Towards a Real-Time Model of Trust in Human-Machine Team Paradigms

Sydney Gibbs, Veronica Tanner, Eric Larson
Southern Methodist University
Dallas, TX
{sigibbs, vtanner, eclarson}@mail.smu.edu

Sandro Scielzo, Alvin Abraham
CAE USA
Arlington TX
{sandro.scielzo, alvin.abraham}@caemilusa.com

ABSTRACT

Trust in automation has long been a key human factor affecting individual and team performance. With the advent of synthetic teammates, where humans and machines must collaborate to ensure mission success, trust takes on an even more prominent role in mediating human-machine team interactions. Achieving calibrated trust—whereby humans appropriately trust an automated system or a synthetic teammate—is paramount to achieving team cohesion and performance. However, to be effective, trust needs to be measured in real-time and objectively to inform the system and automatically engage appropriate trust calibration methods to regain operator confidence. The current study describes the development of a preliminary model of trust based on a variety of operator biobehavioral markers engaged in a search and rescue mission where human operators supervise intelligent unmanned aerial vehicles assets in a constructive synthetic environment. Biometric sensor data from an eye tracker were used to develop a preliminary machine learning model of trust using dimensionality reduction and clustering analysis. Trust labels came from the Continuous Online Numerical Score measure and subject matter expert observations. Twenty-two participants, including unmanned aerial vehicle operators and novices, participated in this experiment. Operator trust was impacted by varying the quality of flight path recommendations provided by the system when searching for survivors. Results describe the reliability of a preliminary trust model, as well as correlations between biobehavioral markers, trust-related behaviors, and performance. This experiment is unique as it provides insights for developing effective ML-driven real-time and objective measures of trust. Study findings discuss additional steps required to validate and generalize a real-time model of trust.

ABOUT THE AUTHORS

Sydney Gibbs is a recent graduate from Southern Methodist University (SMU) with a B.S. in computer science and an M.S. in computer science with a cybersecurity focus. She is also a recent alumni of the professional engineering organization, “Theta Tau”. She has been on various projects with the Darwin Deason Institute since 2021.

Veronica Tanner is a recent graduate from Southern Methodist University (SMU) with a B.S. in computer science. She has been on various projects with the Darwin Deason Institute.

Dr. Eric Larson is a Bobby B. Lyle Endowed Professor in Engineering Innovation in the Computer Science Department in the Bobby B. Lyle School of Engineering, Southern Methodist University. His main research is in machine learning, sensing, and signal & image processing for ubiquitous computing applications, in particular, for biometrics and human cognition.

Dr. Sandro Scielzo is a Human Systems Technical Authority and Learning Science Fellow at CAE USA. Dr. Scielzo has over 20 years of experience researching next-generation training solutions for military and commercial applications. His current focus is on developing a virtual test range for validating human-machine team technologies.

Alvin Abraham is a modeling and simulation engineer at CAE USA supporting Human-Machine Teaming research focusing on human factors engineering and human systems integration with constructive synthetic environments to promote and accelerate applied human subject research.

Towards a Real-Time Model of Trust in Human-Machine Team Paradigms

Sydney Gibbs, Veronica Tanner, Eric Larson
Southern Methodist University
Dallas, TX
{sigibbs, vtanner, eclarson}@mail.smu.edu

Sandro Scielzo, Alvin Abraham
CAE USA
Arlington TX
{sandro.scielzo, alvin.abraham}@caemilusa.com

INTRODUCTION

The relationship between humans and machines is in the middle of an accelerated paradigm shift, going from machines seen as tools to automate any number of tasks, as determined by a human operator, to collaborative interactions where machines are more and more seen as teammates supporting complex mission goals. These interactions range from low-risk situations, such as college students using autonomous food delivery robots, to highly dynamic military environments, where operators use complex collaborative interfaces with synthetic agents to accomplish mission goals. All these interactions require some form of trust in these systems by humans. However, defining and measuring trust in an actionable way between humans and machines represents an existing gap in the research community (Gebru et al., 2022). Measuring trust objectively and reliably is key in mediating effective human-machine teaming (HMT) interactions (Kaplan et al., 2020). Moreover, there is a clear and present need to measure trust in real time to drive future adaptive Human Machine Interfaces (HMI) to optimize HMT collaboration, through, for example, trust-calibration methods (Scielzo et al., 2021). However, existing trust measurement methods can be limiting, such as subjective self-report trust measures, which only offer a small snapshot of trust in time. Objective methodologies can yield reliable trust measures but are time consuming and costly to collect. As a way of addressing these issues with current trust measurement methods, numerous government programs have been created to collect and analyze trust data (e.g., OP TEMPO, 2023). These efforts provide a valuable approach to assessing, in real time, performance-related constructs such as trust via biobehavioral markers and provides a means for developing objective prediction models of proficiency within a team to enhance mission readiness.

