ARL-TN-1294 ● FEB 2026

Addendum to ARL-TR-9623, *Human-Autonomy Teaming Trust Toolkit (HAT^3) Software Development Documentation and User Guide*

by Catherine Neubauer, David Chhan, Sean M. Fitzhugh, Anthony L. Baker, Andrea S. Krausman, Murat Kucukosmanoglu, Bret Kellihan, and Justin Jagielski

NOTICES

Disclaimers

ARL-TN-1294 ● FEB 2026

Addendum to ARL-TR-9623, *Human-Autonomy Teaming Trust Toolkit (HAT^3) Software Development Documentation and User Guide*

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

Murat Kucukosmanoglu
D-Prime LLC

Bret Kellihan and Justin Jagielski
DCS Corporation

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED START DATE	3. DATES COVERED END DATE
February 2026	Technical Note	September 2024	September 2025

4. TITLE AND SUBTITLE
Addendum to ARL-TR-9623, *Human-Autonomy Teaming Trust Toolkit (HAT³) Software Development Documentation and User Guide*

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Catherine Neubauer, David Chhan, Sean M. Fitzhugh, Anthony L. Baker, Andrea S. Krausman, Murat Kucukosmanoglu, Bret Kellihan, and Justin Jagielski

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
DEVCOM Army Research Laboratory ATTN: FCDD-RLA-FA Aberdeen Proving Ground, MD 21005	ARL-TN-1294

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
DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

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

14. ABSTRACT
In ARL-TR-9623, we presented several multimodal measures for assessing team trust that were incorporated into a software tool called the Human-Autonomy Teaming Trust Toolkit (HAT³). The initial software document and user guide described the purpose, scope, assessment modules, and design of the HAT³ and provided instructional details and screenshots regarding how to use the two existing modules: subjective and communication. Since the initial report, several new capabilities for assessing team trust, cohesion, and workload have been added. The purpose of this addendum is to describe these new capabilities.

15. SUBJECT TERMS
team trust, human–autonomy teaming, software development, trust measurement, Humans in Complex Systems

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	C. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UU	24

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
Catherine Neubauer	(520) 691-5046

STANDARD FORM 298 (REV. 5/2020)
Prescribed by ANSI Std. Z39.18

Contents

1. Introduction

In ARL-TR-9623,[1] we presented an overview of a software assessment tool called the Human-Autonomy Team Trust Toolkit (HAT^3) that we developed in response to the need for novel, multimodal assessment methods to ensure proper team trust calibration within the context of human–autonomy teams.[2] Specifically, we provided a software document and user guide describing the purpose, scope, components, and system design of the HAT^3. The objective of the intial Technical Report (TR) was to provide readers, developers, and potential first-time users of this system with background information, highlighting the impetus for the HAT^3, its purpose, and a thorough understanding of its functionality. Finally, we included an instructional guide with screenshots for how to use the two existing modules: subjective and communication. The goal of including these types of measures was to enable a more comprehensive assessment of team trust with methods to capture trust before, during, and after an interaction, using both discrete and continuous measures.

Since the initial TR was published, additional capabilities for assessing and predicting trust were developed, and we expanded the toolkit to model and predict other team states such as workload and cohesion, which are also critical for effective human–autonomy teams. These new models incorporate additional modalities, such as eye tracking and physiological sensing, which have been called out in the literature for their relationship to various team states. Thus, the purpose of this addendum is to provide details about these new capabilities and present an enhanced user guide that describes how to use and understand these additional modules in more detail.

2. Sentiment Analysis Model

The Sentiment Analysis Model was an extension of the original communication module of the HAT^3. This model was designed to quantify the emotional tone of military team communications and link those measures to trust in both teammates and AI systems. The model development leveraged a transformer-based natural language processing (NLP) architecture, specifically DistilBERT,[3] which was fine-tuned for the specialized language of military missions. DistilBERT was ultimately selected as the sole model due to its superior accuracy and computational efficiency. The pipeline consisted of several key stages:

1. Data preparation: Audio communications from simulated combat missions were transcribed using the Whisper speech recognition engine and

segmented into short intervals. Transcripts were cleaned to remove inconsistencies and prepared for annotation.

2. Annotation: Communication segments were manually labeled by trained annotators from the U.S. Army Combat Capabilities Development Command Army Research Laboratory as positive, negative, or neutral. Labels captured both emotional tone (e.g., frustration, encouragement) and operational tone (e.g., mission success or failure). High inter-annotator agreement ensured reliable training data.
3. Model fine-tuning: The DistilBERT model was fine-tuned on the annotated military communication corpus using the Hugging Face Transformers library.[4] Fine-tuning allowed the model to adapt to military-specific jargon and phrasing, which conventional sentiment models typically misinterpret.

The fine-tuned model outputs probabilities for each sentiment class (positive, negative, neutral) for every 30-s communication window. In the HAT[3] implementation, these probabilities are used directly to characterize the distribution of sentiment expressed across teams. This approach allows sentiment predictions to be integrated with other modalities such as eye-tracking and other data streams to support trust monitoring and modeling. Screenshots for this model are shown in the Appendix.

