spent speaking on that network. Click "Back" to return to the overview.





Vehicle

CS01 0%

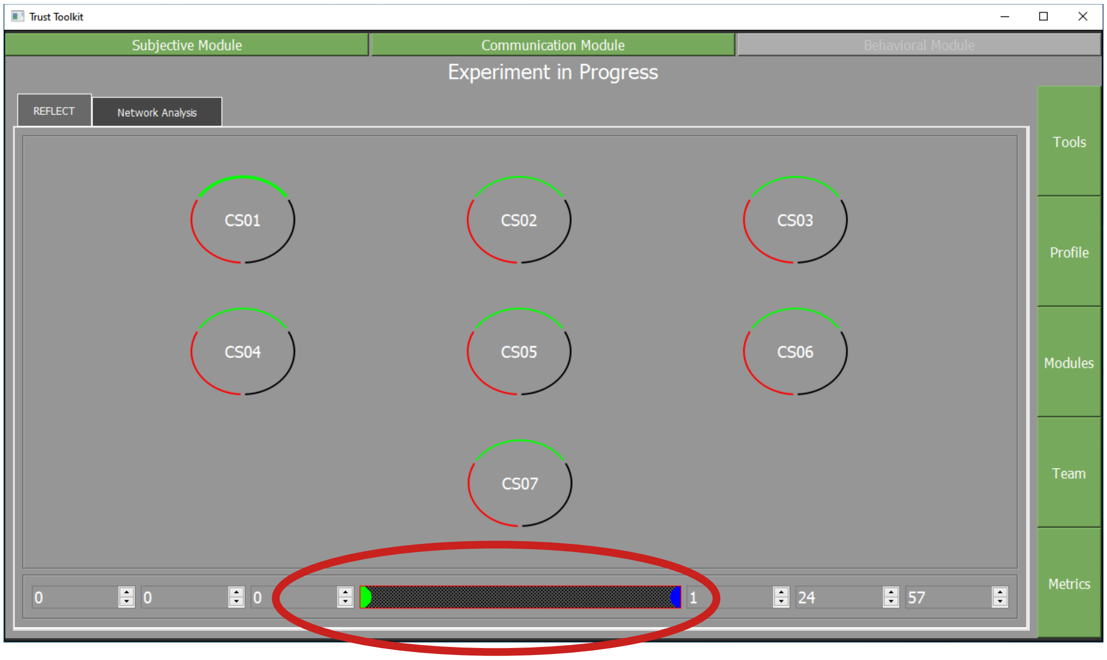
CS02 2%

CS03 5%

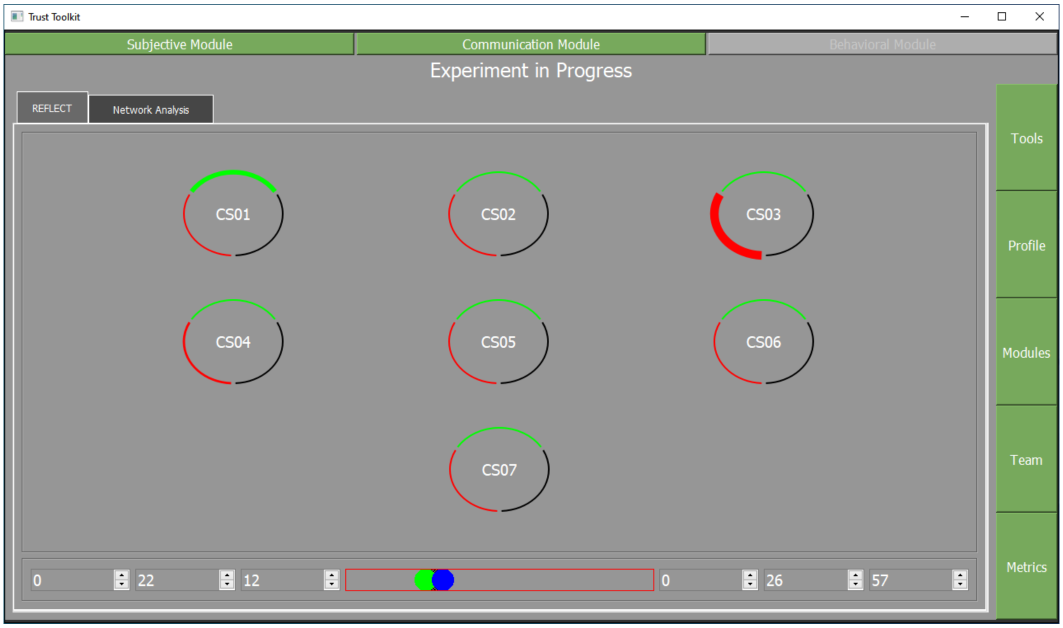
CS04 4%

CS05 3%

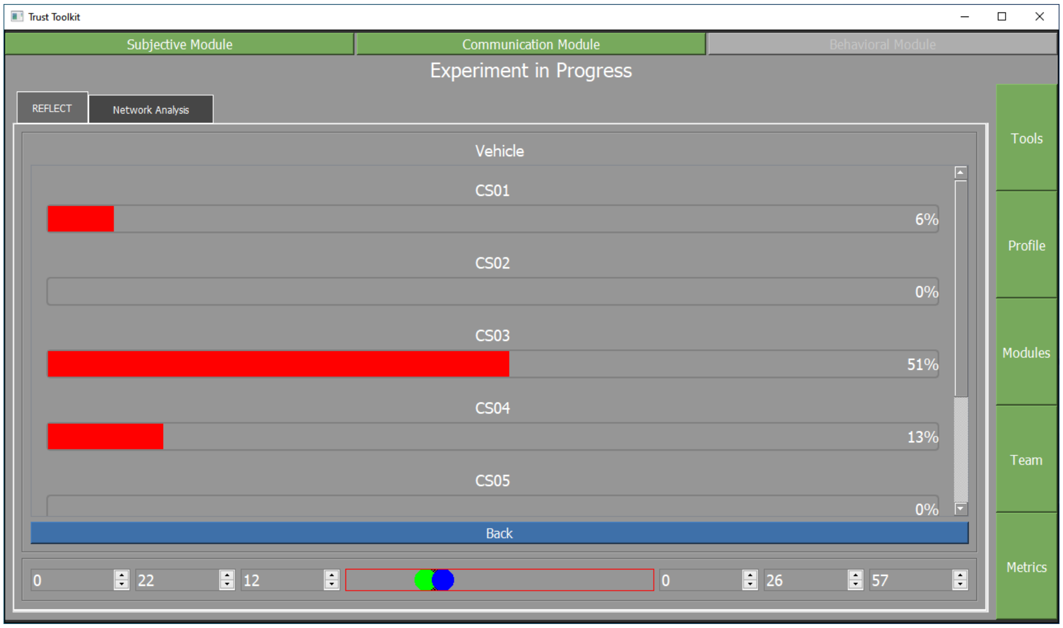
Step 7: Initial statistics for the entire experiment run are displayed. The time selector at the bottom can be used to view statistics for only a section of the experiment. Drag the green and blue sliders to set the start and end time, or manually enter the desired times in the boxes on the left (start) and right (end).



A view with a short time window selected (~5 min). Here, CS03 is speaking over 50% of the time window on the Vehicle network. CS01 is speaking about 28% of the time on the Command network.



More detail is shown by clicking on the Vehicle network.