This paper leverages this promising approach by focusing on operator gaze patterns as objective behavioral markers for trust in a command and control paradigm. Specifically, the research presented here expands on the work of Hartzler et al. (2023) by exploring additional uninvestigated biometric data collected in their experiment aiming to relate operator gaze behaviors to levels of trust. Advancements in leveraging biometrics to infer mental constructs such as cognitive load and situational awareness (Scielzo et al., 2020; Wilson et al., 2021; Crothers et al., 2022) provide compelling evidence for the potential measurability of trust. Trust is influenced by numerous factors, including cognitive processes—therefore biometric indicators reflecting changes in cognitive load and/or situational awareness may also signify shifts in trust levels. For instance, heightened cognitive load or decreased situational awareness may indicate decreased trust in the system, as users may feel uncertain or overwhelmed by the task or environment. Conversely, reduced cognitive load or increased situational awareness may suggest higher levels of trust, as users feel confident in their interactions (Hartzler et al., 2023). By leveraging similar biometric cues that capture cognitive and situational states, this research explores the feasibility of inferring trust in human-machine teaming scenarios. Specifically, it focuses on eye tracking as a primary biometric quantity for understanding trust. This paper analyzes data of human subjects experiments from Hartzler et al. (2023) that measure user trust in a system while working to command unmanned aerial vehicles (UAVs) in search and rescue missions. In the context of these experiments, the following hypotheses regarding gaze tracking and trust are investigated:

Hypothesis One, Fixation Duration: Individuals with higher trust in the system will exhibit longer gaze fixation durations on critical interface elements. That is, higher trust levels result in deliberate and prolonged visual engagement with key interface elements. Conversely, lower trust levels may result in shorter gaze fixation durations or frequent shifts in gaze, reflecting hesitancy or skepticism in the system.

Hypothesis Two, Gaze Distribution and Gaze Transition: Individuals with higher trust in the system will demonstrate a more focused and centralized distribution of gaze across task-related interface elements, indicating efficient utilization of system functionalities. In contrast, lower trust levels may result in erratic gaze distributions, as users explore various interface components in search of reassurance or alternative options. This hypothesis suggests that trust influences the spatial allocation of visual attention during interaction with the interface. In this case, higher trust gaze maps are expected to display consistency, while lower trust levels are more varied and less consistent. Similarly, different levels of trust would be reflected in distinct gaze transition patterns between interface elements, with higher trust exhibiting smoother and more predictable gaze transitions.

Hypothesis Three, Classification and Modeling: Different levels of trust can be reliably classified by machine learning models, using the gaze distributions to infer trust into low and high levels. Furthermore, these distributions are similar enough across participants to generalize to unseen individuals. Thus, a scalable automated model of trust can be developed in the context of drone missions.

BACKGROUND

Numerous works have categorized trust. In a 2015 review, authors Hoff and Bashir presented a 3-layer model of trust between humans and machines. Those 3 layers are situational trust (confidence based on external factors such as type of system or perceived risk) and internal factors (such as self-confidence or mood), dispositional trust (belief in automation reliability), and learned trust (belief based on prior experiences (Hoff & Bashir, 2015). In these experiments, the change in situational trust is investigated.

In terms of works that attempted to automate the identification of trust levels, Hu et al. (2016) used a mix of brain computer interface signals (specifically a nine electrode EEG) and electro-dermal activity to assess trust in driving scenarios, showing that trust could be classified in 71% of responses. However, it was unclear how generalizable the results were to other scenarios. Building on this work, Choo and Nam (2022) developed a machine learning model of trust in the context of a system monitoring task. They employed a convolutional neural network (CNN) to analyze EEG signals, achieving 94% accuracy in the task. This gaze-based system leverages a wider array of biometric data to capture real-time trust levels, potentially providing a more comprehensive understanding of trust dynamics.

A recent survey on trust by Dizaji and Hu (2021) has highlighted various facets of trust building on the IMPACTS model, which delineates seven key features of trust: intention, measurability, performance, adaptivity, communication, transparency, and security. Specifically, their work highlights that there is a neglect of adaptivity for HMT in current research. That is, current works rarely incorporate features that allow systems to adjust dynamically to user needs or environmental changes. The work described in this paper attempts to address this through building a model of trust using automation from gaze biometrics data alone.

EXPERIMENTAL METHODOLOGY

Study Design

For the trust analysis experiments, the 2023 collected dataset described by Hartzler et al. (2023) is adopted and summarized as follows: This study consisted of 4 different simulated drone missions where participants used drones to find survivors of a category F-4 tornado. Once they found these survivors, they would then use the same drones to deliver supplies to randomized survivor locations. Every drone had a 100-meter field of view (FOV) radius and a speed of 25 m/s. The drones always started at the southern end of the map. Missions lasted 12 minutes with varying workloads (either low or high) and recommendation quality (either good or poor). Workload was manipulated using different areas of operations (AOs) and number of drones. Low workloads used 4 drones in a 1.85 square mile AO with 4 no-fly zones (NFZs), while high workloads used 8 drones in a 3.60 square mile AO with 6 NFZs. More drones and larger AOs led to an increase in cognitive load.

The second factor, recommendation quality, consisted of the usefulness of a provided drone-recommended flight path and resource delivery suggestion. Poor recommendations showed participants paths that took drones beyond the AO or through NFZs, while good recommendations showed participants clear paths to where survivors in need of resources were located. Missions had 4 distinct 3-minute phases. During each phase, participants would either see good or poor recommendations. Participants were not told that the quality of recommendations would change, and there were no indications on when recommendations would change. Recommendations were shown to participants on the left side of the screen. Participants could see 2-4 recommendations being presented at a time and were given the option to individually accept or reject each recommendation. Participants also had the option to reject all listed recommendations or accept a random one. If a participant didn’t make a choice within one minute of the recommendations being suggested, they were then marked as ignored. Figure 1 shows various user interface element zones and the map interface for a participant. An example gaze map is also overlaid showing the distribution of participant gaze for the experiment.