3. The Physio Trust Cohesion Model

The Physio Trust Cohesion model was developed to add an additional physiological sensing module to HAT[3]. The model estimates team cohesion on a scale from 1 to 7 where a score of 1 indicates the lowest cohesion and a score of 7 indicates the highest cohesion. The model was trained using the open-sourced automated machine learning (AutoML) library called AutoGluon.[5] AutoGluon is a state-of-the-art AutoML framework specializing in hyperparameter optimization and multilayer stacking that combines both the traditional ML models and deep learning techniques to ensure the models' robustness and accuracy. The AutoGluon Python package fits a variety of models that include gradient boosting, random forests, K-nearest neighbors, and neural network approaches. It then fits a weighted ensemble method that combines predictions from each model and stacks them in layers to achieve superior accuracy.

This ensemble approach is used as the final model for predictions. The model was trained using input features extracted from continuous multimodal physiological and behavioral data (i.e., sentiment scores derived from transcribed audio communication, heart rate and heart rate variability-related measures derived from electrocardiogram [ECG] signals, and pupil and gaze statistics derived from eye-tracking data). The cohesion labels used in the model training were obtained from self-reported overall cohesion rating scores ranging from 1 to 7. Screenshots for this model are shown in Appendix A.

4. Physio Workload

The Physio Workload Model was developed to estimate high cognitive workload from ECG features using a neural network classifier. Input features consisted of descriptive statistics of the ECG, including mean, maximum, and minimum heart rate as well as heart rate variability indices. These features were extracted in real time from 30-s sliding windows, standardized, and then passed to the workload classifier.

The classifier was implemented as a feedforward neural network designed for binary classification of workload state (high vs. low workload). The network architecture comprised two hidden layers of 20 nodes each, with nonlinear activations, and a final sigmoid output layer producing probabilities between 0 and 1. This probability represents the model's estimate of whether the individual is experiencing a high workload state.

In the HAT^3 implementation, ECG features are continuously extracted and standardized in real time, and these processed features are provided to the model to generate workload probabilities. This approach allows workload predictions to be integrated with other physiological and behavioral data streams, supporting real-time monitoring and modeling of trust and cognitive states. Screenshots for this model are shown in the Appendix.

5. Physio Trust Model

The Physio Trust Model was developed to predict subjective team trust ratings using eye features. Trust labels were derived from post-mission questionnaires and binarized: responses ≥5 were classified as "high trust," whereas responses <5 were labeled as "low trust." This framing enabled the use of a supervised ML classifier targeting binary trust outcomes.

Model inputs consisted of eye-tracking features, including pupil size, blink counts, saccade velocity, and fixation metrics. In the HAT^3 implementation, eye data are processed in 10-window segments to generate features, which are then standardized and passed to the classifier. The classifier was implemented using a random forest algorithm.

The model outputs a probability between 0 and 1, representing the likelihood of high trust, which can be monitored continuously in real time. In practice, the Physio Trust Model integrates eye-tracking inputs and produces real-time probabilities of high trust, enabling monitoring of team trust and supporting adaptive feedback and interventions within HAT^3. Screenshots for this model are shown in the Appendix.

6. Conclusion

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. Part of this work will involve continually adding measures and modules to HAT3 to better equip users with appropriate and robust measures of individual and team-level latent states. This addendum serves as a document and user guide to further outline the additional capabilities for assessing and predicting trust as well as models of other team states such as workload and cohesion, which are also critical for effective human–autonomy teams. These new models incorporate additional modalities, such as eye tracking and physiological sensing, for their fieldability and ease of use, and also for their relationship to various team states that have been identified in previous research and literature. Thus, the purpose of this addendum is to provide details about these new capabilitlies and present an enhanced user guide that describes how to use and understand these additional modules in more detail.

Finally, continued research building on this toolkit has supported 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.

7. References

1. Neubauer C et al. Human-autonomy teaming (HAT3) trust toolkit software development documentation and user guide. DEVCOM Army Research Laboratory (US); 2022 Dec. Report No.: ARL-TR-9623.

2. Krausman A et al. Team trust measurement in human-autonomy teams. Trans Hum-Robot Interact. 2022;11(3):1–58.

3. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv; 2020 Mar 1. arXiv:1910.01108. https://doi.org/10.48550/arXiv.1910.01108

4. Wolf T et al. HuggingFace's transformers: state-of-the-art natural language processing. arXiv; 2020 July 14. arXiv:1910.03771. https://doi.org/10.48550/arXiv.1910.03771

5. Erickson N et al. AutoGluon-Tabular: robust and accurate AutoML for structured data. arXiv; 2020 Mar 13. arXiv:2003.06505. https://doi.org/10.48550/arXiv.2003.06505

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

A.1 Introduction to HAT^3 User Guide

The U.S. Army Combat Capabilities Development Command Army Research Laboratory HAT^3 is a collection of tools for accessing 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. In recent efforts, a physiological module that provides 1) visualizations of raw physiological data, 2) predictive models of trust and workload, and 3) sentiment analysis have also been added. This addendum to the original user's guide[1] provides screenshots of these newly developed modules and models with explanations. It assumes that initial installation of the toolkit and associated dependencies have already been performed on all necessary computers.