Vehicle

CS01 6%

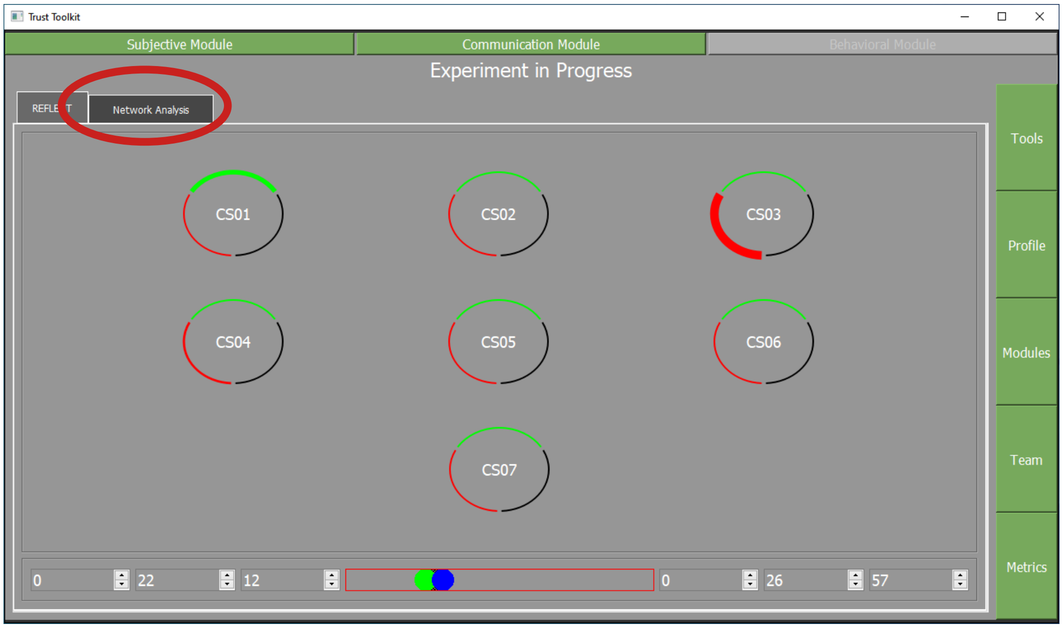
CS02 0%

CS03 51%

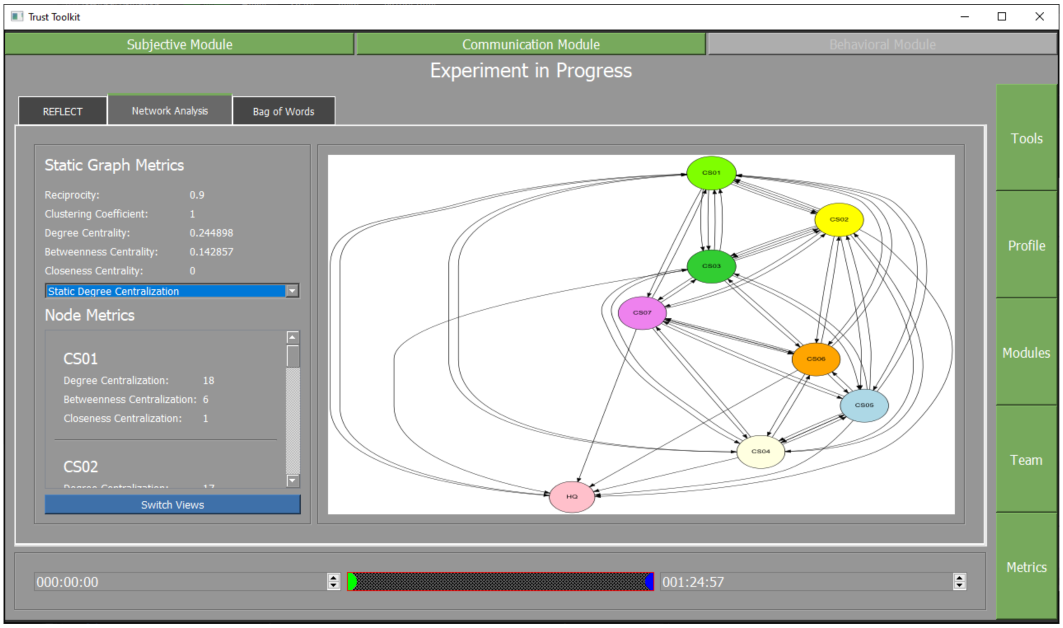
CS04 13%

CS05 0%

Step 8: Click on the “Network Analysis” tab to view the “Social Network Analysis” of the communications traffic.



The “Network Analysis” tab provides a view of the network topology and “Social Network” measures for each node including Degree Centralization, Betweenness Centralization, and Closeness Centralization. It also displays overall network metrics including Reciprocity, Clustering Coefficient, Degree Centrality, Betweenness Centrality, and Closeness Centrality. The drop-down box will scale the nodes based on their individual scores in one of the individual metrics.