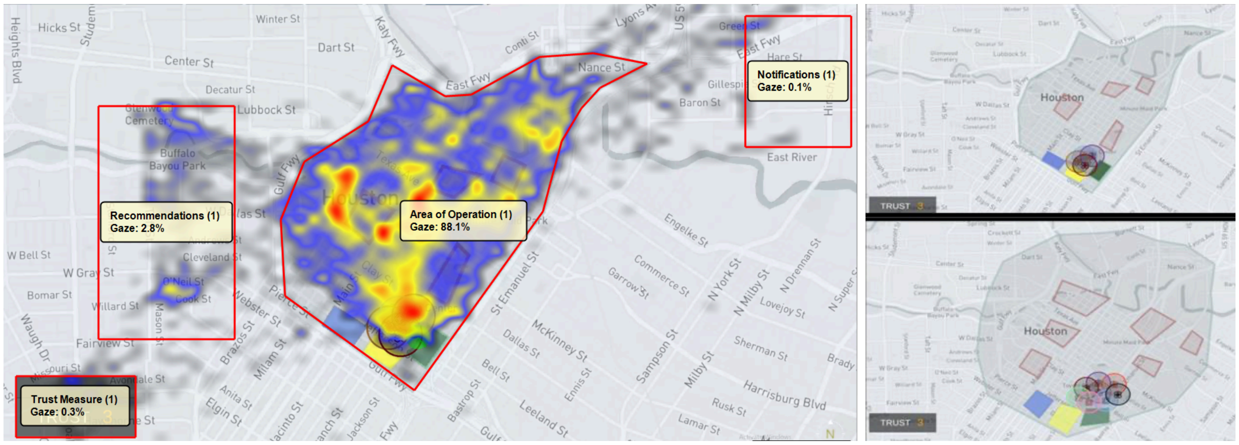

Figure 1: Overlay of gaze distribution for one participant and map user interface with interface regions in callouts (left) and the interface without gaze overlay for two different missions (right). Rightmost image reproduced with permission from Hartzler et al. (2023).

As shown in Table 1 below, the combination of workload (either low or high) and recommendations (either poor or good) led to four possible mission profiles. Mission 1 consisted of low workload, and a good starting recommendation ("G"). Mission 2 consisted of low workload, and a poor starting recommendation ("P"). Mission 3 consisted of high workload, and a good starting recommendation ("G"). Mission 4 consisted of high workload, and a poor starting recommendation ("P"). Participants completed each of the missions in one of four orders shown below.

Table 1: Summary of Mission Orders and Types, reproduced with permission from Hartzler et al., (2023).

	1st Scenario			**2nd Scenario**			**3rd Scenario**			**4th Scenario**		
Order	**Msn**	**Wrkld**	**Rec. Order**	**Msn**	**Wrkld**	**Rec. Order**	**Msn**	**Wrkld**	**Rec. Order**	**Msn**	**Wrkld**	**Rec. Order**
A	1	Low	G - P - G - P	3	High	G - P - G - P	2	Low	P - G - P - G	1	High	P - G - P - G
B	3	High	G - P - G - P	1	Low	G - P - G - P	4	High	P - G - P - G	3	Low	P - G - P - G
C	2	Low	P - G - P - G	4	High	P - G - P - G	1	Low	G - P - G - P	2	High	G - P - G - P
D	4	High	P - G - P - G	2	Low	P - G - P - G	3	High	G - P - G - P	4	Low	G - P - G - P

Note: Participants completed the missions ("Msn") in one of four orders, A-D. Missions differed by task workload ("Wrkld") and the order of recommendation types ("Rec. Order"), alternating good ("G") and poor ("P") quality.

Procedure

During the simulated missions, participants directed the flight path of the drones by defining waypoints. Every drone was numbered and given a unique, identifying color to indicate both the drone itself and the path it was taking on the map. If the drone located a survivor in its FOV along the path, the survivors were marked on the map with a small pin showing their location. If those survivors were in critical need of resources (that is, if they did not receive supplies within 2 minutes of being found they would die) their location would be indicated by a small star. The pin or star color denoted what type of resource the survivors needed. Blue indicated water, yellow food, and green medical supplies. The hubs to fetch these resources were represented by small rectangles, color coordinated to match the supply type. They were all located at the southernmost end of the map. Each drone only had capacity for one supply kit at a time, but there were no restrictions on which drones could carry which supplies.

Before starting the missions, participants completed a series of surveys and watched a short video outlining the missions. They also received two 6-minute practice sessions. During the practice sessions, they experienced both high and low workloads, and always saw good initial recommendations. They then did the four real missions with pre-determined sequences. Participants received a 5-minute break between missions 2 and 3. After each mission, participants completed two self-reporting measurements describing their experience and perception of the system. They also completed a summary regarding their overall experience after mission 4. Each experiment took roughly 2 hours to complete. Performance, trust ratings, and recommendation response data from two participants were not properly saved for Mission 3.

Participants

A total of 31 adults were recruited for the experiment. However three subjects were omitted due to their failure to comply with some of the task guidelines. For example, one participant refused to update their trust level and two others did not reliably respond to the trust measurement prompt. Six other subjects were also omitted due to their range of reported trust values (which is elaborated on more in a later section). Thus, 22 subjects were included with reliable trust measures for further analysis. The participants were asked to self-report their previous education, experiences with UAVs, military experience, and how much time they spend playing video games per week. 39.3% of participants reported a history of military service. Ten of the 22 included participants served in either the Air Force or Navy, and one of the participants had served in the army. Thirteen of the included participants had some form of aviation experience, with 5 of those 13 participants having experience as either a pilot or sensor operator for UAVs. Five of the included participants had reported having experience with either C2 (Command and Control) or C3 (Command, Control, and Communication) systems. 71.4% of participants reported that they had spent at least 1 hour of playing video games in an average week.

