Northeast Journal of Complex Systems (NEJCS)

Volume 7
Number 2 *Special Issue: NERCCS 2025 Papers*

Article 4

June 2025

Classification of human trust in AI using brain activity data

Danushka Bandara
Fairfield University, danushkabandara@gmail.com

Ruhuan Liao
Fairfield University, ruhuan.liao@student.fairfield.edu

Fatima Chowdhury
Fairfield University, fatima.chowdhury@student.fairfield.edu

Leslie Abbott
Fairfield University, leslie.abbott@student.fairfield.edu

Follow this and additional works at: https://orb.binghamton.edu/nejcs

 Part of the Bioinformatics Commons, Cognitive Neuroscience Commons, and the Graphics and Human Computer Interfaces Commons

Recommended Citation

Bandara, Danushka; Liao, Ruhuan; Chowdhury, Fatima; and Abbott, Leslie (2025) "Classification of human trust in AI using brain activity data," *Northeast Journal of Complex Systems (NEJCS)*: Vol. 7 : No. 2 , Article 4.
DOI: https://doi.org/10.63562/2577-8439.1108
Available at: https://orb.binghamton.edu/nejcs/vol7/iss2/4

This Article is brought to you for free and open access by The Open Repository @ Binghamton (The ORB). It has been accepted for inclusion in Northeast Journal of Complex Systems (NEJCS) by an authorized editor of The Open Repository @ Binghamton (The ORB). For more information, please contact ORB@binghamton.edu.

Classification of human trust in AI using brain activity data

Danushka Bandara, Ruhuan Liao, Fatima Chowdhury and Leslie Abbott
Department of Computer Science and Engineering, Fairfield University, Fairfield, CT, 06825
* danushkabandara@gmail.com

Abstract

Trust plays a crucial role in human-computer interaction, particularly in scenarios involving artificial intelligence (AI) systems. This study explores the feasibility of using functional near-infrared spectroscopy (fNIRS) data to classify trust levels in human-AI interaction scenarios. A total of 18 participants completed an image classification task with an AI team member while their hemodynamic responses were recorded using fNIRS. Preprocessing of fNIRS data involved motion artifact removal, filtering, and normalization. Exploratory analysis identified significant associations between hemodynamic responses in the prefrontal cortex and trust levels. An across-subject binary trust classification model was developed using machine learning techniques, achieving an F1 score of 0.77. Receiver Operating Characteristic (ROC) analysis revealed an Area Under the Curve (AUC) of 0.81, achieving improved F1-score and AUC compared to comparable methods. The brain activity-based classifiers were found be be better at classifying the self-report trust level than the objective measure of trust. These findings demonstrate the potential of fNIRS-based approaches for real-time classification of trust levels in human-AI interaction, with implications for improving user experience and trustworthiness of AI systems.

1 Introduction

Trust in human-AI interaction is a crucial factor that significantly impacts the success of collaborations between humans and artificial intelligence systems. Research has demonstrated that trust plays a pivotal role in human-technology interactions [25]. According to literature, there are two levels when it comes to human trust: dispositional trust does not change over the adult lifespan, however, situational trust fluctuates with experience[76]. Understanding the temporal evolution of trust in AI-supported decision-making can offer deeper insights into how trust evolves and

changes over time [26, 27, 28]. However, measuring trust in real world settings is difficult because of two reasons, 1) There is no relaible objective measure of trust. 2) Self reported trust is usually not at a sufficiently fine-grained temporal resolution to analyze the dynamics of trust. Newer sensing methods such as wearable sensors and audio/video processing can offer a solution to this problem [77]. In particular, Brain activity is a valuable window into human cognitive states. There have been numerous studies using Electroencephalography(EEG) sensors to measure trust in human-computer interactions[29, 30, 33, 34, 36] and in particular, human AI interactions [33, 32, 31]. Functional Magnetic Resonance imaging is also widely used to measure trust in controlled settings [35, 37, 38, 39]. However, the high cost and the restrictive nature of the fMRI device make it less preferable to EEG. EEG also has drawbacks such as high susceptibility to noise and low spatial resolution.

Functional near-infrared spectroscopy (fNIRS) is a valuable tool for measuring brain activity non-invasively. It has been extensively used in various fields such as neuro ergonomics, social cognition, and human-computer interaction [1, 2, 3, 4]. fNIRS has better spatial resolution than EEG, better temporal resolution compared to fMRI, and lower cost compared to fMRI. fNIRS also offers greater mobility and comfort for participants compared to fMRI, making it suitable for capturing naturalistic behaviors relevant to trust. Even though fNIRS has lower temporal resolution than EEG, it could be suitable for measuring trust because trust-related cognitive processes unfold at a gradual rate [78]. Recently developed techniques such as short-channel regression [40] have improved the signal-to-noise ratio in fNIRS data, making it increasingly reliable for studying cognitive processes like trust. fNIRS works by measuring changes in oxygenated and deoxygenated hemoglobin levels in the brain, providing insights into neuronal activity [3]. The portability and versatility of fNIRS make it a preferred choice for studying brain function in various naturalistic environments [7]. Researchers have successfully used fNIRS to study trust in different contexts. fNIRS has recently been used to measure trust in automated driving scenarios [5]. Eloy et. al [6] employed multimodal data collection (fNIRS, GSR, eye-tracking, detailed surveys) and statistical analyses to examine trust in a manipulated, team-based context. Their exploration of trust occurred in a realistic human-agent teaming scenario, whereas this study aims to measure trust dynamics in a controlled environment that has fewer confounding factors. This research aims to classify human trust in a human-AI team scenario using brain activity data collected during an image classification task. the aim is to understand how trust manifests in the human brain and to develop a machine learning-based model to classify trust vs mistrust conditions.

https://orb.binghamton.edu/nejcs/vol7/iss2/4
DOI: 10.63562/2577-8439.1108

2 *Related work*

Trust is a fundamental aspect of huaman nature that significantly influences human interactions and decision-making processes. Research has shown that trust plays a crucial role in group decision-making [42, 43, 46, 45]. Trust is underscored as a fundamental element in human-machine relationships as well, highlighting the importance of improving trust in AI systems [44, 47]. Despite the growing attention towards human-AI trust, a comprehensive understanding of trust-based relationships with AI is still developing [48]. Many studies stress the significance of trust in human-technology interactions, especially as AI agents assume tasks traditionally carried out by humans [25]. Trust is commonly utilized as a metric to assess these interactions [49]. The safety and effectiveness of human-AI collaborations rely on humans appropriately adjusting their trust towards AI agents [50]. Exploring trust as a dynamic variable in human-AI interactions can guide designs and enhance decision quality [26]. Moreover, trust is vital in human-AI relationships, with interaction trust playing a crucial role in managing these interactions [51]. With the increased integration of AI systems in day-to-day life, the optimization of trust dynamics becomes more and more vital [52, 53, 54, 55]. Another aspect to consider is that trust is multidimensional. Cognitive trust involves beliefs about the reliability, competence, and predictability of the trustee [57]. Affective trust, on the other hand, is rooted in emotions and feelings, reflecting the emotional bond and comfort felt towards the trustee [58]. Behavioral trust pertains to actions and behaviors that demonstrate trust, such as reliance on the trustee and willingness to be vulnerable in the relationship [59]. The cognitive, affective, and behavioral dimensions of trust have been widely explored by researchers. For example, in e-commerce settings, trust has cognitive, affective, and behavioral facets influencing consumer loyalty and purchase intentions [60]. Similarly, in team dynamics, trust development involves cognitive and affective dimensions that mediate the relationship between leader behavior and team performance [61]. Studies have also investigated the effects of affective and cognitive trust on user continuance intentions in mobile health services [62], highlighting the asymmetric effects of these dimensions [63]. Additionally, the foundational role of cognitive trust in establishing affective trust has been established by researchers [64]. Measuring trust accurately, especially in complex social contexts, presents significant challenges due to the multidimensional nature of trust and the intricacies involved in capturing its essence. The difficulty in measuring trust arises from various factors, as highlighted in the literature. One critical challenge is differentiating between predictors of trust and actual measures of trust [65]. Moreover, the temporal aspect of trust measurement adds another layer of complexity. Most research on trust has traditionally taken a static view, capturing trust at a single point in time, which may not