Trust Toolkit

Experiment in Progress

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

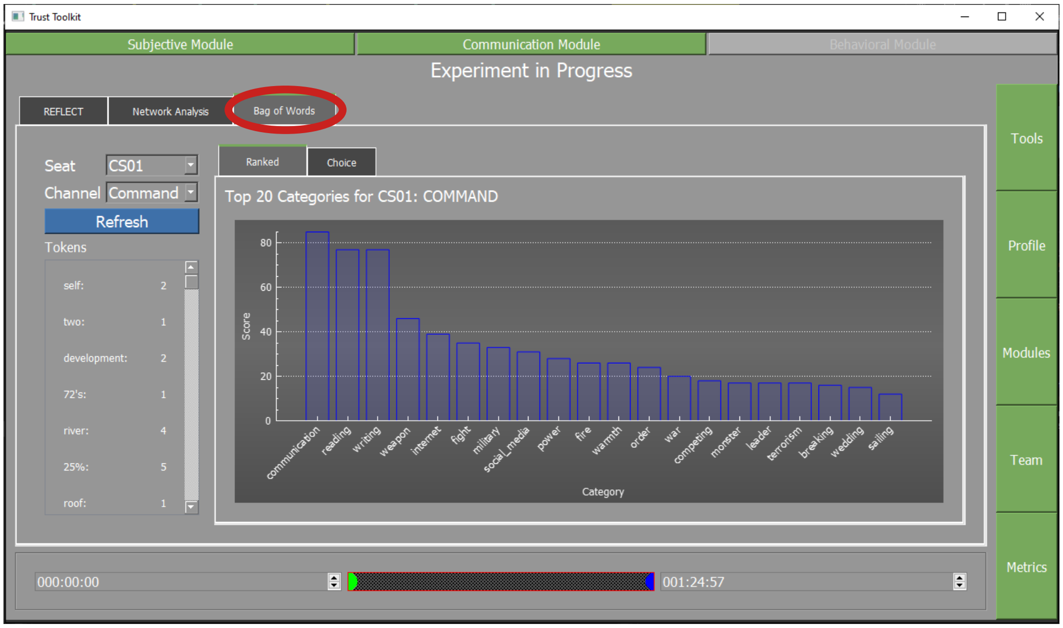
CS02

000:00:00

001:24:57

The “Bag of Words” tab provides analysis of the semantic content of the words uttered by participants by categorizing the words and showing the word count for each category. This can be accessed by clicking on the “Bag of Words” tab.

Initially the toolkit will display the scores for the top 20 categories for the first seat (CS01 or Crew Station 01 in this case) and the first network (the Command network in this case). The left side of the graph shows the individual words (i.e., Tokens) that were uttered and their count.



Seat CS01

Channel Command

Tokens

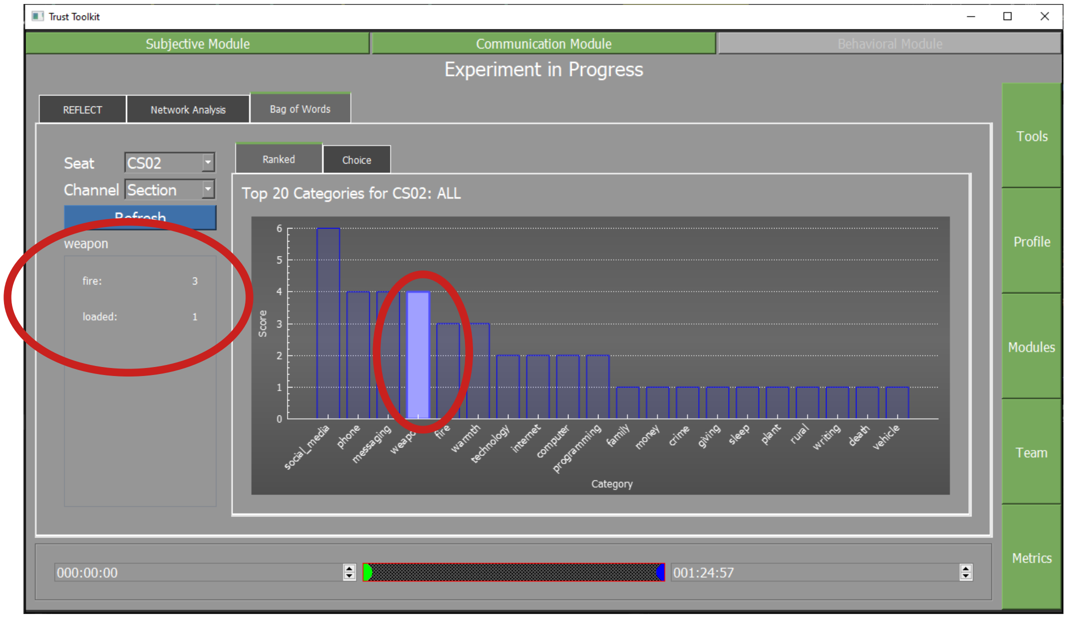
self:	2
two:	1
development:	2
72's:	1
river:	4
25%:	5
roof:	1