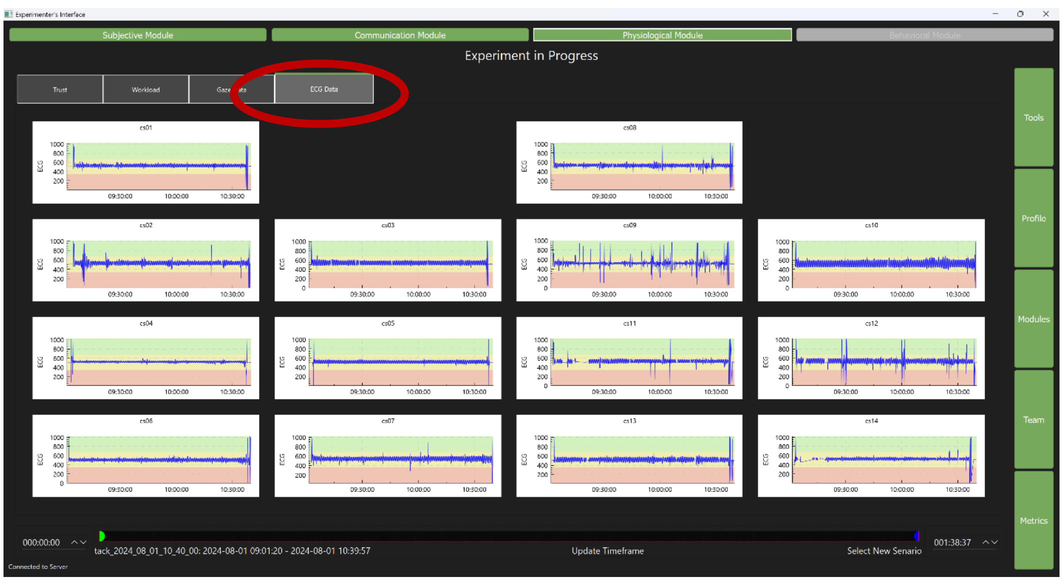
Several enhancements covered in this addendum are contained in the new Physiological Module of the HAT^3. To access this module, click on the "Physiological Module" button at the top of the screen.

[1] Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv; 2020 Mar 1. arXiv:1910.01108. https://doi.org/10.48550/arXiv.1910.01108

A.2 Raw Electrocardiogram (ECG) Data

Physio data from the BioHarness devices is ingested into the HAT^3 for use in computing metrics. The ECG waveform is one of the primary modalities used in derived metrics. For this reason, the option to view the raw waveform is included in the toolkit. The time window of data viewed is controlled by the time slider at the bottom of the HAT^3 window. At longer timescales, the main use for this view is identifying gaps in recorded data. During an experiment, this can also be used to monitor incoming data and detect when a sensor goes offline, providing the opportunity to address the issue and minimize data loss.

To view the raw ECG waveform, click the “ECG Data” button in the physio module.

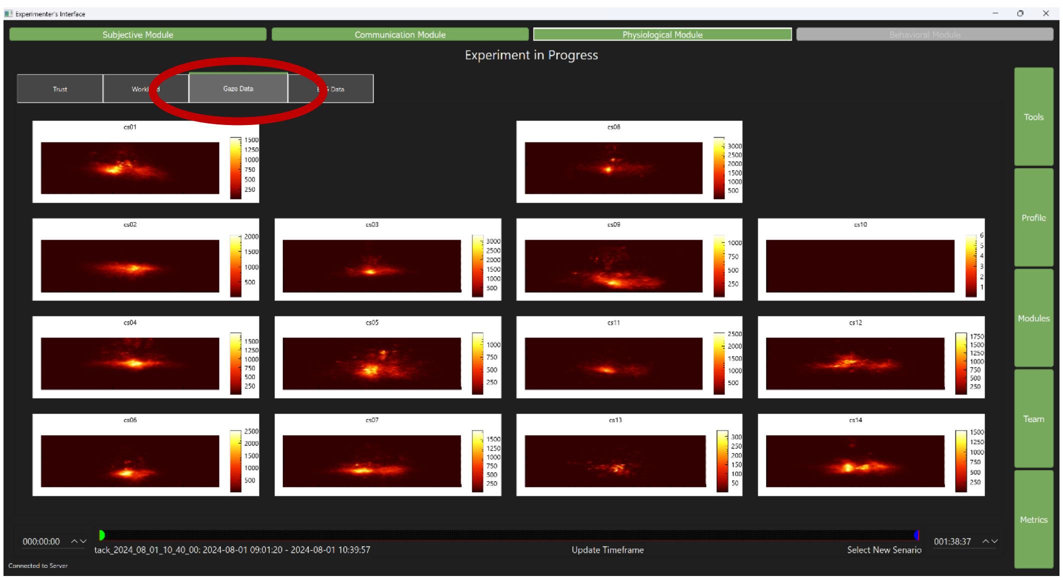


Due to the large amount of data points, loading the raw ECG data for display here can impact performance of HAT^3 on some systems. For this reason, there is a configuration option to disable this tab in the configuration file. If the ECG Data tab is not visible, it will need to be enabled in the configuration file.