fully capture the dynamic nature of trust development and fluctuations over time [66]. Furthermore, the social and relational dimensions of trust in specific contexts, such as human-AI interactions, pose unique measurement challenges [67].

2.1 Definition of trust

We consider the following definition of trust in this paper,

- "The extent to which a user is confident in, and willing to act on the basis of, the recommendations, actions, and decisions of an artificially intelligent decision aid"[9, 10]

Based on the above definition of trust, we arrived at two types of trust. These two aspects are also supported by the work of Ferrario et al. [48].

- Simple trust: reliance on AI to achieve a goal without consideration of AI's capabilities to achieve said goal.
- Reflective trust: AI has properties that are objective reasons for the human to simply trust the AI to achieve the goal.

2.2 Empirical measurement of trust

Many studies measure trust using a questionnaire before the interaction with a system. This initial trust is used as a baseline measure. Following the interaction it assesses participants' final trust level in the system, influenced by the most recent interaction. However, these methods fail to account for shifts in participants' trust levels toward the system. Trust is dynamic, it can be increased, decreased, repaired, and maintained [11]. If we view an experiment as a series of repetitive occurrences, one such occurrence is termed a trial. In our experiment, a trial typically involves participants making decisions in response to an AI recommendation. Surveys may be administered after each individual trial or after each group of trials (block). While this method may offer a more detailed understanding of trust dynamics, such as identifying any sudden spikes in trust levels and pinpointing the specific trial responsible for such fluctuations, it also extends the duration of the experiment and/or necessitates shorter questionnaires. Another choice researchers have to make is the type of questions posed to participants. Trust can be measured on the Likert scale (e.g. from 1 to 5) or as a binary trust level. We chose a binary trust question because it reduces participants' subjectivity in rating their trust level.

2.3 Experiment design for measuring trust

There are many experimental designs used by researchers in the trust domain. Many studies have used a crowd-sourcing approach [17]. Other approaches include in-person studies using various platforms such as Virtual Reality or Mixed Reality, standard psychology testbeds, and custom testbeds. The challenge in in-person studies is the inability to get a large number of participants due to time and resource constraints. In this study, we propose an across-subject analysis that enables researchers to use a larger pool of participants to train the model.

2.4 Physiological measures of trust

Physiological indicators are essential in evaluating trust across various contexts, such as human-robot interactions, automated vehicles, and collaborative settings. These indicators offer valuable insights into individuals' emotional states, cognitive processes, and behavioral responses related to trust. Common physiological indicators used to assess trust include heart rate, skin conductance, facial expressions, and eye movements. Heart rate variability is a well-researched physiological indicator that reflects emotional arousal and cognitive load, making it a valuable measure of trust in human-machine interactions [68]. Changes in heart rate can signal shifts in emotional states and stress levels, providing insights into individuals' trust levels in automated systems [69]. Additionally, heart rate variability has been associated with group cohesion and team trust, underscoring its importance in interpersonal coordination and collaborative environments [70]. Skin conductance responses can indicate alterations in emotional arousal and engagement, offering valuable information about individuals' trust and comfort levels in different situations [71]. Facial expressions and eye movements serve as non-verbal cues that can act as indicators of trust in human-machine interactions [71, 72, 73]. Eye movements, including gaze behavior and pupil mimicry, have been linked to trust and social cognition, offering valuable cues about individuals' trust in automated vehicles and agents [74].

2.5 Brain activity and trust

Brain activity provides an objective method of understanding human psychological state. Thus, brain activity measures have been explored by researchers in the trust measurement paradigm. Wang et al. [12] identified neural correlates of trust in human-autonomy interaction using EEG signals, with significant correlations found in the frontal and occipital areas. Another study by Filkowski et al. [13] investigated the use of a control-of-attitudes fMRI task, which involved explicit instructions to control attitudes of interpersonal trust and distrust. They found that TPJ, mPFC,

insula, and inferior and lateral frontal cortices were associated with trust/mistrust conditions. Firoz et al. [23] obtained 72% accuracy for binary trust classification using an EEG sensor. Akash et al. used EEG in combination with Galvanic Skin Response (GSR) sensors to classify trust level [15] and achieved a 78.55% accuracy level for a binary classification. fNIRS has garnered increasing interest in the area of trust measurment [14, 16]. Noah et al. [75] utilized fNIRS adapted for hyperscanning to investigate neural mechanisms associated with eye-to-eye gaze during social interactions. The study found that neural coherence between participants was greater during reciprocal eye-to-eye contact compared to direct eye-gaze at a dynamic video, particularly in social and face processing systems. The use of physiological response to monitor and understand trust is currently limited due to a lack of knowledge on physiological indicators of trust. This study examines neural responses to trust within a human-AI collaboration task. Researchers have found evidence that increased activation of medial and dorsolateral prefrontal cortex was detected in response to mistrust caused by disruptions in human robot teams [79]. Literature also shows the involvement of the Ventromedial Prefrontal cortex in decision making scenarios [80, 81]. This study extends such work by studying the prefrontal cortex neural activation with relation to trust in a human-AI team scenario.

3 Methods

3.1 Participants

The participants in this study (n=18) were aged 18-22 from a university in the northeast United States (nine male, nine female). Only one participant was left-handed. The participants' experience with computers ranged from two to five (mean=2.58, st. dev.= 0.69) on a scale from one to five. Their propensity to trust [18] ranged from 1.67 to 2.91 (mean=2.58, st. dev.= 0.69) on a scale from one to five.

3.2 Experiment design

We adopted the disposition to trust inventory [19] and adjusted the questions to reflect human-AI trust. When participants arrived for the experiment, they first signed the consent form and were given instructions on how to complete the tasks. They were then seated in front of the computer and completed a demo of the task to remove any novelty effects. The fNIRS sensor was then placed on the participant's scalp and adjusted so that it was centered on the head and the probes were orthogonal to the scalp surface. Then the fNIRS was calibrated and recording started. The experiment presented various images consecutively with 10 seconds of rest

between each trial. The participants' job was to judge whether the images were authentic or 'doctored' (manipulated in any way). We used the CASIA [20] dataset as the source for the images. In addition to the images, participants were shown the AI decision of whether the image was authentic or not authentic. We used a 'wizard of Oz' methodology for this study (i.e., we simulated the AI response) to show the participants a variety of AI responses. The AI decision was designed so that it provided 75% accuracy level (meaning it would provide the correct answer 75% of the time). This decision was made to simulate a realistic AI accuracy level to assess trust under uncertainty. Taking the group as a whole, we found that the average user accuracy was 60% when the users reported mistrust of the AI, and 67% when they reported trust of the AI (p=0.087). We also administered a trust measurement survey during the experiment to gauge if users accepted the AI decision or not. The collected data per each trial of the experiment are as follows,

- The AI system has determined that the given image is xxx (authentic/not-authentic). Do you accept or deny this claim? Response options: Radio button with choices Accept or Deny.
- Level of confidence in the answer. Response options: Slider ranging from 0 to 100.
- Do you trust or distrust the AI? Response options: Keyboard Left arrow for distrust, Right arrow for trust.