MATERIALS AND MEASUREMENTS

Measures of Performance and Effectiveness

The objective measurements of participant performance were based on the number of survivors located, the number of supply kits delivered, and the number of survivors who died without receiving supplies in time. Other measurements to better understand participant engagement included the number of times participants interacted with each drone. Task effectiveness was assessed by taking the percentage of recommendations that participants selected against the number of recommendations the system provided. This metric also contributed to measuring participants' overall reliance on a system. Lastly, the proportion of the number of good versus poor recommendations selected by a participant was measured. An analysis of performance of subjects is given in Hertzler et al., 2023 showing that performance significantly decreases as cognitive load increases. There was no effect on mission performance when considering if a subject accepted or rejected the recommendation.

Trust Measure

For a measure of trust, the Continuous Online Numerical Score (CONS) was developed and validated as a self-reporting trust measurement (Hertzler et al., 2023). It examines factors captured with behavioral measurements according to common trust models (Lee & See, 2004; Mayer et al., 1995; Hoff & Bashir, 2015; Kohn et al., 2021). Participants continuously reported their trust throughout the experiment on a 5-point Likert scale. The trust measurement survey was modeled after Lee and Moray (1994) to create the CONS measurement and administered approximately every 30 seconds up to one minute. A Likert response of "1" meant the lowest level of trust, while "5" meant complete trust in the system. This methodology allowed participants to self-report their trust without substantially disrupting their performance. Due to the nature of this measurement methodology, a trust measurement can be assumed valid for any successive behavior following it (i.e., either accepting or rejecting recommendations). A behavioral measure of trust was also used to validate the CONS measure. When the participant received a recommendation from the system, they could choose one of three options: compliance (accepting a recommendation without reading, and indicative of a high trust state), verification (reviewing the quality of a recommendation before accepting is, which suggests moderately low trust), or rejection (denying the recommendations and indicative of very low trust) (see: Ezer et al., 2008; Lee & Moray, 1994; Moray et al., 2000). Hertzler et al. (2023) found that the CONS measure significantly overlapped with the participant's acceptance or rejection of the recommendation, validating the measure.

Simulation Device and Environment

The simulated mission environment utilized the Vortex interface (developed and customized through a partnership with Perceptronics Solutions Inc.), which is a modular autonomy integrated framework. Originally demonstrated with the U.S. Army's Advanced Teaming Demonstration Program (A-Team), it is designed to moderate levels of autonomy and help optimize HMT performances by assessing in real-time an operator's workload and trust.

Physiologic Sensors

To track eye movement and measure pupillary response, the FX3 Remote Eye Tracker (Eve Tracking LLC) was used. This device, in conjunction with the EyeWorksCognitive Workload Module, was used to calculate the Index of Cognitive Activity (ICA). The sensor was directly below the primary display, approximately 36 inches away from the participant.

DATA OVERVIEW

The work of Hartzler et al. (2023) describes various correlations and performances versus cognitive load and trust. In this work, the focus was aimed more on the aspects of gaze that can be used to infer trust levels. As such, an overview of trust and correlations different from the previous work are presented in the following sections. Hartzler et al. (2023) established a relationship between users' trust on a five point scale and accepting or rejecting system recommendations. To facilitate groupings of trust and expand the amount of gaze data, this research opts to reduce the levels of trust to a binary value (low and high). The threshold for low and high is calculated for each subject, using their median trust score. This ensures that subject specific variability or bias in trust scoring is mitigated, similar to the methods of Wilson et al. (2021). The distribution of trust values collected in the experiment for each subject is shown in Figure 2. Six participants without at least two reported measures of trust were removed from analysis.

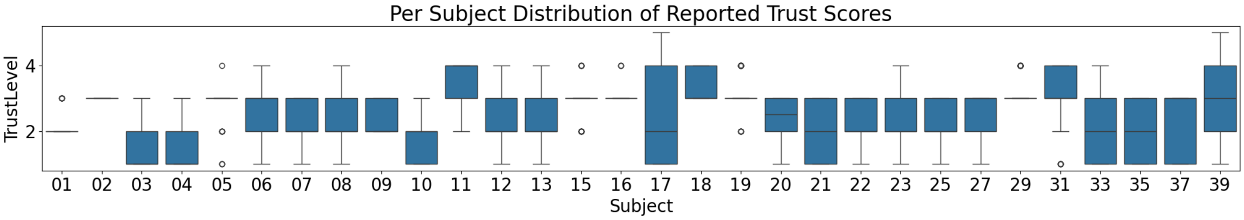

Figure 2. The distribution of reported CONS for each subject in the dataset. Subjects without enough variability in their reported scores are removed from further analysis.

In these analyses a significant correlation of trust levels with the index of cognitive activity ($p>0.5$) was not found. However, the inclusion of ICA as a feature for trust could still be warranted when analyzed in conjunction with the gaze features. Therefore, the inclusion of ICA in the models was investigated in the results section but only as a secondary variable to the gaze feature data.