A.3 Raw Eye Gaze Data

Like the BioHarness, data from Tobii eye-tracking devices is ingested into the HAT^3 for use in computing metrics. The eye data is one of the primary modalities used in derived metrics. For this reason, the option to view the raw plot of eye gaze is included in the toolkit. This is displayed as a heatmap of eye gaze over the time window selected using the time slider at the bottom of the HAT^3 window. Because this is a head-mounted device and the gaze here is not referenced to the environment but relative to the head position, the main use for this view is to identify gaps in recorded data. During an experiment, this can be used to monitor incoming data and detect when a sensor goes offline, providing the opportunity to address the issue and minimize data loss.

To view the raw gaze data, click the "Gaze Data" button in the physio module.

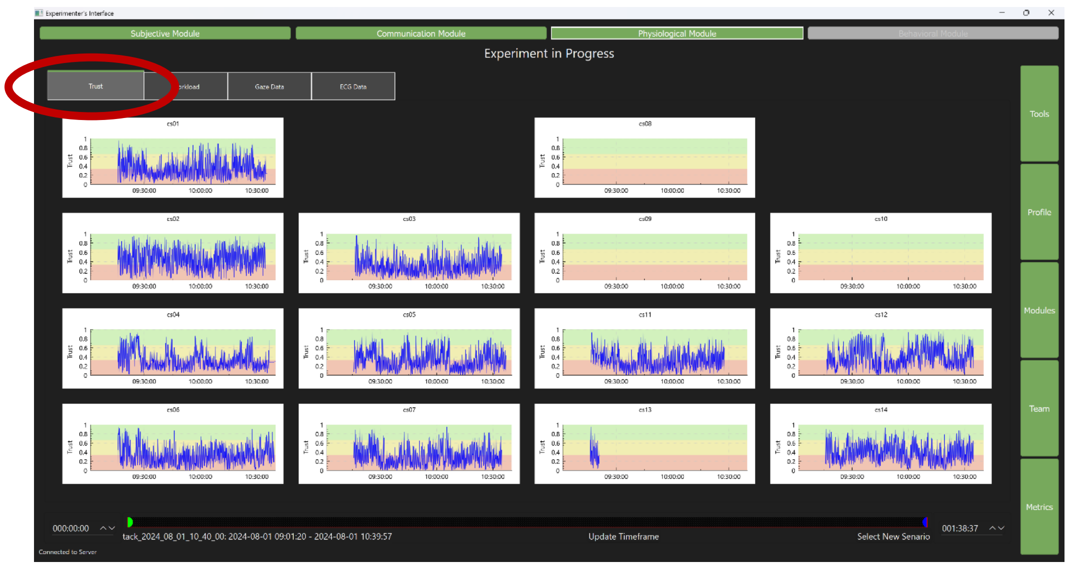


Due the large amount of data points, loading the raw gaze data for display here can impact performance of HAT^3 on some systems. For this reason, there is a configuration option to disable this tab in the configuration file. If the Gaze Data tab is not visible, it must be enabled in the configuration file.

A.4 Physio Trust Model

The Physio Trust Model predicts subjective team trust ratings using eye features. Model inputs consist of eye-tracking features, including pupil size, blink counts, saccade velocity, and fixation metrics. In the HAT^3 implementation, eye data are processed in 10-window segments. The model outputs a probability between 0 and 1, representing the likelihood of high trust, which can be monitored continuously in real time.

To view the Physio Trust Module, click the "Trust" button within the Physiological Module.



A.5 Physio Workload

The Physio Workload Model was developed to estimate high cognitive workload from ECG features using a neural network classifier producing probabilities that an individual is experiencing a high workload state between 0 and 1.

In the HAT^3 implementation, ECG features are continuously extracted and standardized in real time, and these processed features are provided to the model to generate workload probabilities.

To view the Physio Workload model, click the “Workload” button within the Physiological Module.