The above multimodal approach was used to obtain user responses to reduce survey fatigue [22]. The binary trust measures served as the primary target for statistical and machine learning analyses, while the confidence scores were collected for potential secondary investigations (e.g., correlating confidence with hemodynamic responses) but were not used in this study. Each subject completed 100 such trials (labeling decisions). Each trial obtained 5 responses. The user responses were tracked by the psychopy [21] software, which was also utilized to present the experiment materials; it tracked the time of presentation of each question as well. As indicated above, we combined fNIRS data with the user responses to evaluate fNIRS data as trust indicators.

3.3 Data collection

In this research, we utilized a NIRx NIRSport2 fNIRS device operating at a sampling rate of 10 Hz to capture brain activity data. The device was placed on the participant's forehead area to cover the prefrontal cortex region of the brain. The experimental setup is shown in figure 1.

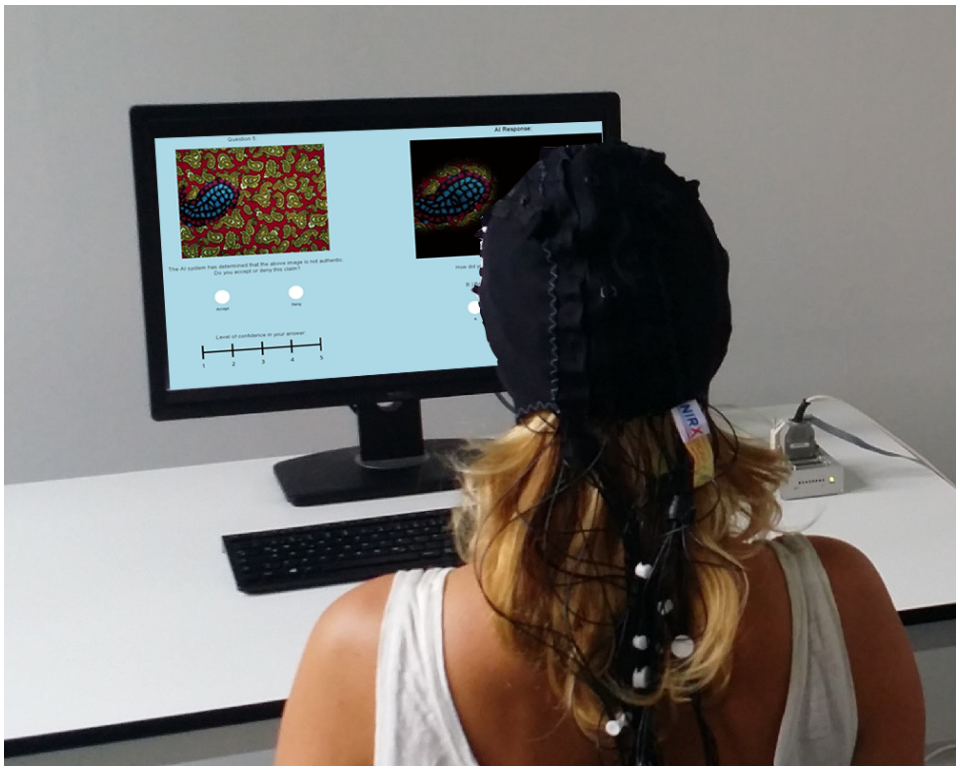

Figure 1: fNIRS experimental setup.

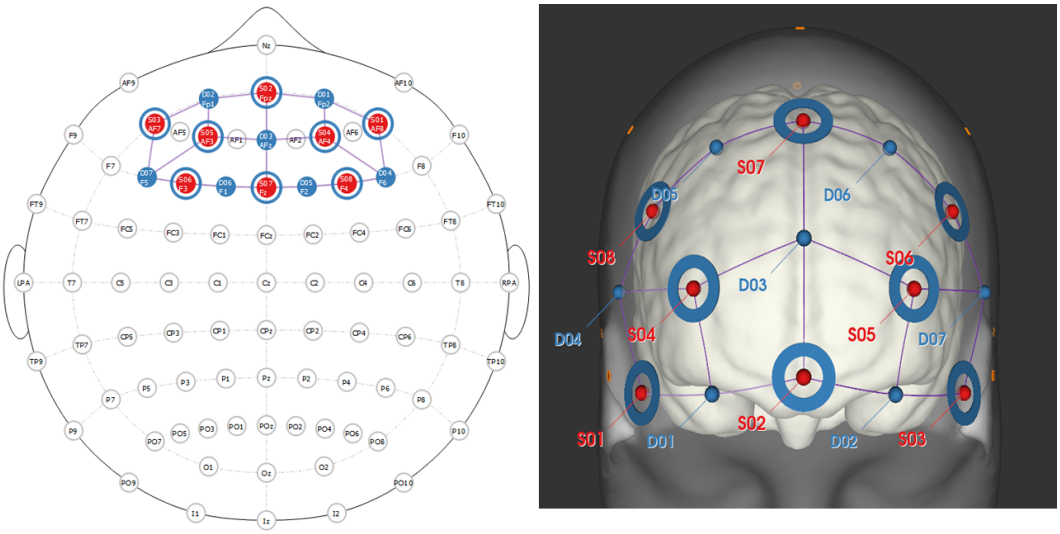

Figure 2: The left-hand side image shows fNIRS optodes and channel locations on a 10-10 coordinate system, The right-hand side image is the optodes and channels shown on the anterior view of the prefrontal cortex. The emitters are shown as red dots and the detectors as blue dots. The blue circles around the red dots are the short-distance channels used in short-channel regression. Each emitter detector pair constitutes a channel (Shown by the lines between the optodes)

3.4 Optode configuration

The fNIRS sensor consists of optodes (light emitters and detectors). These can be custom-configured on the head. The NIRsport 2 sensor we used was configured as shown in figure 2. The optodes were placed on the 10-10 coordinate system to cover the prefrontal cortex.

3.5 Format of the fNIRS data

The fNIRS light emitters (shown in red in figure 2) pulse near-infrared light at 760nm and 850nm wavelengths into the scalp. The reflected intensity of near-infrared light at different wavelengths is measured by the fNIRS detectors (shown in blue in figure 2). This data is collected over time and is stored as a matrix where each row represents a time point and each column represents a measurement channel or optode pair. The channel locations relative to the optode locations are shown in figure 2.

3.6 Data preprocessing

The raw light intensity signals were then processed to calculate the concentration changes of oxygenated hemoglobin ($\triangle$HbO) and deoxygenated hemoglobin ($\triangle$HbR) in the micro-vessels of the prefrontal cortex. Then we applied short-channel regression (SCR) [41] to the $\triangle$HbO and $\triangle$HbR data.