000:00:00

001:24:57

Initially the toolkit will display the scores for the top 20 categories for the first seat (CS02 or Crew Station 02 in this case) and the first network (the Command network in this case). The left side of the graph shows the individual words (i.e., tokens) that were uttered and their count.

By clicking on a category, the token display will change to show only words that were counted in that category.



Seat CS02

Channel Section

Top 20 Categories for CS02: ALL

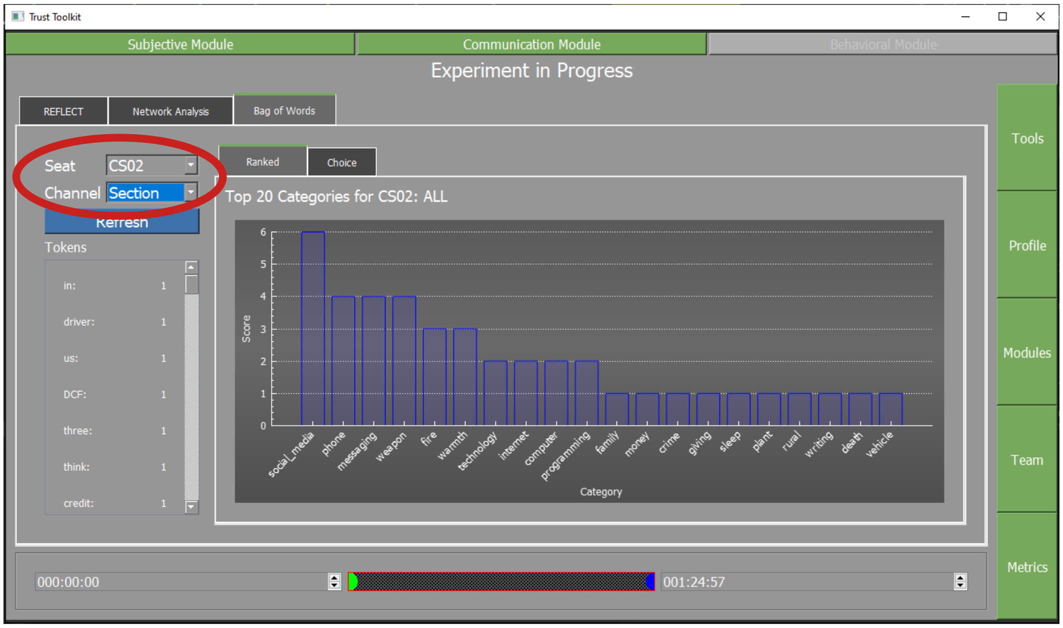
weapon

fire:	3
loaded:	1

000:00:00

001:24:57

Displays for each participant and each network can be selected using the “Seat” and “Network” drop-down menus.