A.6 Physio Trust Cohesion

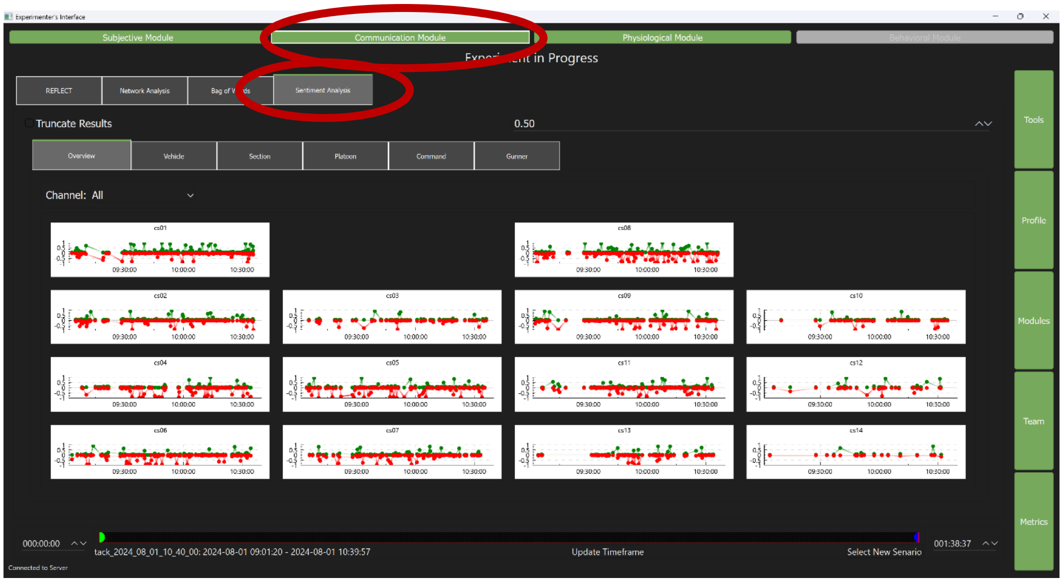
The Physio Trust Cohesion model was developed to add an additional physiological sensing module to HAT[3]. The model estimates team cohesion on a scale from 1 to 7 where a score of 1 indicates the lowest cohesion and a score of 7 indicates the highest cohesion.

To view the Physio Trust Cohesion model, click the “Trust Cohesion” button in the Physiological Module.

A.7 Sentiment Analysis Model

The Sentiment Analysis model was an extension of the original communication module of HAT[3]. This model was designed to quantify the emotional tone of military team communications and link those measures to trust in both teammates and AI systems.

To access the Sentiment Analysis model, first select the Communications Module using the button at the top of the Trust Toolkit. Then click on the Sentiment Analysis tab.



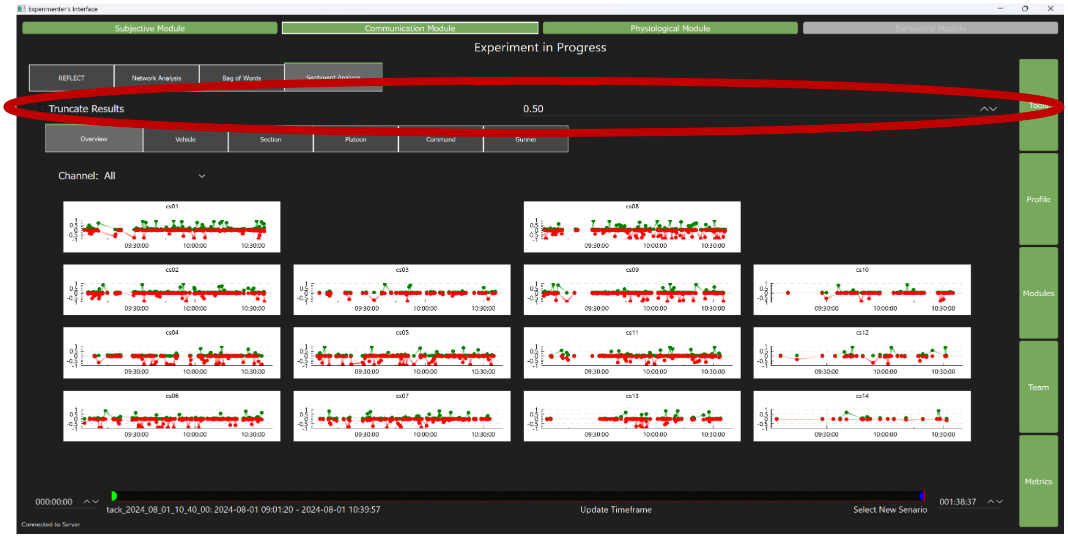
A.7.1 Sentiment Analysis Overview

Recall that the Sentiment Analysis model computes three probabilities for each group of utterances being analyzed: positive, negative, and neutral. These probabilities range from 0 to 1 and all three probabilities sum to 1.

To view this data in an intuitive manner, we chose to plot only the probabilities that each sentiment is positive and negative. The neutral probability is not displayed in this module; however, it can be easily inferred from the plot.