3.6.1 Short channel regression

SCR assumes that physiological noise, such as systemic fluctuations in heart rate and respiration, affects both the short-distance (i.e., superficial) and long-distance (i.e., deeper) channels of the fNIRS sensor array similarly. Therefore, SCR involves regressing out the short-distance channel signal from the signal of each long-distance channel to remove common physiological noise components. This results in a cleaner blood flow signal devoid of any systemic fluctuations due to respiration and other factors.

3.6.2 Feature extraction

After SCR, the $\triangle$HbO and $\triangle$HbR data were further filtered using a third-order zero-phase Butterworth filter (filter range between 0.01 Hz and 0.5 Hz). We had to remove subject IDs 5, 8, and 9 from the analysis due to errors in the data collection process of the fNIRS device. The remaining 15 subjects' fNIRS data was processed by extracting the relevant time series for each trial. We extracted five statistical features (mean, standard deviation, kurtosis [24], min, and max) values as feature values for each measurement channel for each trial. Let's define the features for a given trial:

- x_{ij} as the jth measurement channel's ith data point for the trial,
- μ_j as the mean of the jth measurement channel for the trial,
- σ_j as the standard deviation of the jth measurement channel for the trial,
- κ_j as the kurtosis of the jth measurement channel for the trial,
- $\min_j$ as the minimum value of the jth measurement channel,
- $\max_j$ as the maximum value of the jth measurement channel for the trial.

Then, the formal representation of the statistical features extracted for each measurement channel for the given trial can be represented as follows:

https://orb.binghamton.edu/nejcs/vol7/iss2/4
DOI: 10.63562/2577-8439.1108

1. Mean:

$$\mu_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}$$

2. Standard Deviation:

$$\sigma_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{ij}-\mu_j)^2}$$

3. Kurtosis:

$$\kappa_j = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_{ij}-\mu_j)^4}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_{ij}-\mu_j)^2\right)^2}$$

4. Minimum:

$$\min_j = \min(x_{ij})$$

5. Maximum:

$$\max_j = \max(x_{ij})$$

The combination of the above five statistical features times the 40 channels (20 $\triangle$HbO and 20 $\triangle$HbR) represents the feature vector for a given trial. The experiment had 15 subjects, 100 trials per subject, 40 channels, and 5 features per channel. The design matrix thus contains 15 X 100 = 1500 rows and 40 X 5 = 200 columns. The collection of these vectors along with the self-report (subjective) or objective trust labels creates the full dataset analyzed in the rest of the paper. (i.e. in the subjective vs objective analysis, the brain activity features remained the same, while the labels were changed)

3.6.3 Statistical and Machine Learning Analysis

In the statistical analysis, a two-tailed, paired t-test of unequal variance was conducted utilizing a comparative p-value of 0.05. This test aimed to determine whether there was a statistically significant difference in $\triangle$HbO and $\triangle$HbR between the accept and deny conditions (objective measure of trust). In other words, The null hypothesis was that there is no difference in mean $\triangle$HbO, $\triangle$HbR between trust and mistrust conditions, while the alternative hypothesis was that there is a difference. The statistical analysis was run on each of the statistical features extracted from the data (mean, standard deviation, kurtosis, min, and max).

In the machine learning analysis, several machine learning methods were tested on the resulting dataset to assess the binary trust classification performance. We conducted across-subject classification using leave one subject out cross-validation to test the generalizability of our model. The preprocessing pipeline is shown in

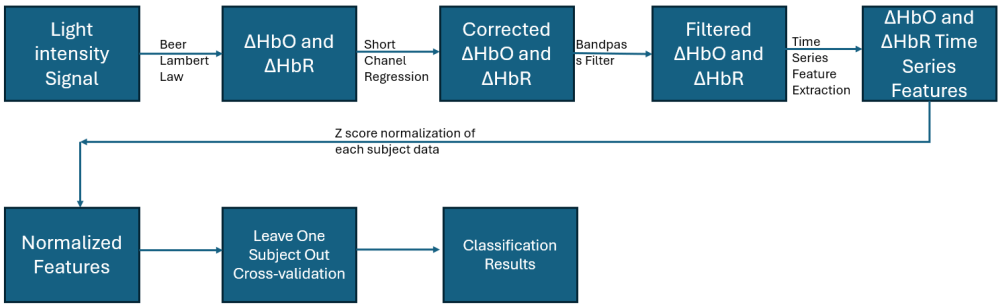

Figure 3: fNIRS data preprocessing and machine learning pipeline.)

figure 3.

We conducted an across-subject binary trust classification analysis to evaluate whether the fNIRS model can perform trust classification. Specifically, we aimed to answer the following questions:

- RQ1: How do hemodynamic responses in the prefrontal cortex correlate with trust levels in human-AI interaction scenarios?
- RQ2: Can machine learning be used to classify trust levels based on fNIRS data?
- RQ3: How do subjective perceptions of trust in AI systems compare with objective measurements when it comes to classification performance?

Exploratory analysis of fNIRS features revealed that hemodynamic responses in the prefrontal cortex were significantly associated with trust levels. Specifically, increased oxygenation levels in the prefrontal cortex were observed during trust conditions compared to mistrust conditions.

3.7 Objective (simple) trust level statistical analysis

Statistical analysis of accept/deny response data with null hypothesis being there is no significant difference in deoxygenated hemoglobin ($\triangle$HbR) levels within the S1-D1 channel between the accept and deny conditions and alternate hypothesis being there is a significant difference indicates a statistically significant difference in deoxygenated hemoglobin levels within the S1-D1 channel ($p<0.05$). This channel consistently displays significant variations across key statistical features, such as mean, standard deviation, kurtosis, minimum, and maximum values. As seen in

figure 4, the higher trust (accept) condition is associated with higher △HbR for all subjects.

3.8 Self-report (reflective) trust level statistical analysis

The t-test results with null hypothesis being There is no significant difference in deoxygenated hemoglobin (△HbR) and oxygenated hemoglobin (△HbO) levels between the trust and mistrust conditions in the S2-D1 and S5-D7 channels. And alternate hypothesis being there is a significant difference in △HbR and △HbO levels between the trust and mistrust conditions indicate that the △HbR values in the S2-D1 and S5-D7 channels consistently showed a statistically significant difference between the trust and mistrust conditions ($p<0.05$) across all the measured features. This indicates that these channels are connected with the trust levels. The △HbO values in the S5-D7 consistently showed a significant difference ($p<0.05$). Channels showing statistically significant differences between trust and mistrust conditions likely play a crucial role in decision-making concerning trust in AI. The box plots in figure 4 further illustrate this point. Figure 4 b shows a significant decrease in △HbO for the 'Deny response' condition ($p < 0.05$), while figure 4 c, d, e show significant increases in trust conditions ($p < 0.05$), reflecting regional variations. These statistical testing results indicate that the answer to our RQ1 is affirmative.

3.9 Objective (simple) trust level machine learning analysis

The objective labels (accept/deny) were used as the machine learning targets and the leave one subject out cross-validation results were obtained. We tested Logistic regression, Random Forest, and extreme gradient boosting models on this data. The best results were obtained by the extreme gradient boosting model. The average accuracy from the leave one subject out cross-validation for the extreme gradient boosting model was 57%. The average f1-score was 0.693 and the average AUC was 0.53.