The gaze data from the 30 seconds prior to a subject reporting trust is aggregated, and a Gaussian smoothing is applied to the gaze maps as is customary before visualizing (Wilson et al., 2020). This is achieved using the two-dimensional kernel density estimation in the scientific python library (SciPy). Figure 3 below shows two example gaze heatmaps for a low and high trust reported response from a subject. Darker values represent more concentrated fixations within a given spatial area.

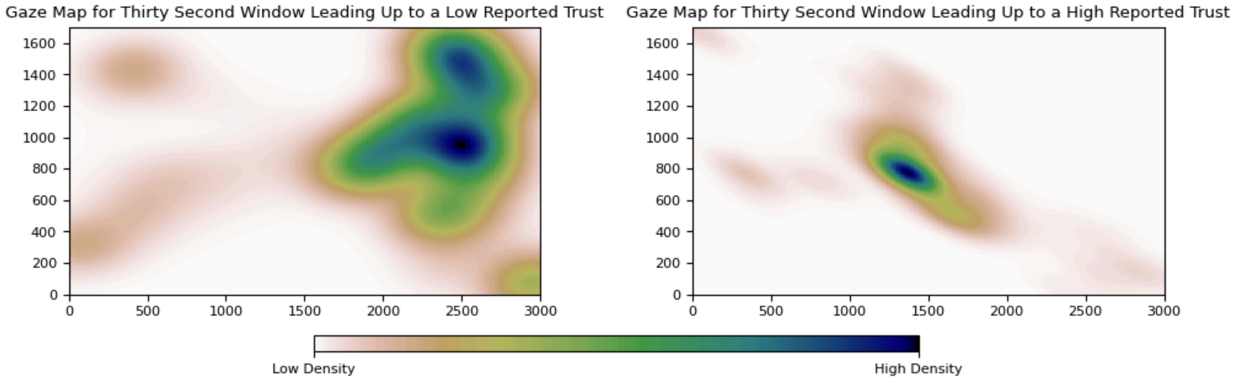

Figure 3. Example gaze heatmaps for a low and hight trust reported response

RESULTS AND DISCUSSION

Gaze Components and CONS Trust Reporting

The first focus is investigating the first two hypotheses that gaze fixation, distribution, and transition significantly influence the CONS self-reported trust measure. The binary labels of low and high trust are used to group the gaze data across participants and the smoothed gaze heatmaps as described previously. For each group of high and low trust gaze heatmaps, principal components analysis (PCA) is performed to understand the major gaze elements that contribute to variation. To understand the number of components in each group that significantly influence variability, a scree plot is displayed for each group in Figure 4. The x-axis shows the index for each component in the PCA ordered by eigen-value and the y-axis shows the relative contribution for each component (i.e., the explained variance ratio). From these scree plots, it is visually apparent that fewer components are needed to capture variability in the high-trust scenarios. This supports the previously stated hypothesis that high trust scenarios exhibit more consistency in gaze patterns because the operator is less concerned with checking the machine quality and can focus on more specific tasks. The lower trust scenarios require additional components because the gaze patterns are more variable and distributed. This is evidence that lower trust scenarios are marked by gaze patterns that verify the system information and therefore fixations and transitions are less consistent as the operator shifts attention across a wider array of tasks.

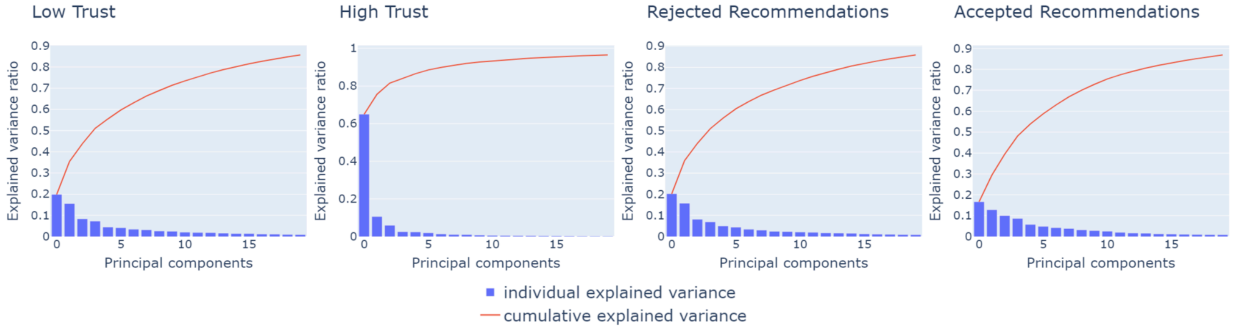

Figure 4: Scree plots of principal components of various groupings of fixation data from the collected dataset.

To further elucidate the scenarios, the top four components for low trust and high trust scenarios are visualized in Figure 5. Here each component is reshaped and visualized so that it can be interpreted as an element that makes up a gaze heatmap. In this way the leftmost component in Figure 5 is the most common component visible in each trust category. The top row shows components from the low trust responses and the bottom row shows components from the high trust responses. From left to right, the components become less influential. Since only the top four components are shown, all components can be considered relatively influential to the reconstruction of the gaze heatmap. A Welch's t-test is used to determine if the mean values in each component are statistically different across groups. The test finds that components one, two, and four in the low trust group are statistically different from the high trust group (p<0.001). However, component three cannot reject the null hypothesis (p=0.44). This can also be confirmed visually. Firstly, it can be observed that there are strikingly different component structures in the high and low trust groupings for the first, second, and fourth components, validating the findings from the scree plots. Even so, the third components have considerable visual similarity, which provides evidence that the gaze maps are not entirely different. Second, the components for the low trust responses are more dispersed, as shown by the relatively larger and separated clusters. This also validates the previously stated second hypothesis that high trust responses result in more focused attention to user interface elements for interpreting outputs, rather than verifying system actions. Finally, the high trust components are concentrated in the upper portion of the screen, validating the previously stated first hypothesis regarding higher fixation duration on specific UI elements for high trust scenarios.