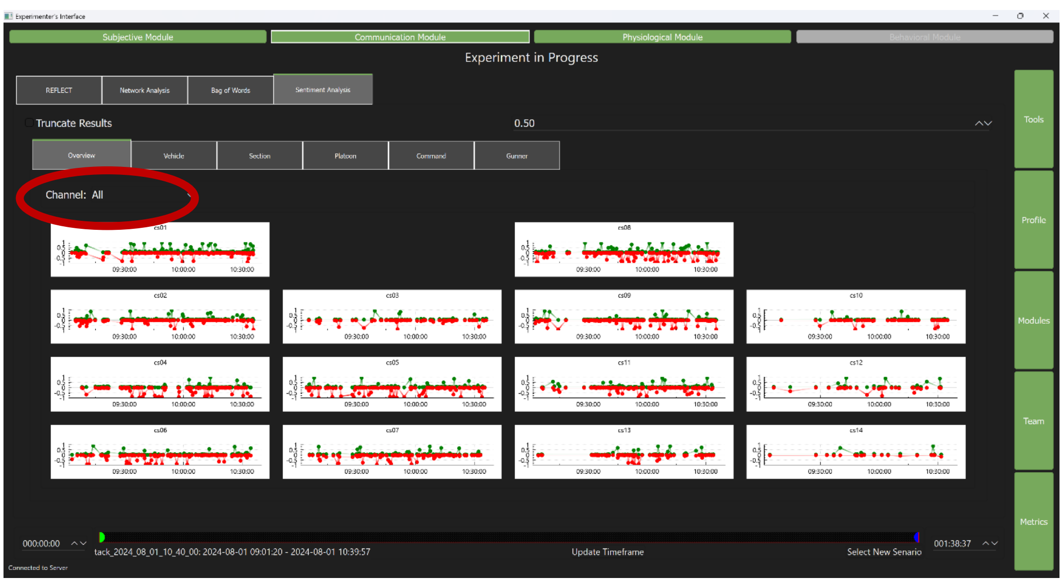
The probability that the sentiment is positive is plotted in green on the positive y-axis. The probability that the sentiment is negative is plotted in red on the negative y-axis. Upward green spikes indicate a point where there is a high probability of positive sentiment, and low red spikes indicate a point where there is a high probability of negative sentiment. Points of both colors clustered along the axis indicate a high probability of neutral sentiment.

To fine tune these graphs, there is an option to collapse probabilities below the specified threshold to 0. This threshold can be changed by changing the "Truncate Results" value.



There are several options for visualizing the data in the Sentiment Analysis model. Initially, an overview is displayed, broken down by individual subject, including data from all communications channels. There are 14 subjects identified by their crewstation ID labeled cs01 through cs14.

To restrict the visualizations to data from a specific channel, select the desired channel in the channel selector.



Like the other tools in the Communications and Physiological Modules, the time window slider at the bottom of the Trust Toolkit window controls the time frame of data being displayed.

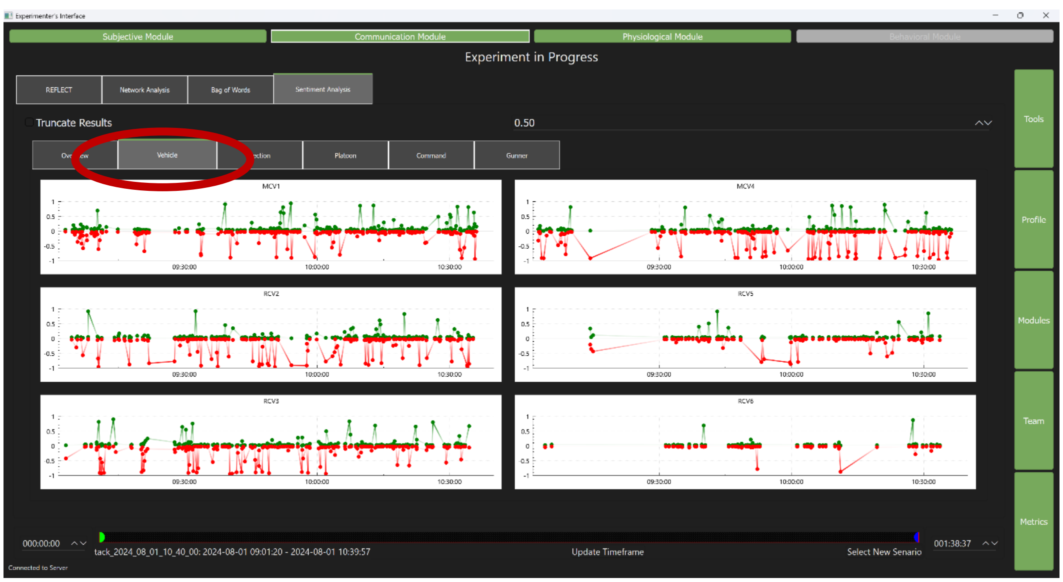
In views with more than one graph, a graph can be clicked to see a larger interactive graph. Clicking on a data point in the interactive graph provides information on the subject, the utterance that was evaluated, and all three probabilities from the

Sentiment Analysis model. In this example, the initial graphs in the Overview, Vehicle, and Section views can be clicked to view a larger interactive graph. The interactive graphs is described in detail in the Platoon section (A.7.4).

Based on the communications channels that are configured during experiment setup, HAT[3] will offer visualizations for groupings of data for each channel. The examples here are specific to the experiment configuration and may be different based on communications channels defined in a specific experimental environment.