3.10 Self-report (reflective) trust level machine learning analysis

We tested several machine-learning classification models, the results of which are shown in table 1. The extreme gradient boosting model achieved the best classification performance. Figure 5 shows the ROC curve for the binary classification of self-report trust level based on a 10 to 5 subject split of the full dataset. According to table 1, the trust classification model achieved best average accuracy of 76% (σ= 11%) and F1 score of 0.77 (σ= 13%) for across subject trust classification. Receiver Operating Characteristic (ROC) analysis revealed an average Area Under the Curve

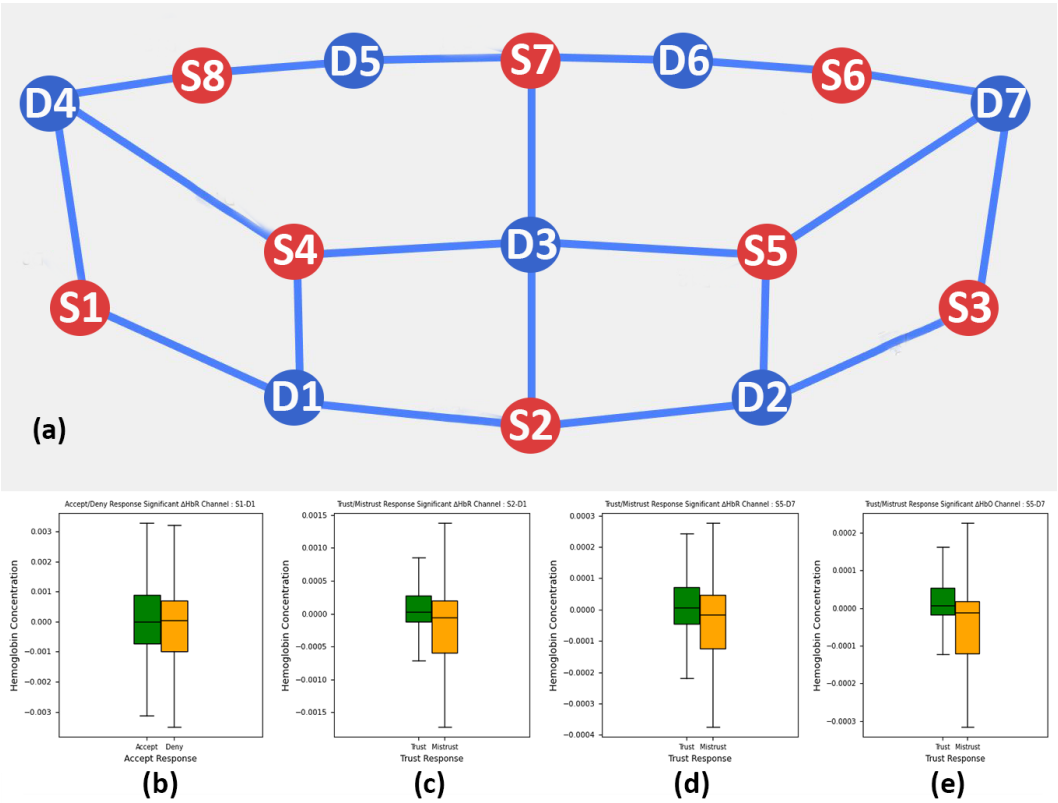

Figure 4: (a) Layout of the probe locations. The blue lines depict the 20 fNIRS channels. Each channel takes two measurements:△HbR and △HbO. These channel data were used in the statistical and machine learning analysis. (b) Box plot for statistically significant channel S1-D1 (△HbR) for the accept deny condition. (c) Box plot for statistically significant channel S2-D1 (△HbR) for the self-report trust condition, (d) Box plot for statistically significant channel S5-D7 (△HbR) for the accept deny condition., (E) Box plot for statistically significant channel S5-D7 (△HbO) for the accept deny condition.)

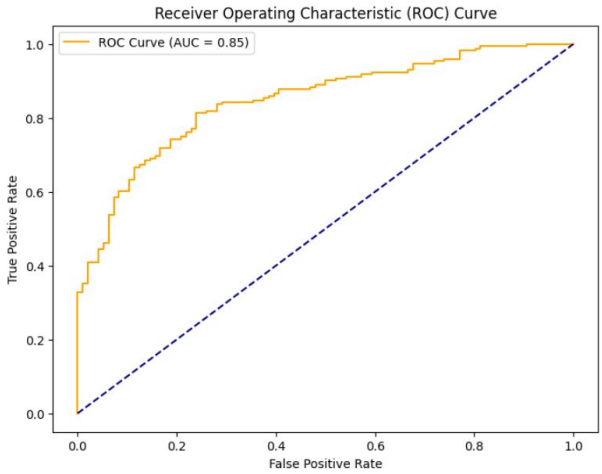

Figure 5: Binary trust classification ROC curve for self-report (reflective) trust classification.

	Subjective Classification			*Objective Classification*		
	Accuracy	*F1-score*	*AUC*	*Accuracy*	*F1-score*	*AUC*
Logistic Regression	69.3%	0.73	0.75	56.9%	0.61	0.53
Random Forest	74.8%	0.77	0.81	**59.7%**	**0.62**	0.51
XGBoost	**76.0%**	**0.77**	**0.81**	57%	0.60	**0.53**

Table 1: Classification results for subjective and objective trust levels in the across-subject setting.

(AUC) of 0.81 (σ= 0.11), indicating good performance of the classification model for this task. Thus the answer to our RQ2 is affirmative.

3.11 Self-report (reflective) trust level most predictive features

The rest of the analysis was done on our best-performing XGBoost model. We tallied the most predictive features for all the cross-validation splits. The features that appeared the most were the mean oxygenated and deoxygenated hemoglobin features shown below,

4 Discussion

With reference to RQ1: we found that increased oxygenation levels in the prefrontal cortex were observed during trust conditions compared to mistrust condi-

Table 2: Most predictive channels for the self-report trust classification

△HbO	**△HbR**
S1-D1	s2-D3
S1-D4	S8-D5
s2-D1	
S3-D2	
S3-D7	
S4-D3	
S4-D4	
S5-D2	
S5-D3	
S5-D7	
S6-D7	
S7-D5	

tions. Increased oxygenation levels in the prefrontal cortex during trust conditions may reflect greater engagement of cognitive and social processes, such as decision-making, emotional regulation, or theory of mind, which are known to activate this region. In contrast, mistrust conditions might involve less prefrontal activity if participants default to skepticism without extensive deliberation. However, as seen in figure 4, this was not consistent for every channel. The difference in this effect may reflect distinct hemodynamic responses in different prefrontal cortex regions. For example, "The drop in △HbO in Fig. 4b during the 'Deny response' condition may indicate reduced cognitive engagement in mistrust states, whereas the increases in figs.4 c,d,e during trust conditions suggest heightened activity related to decision-making or emotional regulation.

To address our RQ3, in the statistical analysis, we found that both simple and reflective trust labels were associated with statistically significant variations in △HbO and △HbR channels.

However, the machine learning models were much more adept at predicting the subjective self-report measure of trust than predicting the objective measure of trust. Here the objective measure is based on the acceptance or denial of AI decision, this could be confounded by other variables like fatigue. Whereas the self-report trust, albeit subjective, gets directly at the trust variable. The models for self-report trust achieved 0.77 F1 score and 0.81 AUC. Surpassing previously obtained results using a combination of EEG and GSR [15].