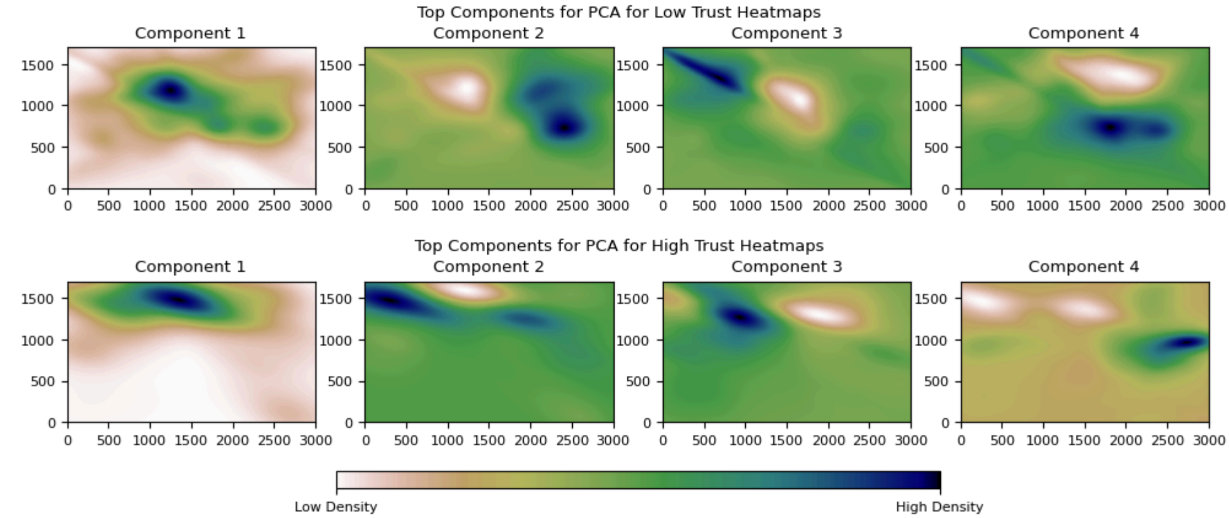

Figure 5. Principal components for low trust (top row) and high trust (bottom row), interpreted as gaze heatmaps. Darker regions correspond to more dense clusters of fixations.

Gaze Components and Recommendation Acceptance

The above methodology was repeated, but responses were grouped by accepted or rejected recommendations. Hartzler et al. (2023) showed considerable overlap in the CONS subjective reporting and recommendations behavior, which is investigated further here for similarity in gaze components. Figure 6 shows the top components for the rejected (top row) and accepted (bottom row). A Welch's t-test confirmed that all components in the rejected recommendations are statistically different from the accepted ones (*p*<0.01). The results show that the eye gaze components don't fully match with high and low trust scenarios as hypothesized. Accepted components mirror the high trust components in the 1st and 4th components but differ otherwise. Accepted recommendations still have smaller clusters and fixations (as do high trust scenarios). In general, this supports the hypotheses about fixation durations and gaze distributions. However, because high trust and acceptance components do not match entirely, gaze maps alone cannot be used for classifying trust without intelligent pre-processing. That is, the duration of fixation and consistency of clusters within the gaze map are more critical than where the subject looks. Larger, scattered gaze distributions contribute to rejections and low trust, though spatial dimensions of this are not well-defined. Thus, a model for classifying trust might be effective from features based on gaze maps kurtosis (i.e., more peaked spatial data), in contrast to a raw gaze analyzer like a convolutional neural network (CNN).