A.7.2 Sentiment Analysis Vehicle

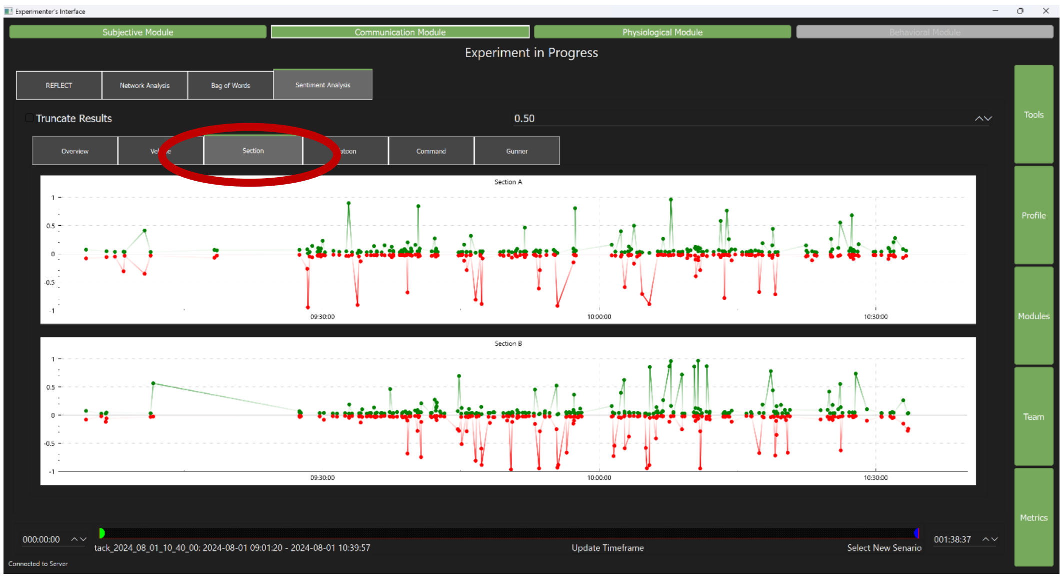
In the experiment shown here, crewmembers (subjects) are assigned to vehicles they are responsible for operating. Everyone assigned to a vehicle has access to a communications channel specific to it that can be used to talk with others assigned to the same vehicle. Therefore, HAT[3] provides visualizations of the sentiment analysis for every utterance on each channel. Multiple speakers are represented on each graph in these channel views.



A.7.3 Sentiment Analysis Section

In this experiment, the vehicles were grouped into units called Section A and Section B, which comprised three vehicles each. Each of these Sections has a separate communications channel for their section. This view graphs sentiment analysis for each utterance on that channel.

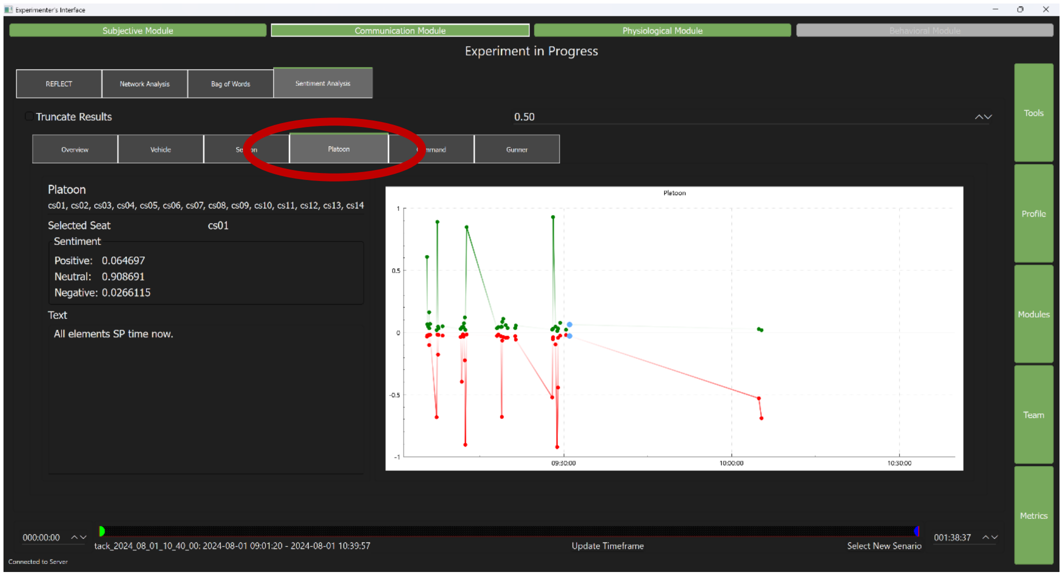
To look at the Section channel view, click the “Section” button within the Sentiment Analysis model.



A.7.4 Sentiment Analysis Platoon

In this experiment, the two Sections were further grouped into a single Platoon and there is a communications channel for talking to the entire platoon. All utterances on that channel are visualized here.

To view the Platoon channel graph, click the “Platoon” button in the Sentiment Analysis model.