4.1 Limitations and future work

Our study was limited to a narrow demographic which is students from a university in the northeast. The variability in fNIRS signals across participants means that the model might not generalize to the general population. Further studies are needed to explore the effect of demographic factors on trust measurement. Collecting a larger dataset with various demographic participants can help test the generalizability of the model. A larger dataset would also permit the use of deep learning models that could provide better classification performance. Our study tasks were kept simple to remove any confounders, however, this might limit the ecological validity of the study. Further research can test this approach on ecologically valid tasks such as interactive applications, games, etc.

4.2 Implications

A brain activity-based trust classifier can improve the design and adaptation of AI systems to better meet user needs and preferences. By accurately detecting users' trust levels in real time, systems can dynamically adjust their behavior, interface, and decision-making processes to foster trust and enhance user experience. Such a classifier can also help conduct in-depth studies into trust dynamics in human-AI interactions by removing the reliance on self-report trust surveys. In workplace environments, fNIRS-based trust classification can inform the design of human-machine interfaces, automation systems, and collaborative workspaces. Understanding how trust evolves over time and in response to different stimuli can help optimize task allocation, decision support, and team coordination to enhance productivity and safety. In educational settings, such a model can inform the development of adaptive learning systems and intelligent tutoring systems. By monitoring students' trust levels during learning tasks, systems can provide personalized feedback, scaffold learning experiences, and improve teacher training to optimize educational outcomes.

5 Conclusions

This study utilized functional Near-Infrared Spectroscopy (fNIRS) as an emerging neuroimaging tool to investigate the neural correlates of trust. Through a series of experiments, we successfully demonstrated the feasibility of using fNIRS to classify trust levels based on hemodynamic responses in the prefrontal cortex. Our results contribute to the growing body of literature on the neural basis of trust. This deeper understanding of trust dynamics has significant implications for various fields, including psychology, marketing, human-computer interaction, and so-

cial neuroscience. Furthermore, our findings hold promise for the development of trust-based interventions and applications, such as trust-aware human-computer interfaces and trust-building interventions in clinical settings. The non-invasive nature and portability of fNIRS make it a practical tool for studying trust dynamics in real-world settings. It could be deployed on a computer or other wearable devices to measure the wearer's trust. Future work in this study can include combining fNIRS data with other behavioral metrics to improve accuracy. We will also explore more advanced classification models to improve the results.

Acknowledgments

Author Contributions: Conceptualization, D.B.; methodology, D.B.; software, D.B. and R.L.; validation, D.B, R.L. and L.A.; formal analysis, D.B.; investigation, D.B. and R.L.; resources, D.B.; data curation, D.B.; writing—original draft preparation, F.C.; writing—review and editing, R.L. and L.A.; visualization, D.B.; supervision, D.B.; project administration, D.B.; funding acquisition, D.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by a Fairfield University School of Engineering and Computing faculty startup grant (D.B.) and a NASA Connecticut Space Grant Consortium grant (D.B.)

Institutional Review: The study was approved by the Institutional Review Board Fairfield University (protocol code 3845 and approved on 5/13/2022).

Informed Consent: Informed consent was obtained from all subjects involved in the study.

Data Availability: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

[1] Ayaz, H., Onaral, B., İzzetoğlu, K., Shewokis, P. A., McKendrick, R., & Parasuraman, R. (2013). Continuous monitoring of brain dynamics with functional near infrared spectroscopy as a tool for neuroergonomic research: empirical examples and a technological development. *Frontiers in Human Neuroscience*, 7.

[2] Pan, Y., Borragán, G., & Peigneux, P. (2019). Applications of functional near-infrared spectroscopy in fatigue, sleep deprivation, and social cognition. *Brain Topography*, 32(6), 998-1012.

[3] Bandara, D., Hirshfield, L., & Velipasalar, S. (2019). Classification of affect using deep learning on brain blood flow data. *Journal of Near Infrared Spectroscopy*, 27(3), 206-219.

[4] Bandara, D., Velipasalar, S., Bratt, S., & Hirshfield, L. (2018). Building predictive models of emotion with functional near-infrared spectroscopy. *International Journal of Human-Computer Studies*, 110, 75-85.

[5] Perelló-March, J., Burns, C., Woodman, R., Elliott, M. T., & Birrell, S. A. (2022). Using fnirs to verify trust in highly automated driving. *SSRN Electronic Journal*.

[6] Eloy, L., Doherty, E., Spencer, C. A., Bobko, P., & Hirshfield, L. M. (2022). Using fnirs to identify transparency- and reliability-sensitive markers of trust across multiple timescales in collaborative human-human-agent triads. *Frontiers in Neuroergonomics*, 3.

[7] Pinti, P., Aichelburg, C., Gilbert, S. J., Hamilton, A. F. d. C., Hirsch, J., Burgess, P. W., & Tachtsidis, I. (2018). A review on the use of wearable functional near-infrared spectroscopy in naturalistic environments. *Japanese Psychological Research*, 60(4), 347-373.

[8] Lee, J., & See, K. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 46(1), 50-80.

[9] Madsen, M. A., & Gregor, S. (2000). Measuring Human-Computer Trust. In *Proceedings of the 11th Australasian Conference on Information Systems*. Brisbane, Australia.

[10] McAllister, D. J. (1995). Affect- and Cognition-Based Trust as Foundations for Interpersonal Cooperation in Organizations. *The Academy of Management Journal*, 38(1), 24-59.

[11] Lewicki, R., & Brinsfield, C. (2011). Measuring Trust Beliefs and Behaviours. In F. Lyon, G. Möllering, & M. Saunders (Eds.), *Handbook of Research Methods on Trust* (pp. 29-39). Edward Elgar.

[12] Wang, M., Hussein, A., Rojas, R. F., Shafi, K., & Abbass, H. A. (2018). EEG-based neural correlates of trust in human-autonomy interaction. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 350-357). IEEE.

[13] Filkowski, M. M., Anderson, I. W., & Haas, B. W. (2016). Trying to trust: Brain activity during interpersonal social attitude change. *Cognitive, Affective, & Behavioral Neuroscience*, 16, 325-338.

[14] Hopko, S. K., & Mehta, R. K. (2024). Trust in Shared-Space Collaborative Robots: Shedding Light on the Human Brain. *Human Factors*, 66(2), 490-509.

[15] Akash, K., Hu, W.-L., Jain, N., & Reid, T. (2018). A classification model for sensing human trust in machines using EEG and GSR. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(4), 1-20.

[16] Eloy, L., Doherty, E. J., Spencer, C. A., Bobko, P., & Hirshfield, L. (2022). Using fNIRS to identify transparency-and reliability-sensitive markers of trust across multiple timescales in collaborative human-human-agent triads. *Frontiers in Neuroergonomics*, 3, 838625.

[17] Yin, M., Vaughan, J. W., & Wallach, H. (2019). Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19)* (pp. 1-12). Association for Computing Machinery.

[18] Frazier, M. L., Johnson, P. D., & Fainshmidt, S. (2013). Development and validation of a propensity to trust scale. *Journal of Trust Research*, 3(2), 76-97.

[19] McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, 13(3), 334-359.

[20] Dong, J., Wang, W., & Tan, T. (2013). CASIA Image Tampering Detection Evaluation Database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*. IEEE.

[21] Peirce, J. W. (2007). PsychoPy–Psychophysics software in Python. *Journal of Neuroscience Methods*, 162(1-2), 8-13.