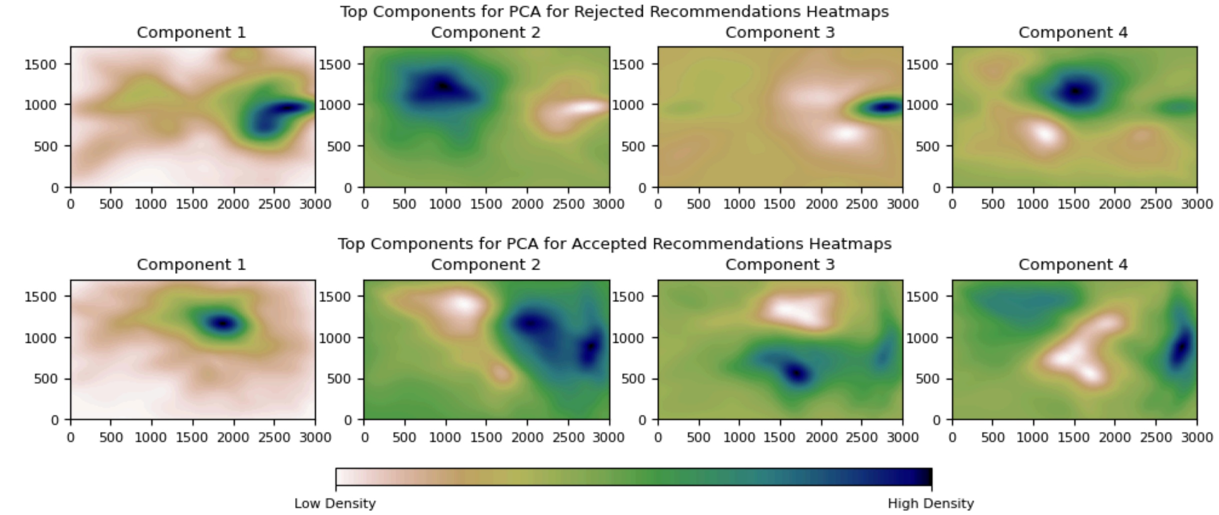

Figure 6. Principal components for rejected recommendations (top row) and accepted recommendations (bottom row), interpreted as gaze heatmaps.

Automated Classification of Trust from Gaze Heatmaps

Given the analysis regarding gaze, it has been established that trust should be discernible based upon the gaze maps, but that there is some overlap such that the classification of trust may not be perfect. To investigate if an automated machine learning model can discern trust, several machine learning models are trained including Logistic Regression, Random Forests, and k-Nearest Neighbor (KNN). Each model is chosen for the type of decision boundary that can be learned. Logistic regression can learn linear divisions of the features space, while random forests and KNN classifiers can learn more arbitrary decision boundaries between classes (Breiman, 2001).

For each classifier a new PCA analysis of all gaze maps combined is used. That is, all gaze maps are aggregated and PCA is performed to learn the top 20 components. The projection of each gaze heatmap onto these components is then used as a feature dimensionality reduction. These 20 features are then used to train each of the machine learning models. To train the models, per subject cross validation (PS-CV) is used. With PS-CV one subject is removed from the data and a model is trained on 80% of that subject's data and tested on the remaining 20%. This process is repeated for every subject in the dataset such that every subject's gaze and trust data are used to create a separate model. A "cross subject" model was also investigated, but ultimately did not perform well. This means that, in the current modeling, trust can only be captured when training data from a specific individual is available for calibrating the model. The results of this evaluation are shown in Table 2 below. The overall accuracy of the classifiers, along with the recall, precision, and F1-score is reported below. These last three measures focus more on the tradeoff between false negatives and false positives with recall being more sensitive to false negatives, precision being more sensitive to false positives, and F1-score being sensitive to both equally. For these evaluation metrics, "low trust" is considered the positive class as this is typically the classification of the most interest.

We also investigated the inclusion of pupillary-based measures such as the previously described ICA for a subject. To investigate this, machine learning models are trained with and without the mean ICA leading up to their measure of trust. These results are reported in Table 2, denoted by "with ICA" appended to the model name.

Finally, this paper investigated the use of statistical aggregations of each gaze map as features, replacing the PCA projection weights as features. In this way, the classifier uses the mean, standard deviation, skewness and kurtosis of the heatmaps as features to determine high or low trust CONS scores. These statistical aggregations are chosen because they are related to how clustered or variable the gazemaps are, without being specific to any specific locations of the gaze within the maps. From this perspective, these features might be more amenable to different user interfaces where the UI element locations are unknown beforehand. These models are also shown in Table 2, denoted by "with stats" appended to the model name.

Table 2: Summary of Per Subject Trust Classifier Metrics on Testing Set Data

Model	Accuracy	Recall	Precision	F1-Score	Support
Logistic Regression	**0.76**	**1.00**	0.76	**0.87**	1037
Logistic Regression with ICA	**0.76**	**1.00**	0.76	**0.87**	1037
Random Forest	0.74	0.97	0.76	0.85	1037
Random Forest with ICA	0.74	0.97	0.76	0.85	1037
K-Nearest Neighbor	0.70	0.89	0.76	0.82	1037
K-Nearest Neighbor with ICA	0.72	0.89	**0.77**	0.83	1037
Logistic Regression with Stats	**0.76**	**1.00**	0.76	**0.87**	1037
Random Forest with Stats	0.73	0.92	**0.77**	0.84	1037
K-Nearest Neighbor with Stats	0.72	0.92	0.76	0.83	1037

Based on the results in Table 2, it is apparent that using gaze for trust identification is reliable with high recall and good precision. The best performing model was Logistic Regression with any feature set. Thus, trust classification is viable from gaze data. The addition of pupillary-based measures appeared to either yield the same results, or slightly improve model metrics (as seen with K-Nearest Neighbor). The overall high performing model evaluations partially validates the previously stated third hypothesis that subjective trust can be predicted from biometric gaze data. However, the final result requires user-specific calibration–it is unclear if a model can be created that generalizes across users. In this way, the third hypothesis is not fully validated, and the generalization remains a topic for future research.

CONCLUSION

We presented an analysis regarding how levels of trust relate to eye gaze in a set of four search and rescue operations with UAVs. These results show that there are distinctive eye gaze attributes associated with trust level, with low trust levels having a more varied distribution with fewer fixations and more transitions between fixations. On the contrary, high trust was associated with more consistent fixations and less varied gaze distributions. This paper also presents a preliminary analysis for classifying trust using machine learning models that interpret the gaze heatmaps. These models resulted in the best accuracy of 76% and F1-score of 87% from a model that employed both gaze and ICA features. Statistical aggregations of the heatmaps also had competitive performance.