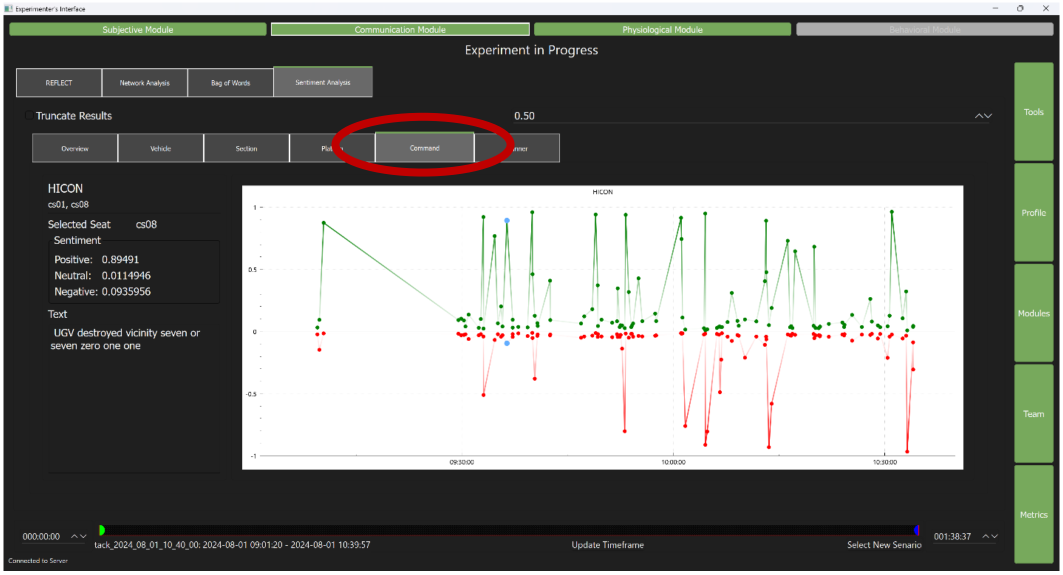


Because there is only a single graph, this is displayed as an interactive graph. To the left of the graph, there is an information display that shows the name of the channel and all subjects/seats that could access the channel. Clicking a data point on the graph will display the seat/subject that spoke the utterance, the three probabilities from the Sentiment Analysis model, and the transcribed text of the utterance itself in the display on the left. Remember, in the previous views with multiple graphs, an interactive graph like this can be viewed by clicking on one of the graphs.

A.7.5 Sentiment Analysis Command

In this experiment, a Higher Command (HICON) channel was available for talking with experimenters playing the role of units above the platoon. All utterances on that channel are visualized in this view.

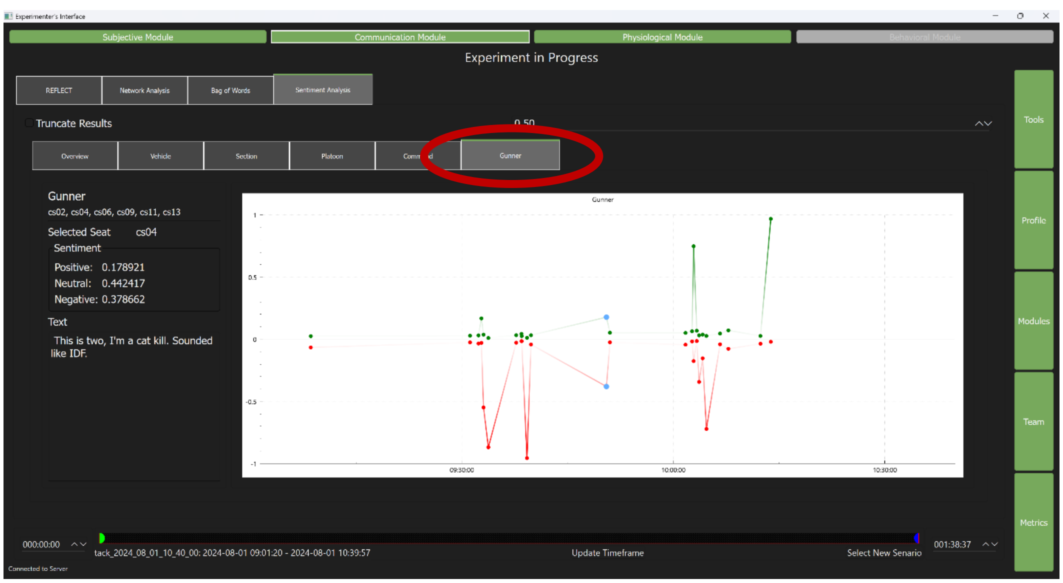
To view this channel, click the “Command” button in the Sentiment Analysis model.



A.7.6 Sentiment Analysis Gunner

In this experiment, a Gunner channel was available for talking with all subjects assigned to the Gunner role in all vehicles. All utterances on that channel are visualized in this view.

To view this channel, click the "Gunner" button in the Sentiment Analysis model.



Like the Platoon view, this is an interactive graph.

List of Symbols, Abbreviations, and Acronyms

AI	Artificial Intelligence
AutoML	Automated Machine Learning
DEVCOM	U.S. Army Combat Capabilities Development Command
ECG	Electrocardiogram
HAT^3	Human-Autonomy Team Trust Toolkit
HICON	Higher Command
ID	Identification
ML	Machine Learning
NLP	Natural Language Processing
TR	Technical Report

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
1 (PDF)	DEVCOM ARL FCDD RLB CI TECH LIB