[22] Oviatt, S., & Cohen, P. R. (2000). Perceptual user interfaces: multimodal interfaces that process what comes naturally. *Communications of the ACM*, 43(3), 45-53.

[23] Firoz, K. F., Seong, Y., & Oh, S. (2022). A neurological approach to classify trust through EEG signals using machine learning techniques. In *2022 IEEE 3rd International Conference on Human-Machine Systems (ICHMS)* (pp. 1-6). IEEE.

[24] Khan, M. J., & Hong, K. S. (2015). Passive BCI based on drowsiness detection: an fNIRS study. *Biomedical Optics Express*, 6(10), 4063-4078.

[25] Oksanen, A., Savela, N., Latikka, R., & Koivula, A. (2020). Trust toward robots and artificial intelligence: an experimental approach to human–technology interactions online. *Frontiers in Psychology*, 11.

[26] Humr, S. A., Canan, M., & Demir, M. (2023). Temporal evolution of trust in artificial intelligence-supported decision-making. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 67(1), 1936-1941. https://doi.org/10.1177/21695067231193672

[27] Kahr, P., Rooks, G., Willemsen, M. C., & Snijders, C. J. (2023). It seems smart, but it acts stupid. development of trust in ai advice in a repeated legal decision-making task. Arxiv. https://doi.org/10.31234/osf.io/9zr3u

[28] Abbaszadeh, H., Sreedharan, S., & Kambhampati, S. (2023). A mental model based theory of trust. Arxiv. https://doi.org/10.48550/arxiv.2301.12569

[29] Choo, S., Sanders, N., Kim, N., Kim, W., Nam, C. S., & Fitts, E. P. (2019). Detecting human trust calibration in automation: a deep learning approach. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 63(1), 88-90. https://doi.org/10.1177/1071181319631298

[30] Kohn, S., Visser, E. d., Wiese, E., Lee, Y., & Shaw, T. H. (2021). Measurement of trust in automation: a narrative review and reference guide. Frontiers in Psychology, 12. https://doi.org/10.3389/fpsyg.2021.604977

[31] Oh, S., Seong, Y., Sun, Y., & Park, S. (2020). Neurological measurement of human trust in automation using electroencephalogram. INTERNATIONAL JOURNAL of FUZZY LOGIC and INTELLIGENT SYSTEMS, 20(4), 261-271. https://doi.org/10.5391/ijfis.2020.20.4.261

[32] Visser, E. J. d., 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, 12. https://doi.org/10.3389/fnhum.2018.00309

[33] Jung, E., Dong, S., & Lee, S. Y. (2019). Neural correlates of variations in human trust in human-like machines during non-reciprocal interactions. Scientific Reports, 9(1). https://doi.org/10.1038/s41598-019-46098-8

[34] Shafiei, S. B., Hussein, A. A., Muldoon, S. F., & Guru, K. A. (2018). Functional brain states measure mentor-trainee trust during robot-assisted surgery. Scientific Reports, 8(1). https://doi.org/10.1038/s41598-018-22025-1

[35] Fouragnan, E., Queirazza, F., Retzler, C., Mullinger, K. J., & Philiastides, M. G. (2017). Spatiotemporal neural characterization of prediction error valence and surprise during reward learning in humans. Scientific Reports, 7(1). https://doi.org/10.1038/s41598-017-04507-w

[36] Sanders, N., Choo, S., Kim, N., Nam, C. S., & Fitts, E. P. (2019). Neural correlates of trust during an automated system monitoring task: preliminary results of an effective connectivity study. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 63(1), 83-87. https://doi.org/10.1177/1071181319631409

[37] Winston, J. S., Strange, B. A., O'Doherty, J., & Dolan, R. J. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces. Nature neuroscience, 5(3), 277-283.

[38] Aimone, J. A., Houser, D., & Weber, B. (2014). Neural signatures of betrayal aversion: an fMRI study of trust. Proceedings of the Royal Society B: Biological Sciences, 281(1782), 20132127.

[39] Stanley, D. A., Sokol-Hessner, P., Fareri, D. S., Perino, M. T., Delgado, M. R., Banaji, M. R., & Phelps, E. A. (2012). Race and reputation: perceived racial group trustworthiness influences the neural correlates of trust decisions. Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1589), 744-753.

[40] Noah, J. A., Zhang, X. Z., Dravida, S., DiCocco, C., Suzuki, T., Aslin, R. N., Tachtsidis, I., & Hirsch, J. (2021). Comparison of short-channel separation and spatial domain filtering for removal of non-neural components in functional near-infrared spectroscopy signals. Neurophotonics, 8(01). https://doi.org/10.1117/1.nph.8.1.015004

[41] Scholkmann, F., Kleiser, S., Metz, A. J., Zimmermann, R., Pavia, J. M., Wolf, U., & Wolf, M. (2014). A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology. Neuroimage, 85, 6-27.

[42] Yao, S., Song, Y., Yu, Y., & Guo, B. (2020). A study of group decision-making for green technology adoption in micro and small enterprises. Journal

of Business & Industrial Marketing, 36(1), 86-96. https://doi.org/10.1108/jbim-02-2020-0093

[43] Whitney, R. L., White, A. E. C., Rosenberg, A. S., Kravitz, R. L., & Kim, K. (2021). Trust and shared decision-making among individuals with multiple myeloma: a qualitative study. Cancer Medicine, 10(22), 8040-8057. https://doi.org/10.1002/cam4.4322

[44] Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. Ergonomics, 35(10), 1243-1270. https://doi.org/10.1080/00140139208967392

[45] Oh, S., & Seong, Y. (2019). Neural investigation of human trust using electroencephalogram. Iarjset, 6(7), 71-78. https://doi.org/10.17148/iarjset.2019.6712

[46] Uslu, S., Kaur, D., Rivera, S. J., Durresi, A., Babbar-Sebens, M., & Tilt, J. H. (2020). Control theoretical modeling of trust-based decision making in food-energy-water management. Complex, Intelligent and Software Intensive Systems, 97-107. https://doi.org/10.1007/978-3-030-50454-0-10

[47] Ryan, M. (2020). In ai we trust: ethics, artificial intelligence, and reliability. Science and Engineering Ethics, 26(5), 2749-2767. https://doi.org/10.1007/s11948-020-00228-y

[48] Ferrario, A., Loi, M., & Viganò, E. (2019). In ai we trust incrementally: a multi-layer model of trust to analyze human-artificial intelligence interactions. Philosophy & Technology, 33(3), 523-539. https://doi.org/10.1007/s13347-019-00378-3

[49] Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., & Tan, C. (2021). Towards a science of human-ai decision making: a survey of empirical studies. Arxiv. https://doi.org/10.48550/arxiv.2112.11471

[50] Okamura, K., & Yamada, S. (2020). Adaptive trust calibration for human-ai collaboration. Plos One, 15(2), e0229132. https://doi.org/10.1371/journal.pone.0229132

[51] Mou, Y., Xu, T., & Hu, Y. (2023). Uniqueness neglect on consumer resistance to ai. Marketing Intelligence & Planning, 41(6), 669-689. https://doi.org/10.1108/mip-11-2022-0505

[52] Vereschak, O., Bailly, G., & Caramiaux, B. (2021). How to evaluate trust in ai-assisted decision making? a survey of empirical methodologies. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1-39. https://doi.org/10.1145/3476068