Seat CS02

Channel Section

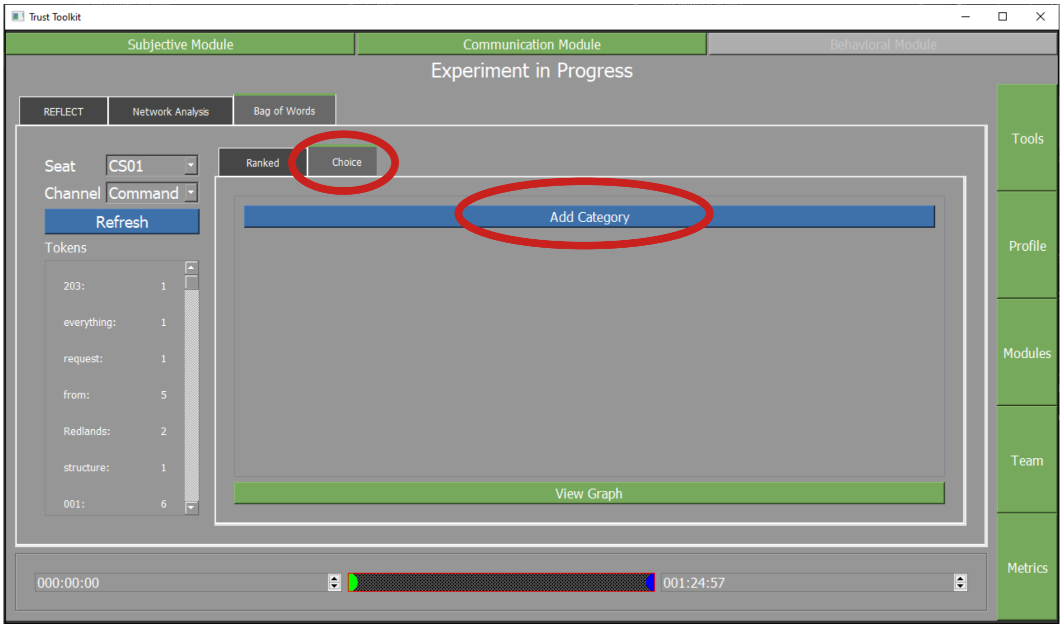
Tokens

```
in:       1
driver:   1
us:       1
DCF:      1
three:    1
think:    1
credit:   1
```

000:00:00

001:24:57

The “Ranked” tab displays the top 20 categories ranked by score. The “Choice” tab allows the user to configure a specific set of categories they would like to view. Click on the “Choice” tab to switch to “Choice” mode and then click “Add Categories” to choose categories to view.

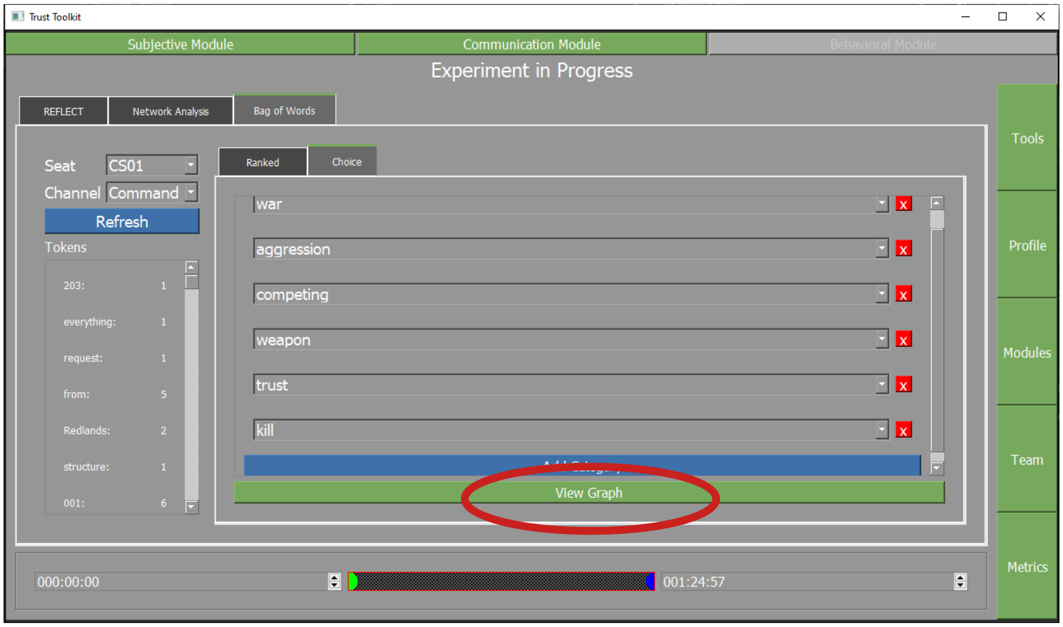


Seat CS01
Channel Command

203:	1
everything:	1
request:	1
from:	5
Redlands:	2
structure:	1
001:	6

000:00:00
001:24:57

Additional categories can be selected by pressing “Add Category” for each category desired and selecting it from the drop-down menu. Click the red “X” button to remove a category. When all desired categories have been selected, press “View Graph”.