One limitation of the study is that gaze data to analyze trust for a given subject was the only factor used. Additional biometrics beyond gaze heatmaps were not investigated in this paper. An interesting avenue for future work would be the use of biometrics that can be sensed from on the body such as electro-dermal activity and brain-computer interfaces. These features may prove to be even more resilient to cognitive processes related to trust or may prove to generalize to a more varied set of mission scenarios than gaze alone. Moreover, these experiments only investigate trust in a set of four related missions on a similar interface. Thus, it's unclear how these automated measures of trust generalize to different user interfaces and different missions. Finally, the current modeling strategy for classifying trust is only reliable when subject-specific training data is available. Thus, some form of calibration is needed to create a model of trust for each individual. Generalizing this model beyond each subject remains an open research topic and likely will require additional features beyond gaze.

REFERENCES

Breiman, Leo (2001). Random forests. *Machine learning* (vol. 45, pp. 5-32).

Choo, S., and C. S. Nam (2022). Detecting human trust calibration in automation: a convolutional neural network approach. I*EEE Transactions on Human-Machine Systems* (Vol 52, no. 4, pp.774-783).

Crothers, N., Y. Sinha, E. C. Larson, and S. Scielzo (2022). Real-Time Situation Awareness Assessment for Pilots via Machine Learning: Constructing an Automated Classification System. *MODSIM World* (no. 14 pp. 1-11).

Dizaji, L. G., and Y. Hu (2021). Building and measuring trust in human-machine systems. In *2021 IEEE International Conference on Autonomous Systems* (ICAS), (pp. 1-5).

Ezer, N., Fisk, A. D., & Rogers, W. A. (2008). Age-related differences in reliance behavior attributable to costs within a human-decision aid system. *Human Factors* 50(6), 853–863. doi: 10.1518/001872008X375018

Gebru, B., L. Zeleke, D. Blankson, M. Nabil, S. Nateghi, A. Homaifar, and E. Tunstel (2022). A review on human–machine trust evaluation: Human-centric and machine-centric perspectives. IE*EE Transactions on Human-Machine Systems* (Vol. 52, no. 5 pp. 952-962).

Hartzler, B. M., S. Scielzo, A. Abraham, R. Wong & S. Kohn (2023). Effects of Trust Calibration on Human-Machine Team Performance in Operational Environments. In *The Interservice/Industry Training, Simulation and Education Conference* (I/ITSEC), Orlando, FL.

Hoff, K. A., & Bashir, M. (2015, September). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors*, (Vol 57 (3), pp. 407-434).

Hu, W., K. Akash, N. Jain, and T. Reid (2016). Real-time sensing of trust in human-machine interactions. In the In*ternational Federation of Automatic Control* (IFAC) (Vol 49, no. 32, pp. 48-53).

Kaplan, A. D., T. T. Kessler, and P. A. Hancock (2020). How Trust is Defined and its use in Human-Human and Human-Machine Interaction. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, (vol. 64, no. 1, pp. 1150-1154). Sage CA: Los Angeles, CA: SAGE Publications, 2020.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, (vol. 40 no (1), pp. 153-184). doi: 10.1006/ijhc.1994.1007

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors* (Vol. 46(1), pp. 50–80). doi: 10.1518/hfes.46.1.50.30392

Lyons, J. B., N. T. Ho, L. C. Hoffmann, G. G. Sadler, A. L. Van Abel, and M. Wilkins (2018, July). Trust in sensing technologies and human wingmen: Analogies for human-machine teams. In *Augmented Cognition: Intelligent Technologies: 12th International Conference*, AC 2018, Held as Part of HCI International, Las Vegas, NV, USA, (Proceedings, Part I, pp. 148-157). Springer International Publishing.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, (vol. 20(3), pp. 709–734). doi: 10.5465/amr.1995.9508080335

Moray, N., Inagaki, T., & Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of Experimental Psychology – Applied*, (Vol. 6, pp. 44–58).

OP TEMPO: Objective Prediction of Team Effectiveness via Models of Performance Outcomes (OP TEMPO) (2023, November). DARPA program HR001124S0006.

Scielzo, S., & Kocak, D. (2021). A Multi-Domain Robotic Teammate Framework: Next Generation Human-Machine Interface Guidelines to Support Trust and Mission Outcomes. *The Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*, Orlando, FL, 2021

Scielzo, S., Wilson, J. C., & Larson, E. C. (2020, November). Towards the development of an automated, real-time, objective measure of situation awareness for pilots. In *The Interservice/Industry Training, Simulation and Education Conference* (I/ITSEC), Orlando, FL.

Visser, de E. J., Beatty, P. J., Estepp, J. R., Kohn, S., Abubshait, A., Fedota, J. R., & McDonald, C. G. (2018). Learning from the slips of others: Neural correlates of trust in automated agents. *Frontiers in Human Neuroscience*, (vol. 12, p. 309).

Wilson, J., Scielzo, S., Nair, S. and Larson, E.C., (2020). Automatic gaze classification for aviators: Using multi-task convolutional networks as a proxy for flight instructor observation. *International Journal of Aviation, Aeronautics, and Aerospace*, (vol 7(3), p. 7).

Wilson, J. C., Nair, S., Scielzo, S., & Larson, E. C. (2021). Objective measures of cognitive load using deep multi-modal learning: A use-case in aviation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, (Vol. 5 no (1), pp. 1-35).