[53] Hou, K., Hou, T., & Cai, L. (2023). Exploring trust in human–ai collaboration in the context of multiplayer online games. Systems, 11(5), 217. https://doi.org/10.3390/systems11050217

[54] Schelble, B. G., Lopez, J., Textor, C., Zhang, R., McNeese, N. J., Pak, R., & Freeman, G. (2022). Towards ethical ai: empirically investigating dimensions of ai ethics, trust repair, and performance in human-ai teaming. Human Factors: The Journal of the Human Factors and Ergonomics Society, 66(4), 1037-1055. https://doi.org/10.1177/00187208221116952

[55] Karran, A. J., Demazure, T., Hudon, A., Sénécal, S., & Léger, P. (2022). Designing for confidence: the impact of visualizing artificial intelligence decisions. Frontiers in Neuroscience, 16. https://doi.org/10.3389/fnins.2022.883385

[56] Garcia, K., Mishler, S., Xiao, Y., Hu, B., Still, J. D., & Chen, J. (2022). Drivers' understanding of artificial intelligence in automated driving systems: a study of a malicious stop sign. Journal of Cognitive Engineering and Decision Making, 16(4), 237-251. https://doi.org/10.1177/15553434221117001

[57] Webber, S. S. (2008). Development of cognitive and affective trust in teams. Small Group Research, 39(6), 746-769. https://doi.org/10.1177/1046496408323569

[58] Chen, X., Eberly, M. B., Chiang, T., Farh, J., & Cheng, B. (2011). Affective trust in chinese leaders. Journal of Management, 40(3), 796-819. https://doi.org/10.1177/0149206311410604

[59] Colquitt, J. A., Scott, B. A., & LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance. Journal of Applied Psychology, 92(4), 909-927. https://doi.org/10.1037/0021-9010.92.4.909

[60] Xiao, L., Guo, Z., D'Ambra, J., & Fu, B. (2016). Building loyalty in e-commerce. Program, 50(4), 431-461. https://doi.org/10.1108/prog-04-2016-0040

[61] Schaubroeck, J., Lam, S. S. K., & Peng, A. C. (2011). Cognition-based and affect-based trust as mediators of leader behavior influences on team performance. Journal of Applied Psychology, 96(4), 863-871. https://doi.org/10.1037/a0022625

[62] Hall, M. A., Dugan, E., Zheng, B., & Mishra, A. K. (2001). Trust in physicians and medical institutions: what is it, can it be measured, and does it matter? The Milbank Quarterly, 79(4), 613-639. https://doi.org/10.1111/1468-0009.00223

[63] Meng, F., Guo, X., Peng, Z., Ye, Q., & Lai, K. H. (2021). Trust and elderly users' continuance intention regarding mobile health services: the contingent role of health and technology anxieties. Information Technology & People, 35(1), 259-280. https://doi.org/10.1108/itp-11-2019-0602

[64] Rahayu, A., & Baridwan, Z. (2020). The influence of sponsored post towards the urge to buy implusively on the information technology system of the social media of instagram. AKRUAL: Jurnal Akuntansi, 11(2), 95. https://doi.org/10.26740/jaj.v11n2.p95-109

[65] Hall, M. A., Zheng, B., Dugan, E., Camacho, F., Kidd, K. E., Mishra, A. K., & Balkrishnan, R. (2002). Measuring patients' trust in their primary care providers. Medical Care Research and Review, 59(3), 293-318. https://doi.org/10.1177/1077558702059003004

[66] Lewicki, R. J., Tomlinson, E. C., & Gillespie, N. (2006). Models of interpersonal trust development: theoretical approaches, empirical evidence, and future directions. Journal of Management, 32(6), 991-1022. https://doi.org/10.1177/0149206306294405

[67] Chita-Tegmark, M., Law, T., Rabb, N., & Scheutz, M. (2021). Can you trust your trust measure? Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. https://doi.org/10.1145/3434073.3444677

[68] Sheng, S., Pakdamanian, E., Han, K., Kim, B., Tiwari, P., Kim, I., & Lu, F. (2019). A case study of trust on autonomous driving. 2019 IEEE Intelligent Transportation Systems Conference (ITSC). https://doi.org/10.1109/itsc.2019.8917251

[69] Zieger, S., Dong, J., Taylor, S., Sanford, C., & Jeon, M. (2023). Happiness and high reliability develop affective trust in in-vehicle agents. Frontiers in Psychology, 14. https://doi.org/10.3389/fpsyg.2023.1129294

[70] Cornejo, C., Cuadros, Z., Morales, R., & Mayor, J. P. (2017). Interpersonal coordination: methods, achievements, and challenges. Frontiers in Psychology, 8. https://doi.org/10.3389/fpsyg.2017.01685

[71] Wang, J. (2018). Gaze behavior, skin conductance, and trust in automation [Master's thesis]. University of Twente.

[72] Chua, J. S. K., Xu, H., & Lye, S. W. (2021). The face of trust: using facial action units (aus) as indicators of trust in automation. Human Interaction, Emerging Technologies and Future Systems V, 265-273. https://doi.org/10.1007/978-3-030-85540-6_34

[73] Walker, F., Wang, J., Martens, M., & Verwey, W. B. (2019). Gaze behaviour and electrodermal activity: objective measures of drivers' trust in automated vehicles. Transportation Research Part F: Traffic Psychology and Behaviour, 64, 401-412. https://doi.org/10.1016/j.trf.2019.05.021

[74] Kret, M. E., Fischer, A. H., & Dreu, C. K. W. D. (2015). Pupil mimicry correlates with trust in in-group partners with dilating pupils. Psychological Science, 26(9), 1401-1410. https://doi.org/10.1177/0956797615588306

[75] Noah, J. A., Zhang, X., Dravida, S., Ono, Y., Naples, A., McPartland, J. C., & Hirsch, J. (2020). Real-time eye-to-eye contact is associated with cross-brain neural coupling in angular gyrus. Frontiers in Human Neuroscience, 14. https://doi.org/10.3389/fnhum.2020.00019

[76] Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human factors, 57(3), 407-434.

[77] Ajenaghughrure, I. B., Sousa, S. D. C., & Lamas, D. (2020). Measuring Trust with Psychophysiological Signals: A Systematic Mapping Study of Approaches Used. Multimodal Technologies and Interaction, 4(3), 63. https://doi.org/10.3390/mti4030063

[78] Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., & Chen, F. (2017). User Trust Dynamics: An Investigation Driven by Differences in System Performance. Proceedings of the 22nd International Conference on Intelligent User Interfaces, 307-317. https://doi.org/10.1145/3025171.3025219

[79] Hopko, S., & Mehta, R. K. (2022). Trust in shared-space collaborative robots: shedding light on the human brain. Human Factors: The Journal of the Human Factors and Ergonomics Society, 66(2), 490-509. https://doi.org/10.1177/00187208221109039

[80] Moretto, G., Sellitto, M., & Pellegrino, G. d. (2013). Investment and repayment in a trust game after ventromedial prefrontal damage. Frontiers in Human Neuroscience, 7. https://doi.org/10.3389/fnhum.2013.00593

[81] Wei, Z., Zhao, Z., & Zheng, Y. (2019). Following the majority: social influence in trusting behavior. Frontiers in Neuroscience, 13. https://doi.org/10.3389/fnins.2019.00089