Seat CS01

Channel Command

203:	1
everything:	1
request:	1
from:	5
Redlands:	2
structure:	1
001:	6

war ☒

aggression ☒

competing ☒

weapon ☒

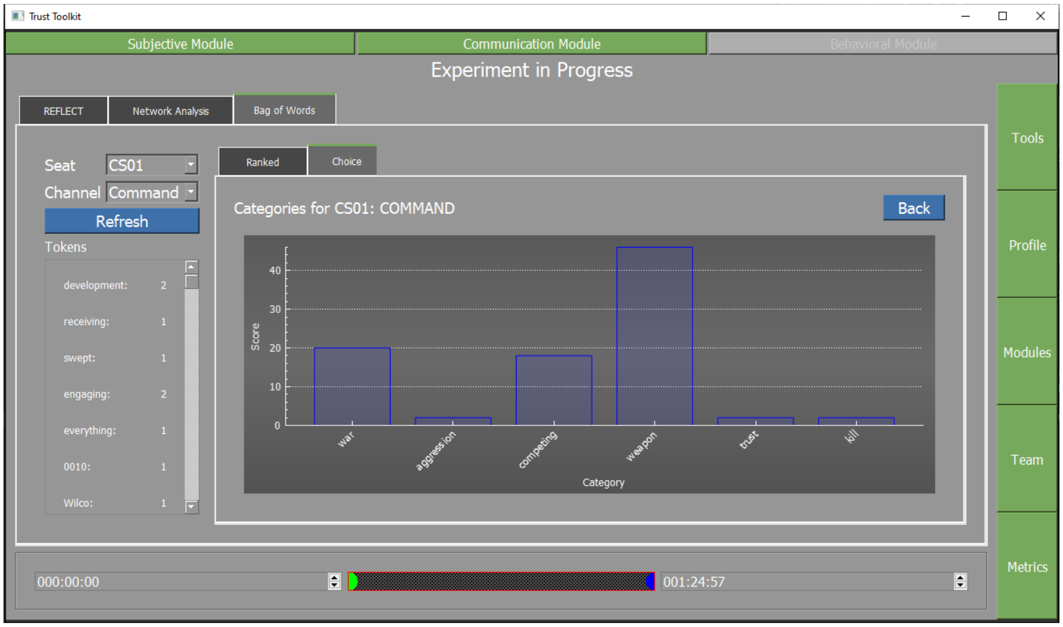
trust ☒

kill ☒

000:00:00

001:24:57

The category “Choice” graph displays.



Seat CS01

Channel Command

Tokens

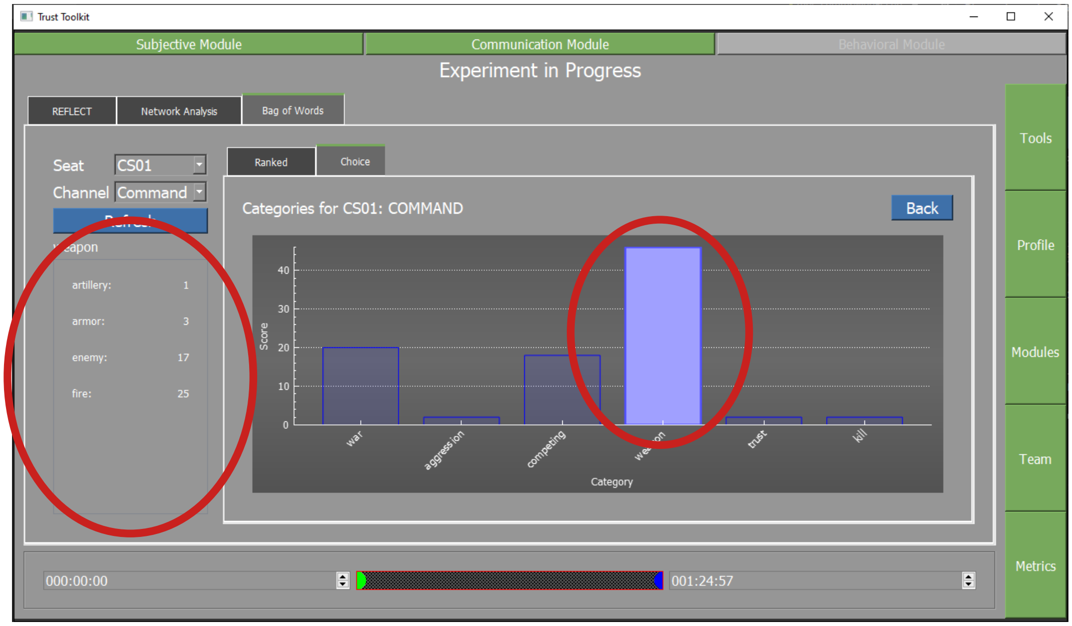
development:	2
receiving:	1
swept:	1
engaging:	2
everything:	1
0010:	1
Wilco:	1

Categories for CS01: COMMAND

000:00:00

001:24:57

As with the “Ranked” graph, clicking on an individual category’s bar will display the words scored in that category in the token area on the left.



Seat CS01

Channel Command

weapon

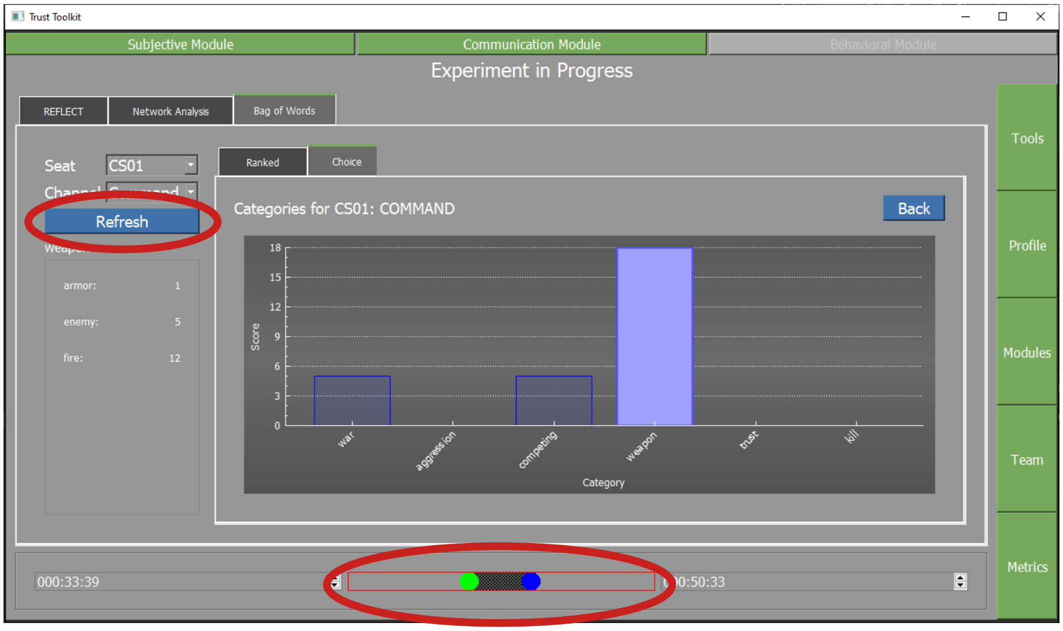
artillery:	1
armor:	3
enemy:	17
fire:	25

Categories for CS01: COMMAND

000:00:00

001:24:57

Initially, the "Bag of Words" module shows statistics for the entire time window of the scenario. Like other tools in the Communication Module, the time window can be scaled using the time slider at the bottom. For performance reasons, the graph does not automatically update when a new time window is selected. After selecting a new time window, press the blue "Refresh" button and the graph will update within a few seconds.



Seat CS01

Channel Command

Weapon

armor: 1

enemy: 5

fire: 12

Categories for CS01: COMMAND

000:33:39

00:50:33

List of Symbols, Abbreviations, and Acronyms

ARL	Army Research Laboratory
DEVCOM	US Army Combat Capabilities Development Command
GAART	Global After-Action Review Technology
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
IA	Intelligent Agent
INFORMS	Information for Mixed-Squads
RCV	robotic combat vehicle
REFLECT	Real-Time Event, Flow, and Coordination Tool
TEV&V	test, evaluation, verification, and validation
WMI	Warfighter Machine Interface

1	DEFENSE TECHNICAL
(PDF)	INFORMATION CTR
	DTIC OCA
1	DEVCOM ARL
(PDF)	FCDD RLD DCI
	TECH LIB
4	DEVCOM ARL
(PDF)	FCDD RLH FA
	S FITZHUGH
	C NEUBAUER
	FCDD RLH FD
	A BAKER
	A KRAUSMAN