International Journal of Human–Computer Interaction

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/hihc20

Investigating Trust in Human-AI Collaboration for a Speech-Based Data Analytics Task

Abdullah Aman Tutul, Ehsanul Haque Nirjhar & Theodora Chaspari

To cite this article: Abdullah Aman Tutul, Ehsanul Haque Nirjhar & Theodora Chaspari (22 Mar 2024): Investigating Trust in Human-AI Collaboration for a Speech-Based Data Analytics Task, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2024.2328910

To link to this article: https://doi.org/10.1080/10447318.2024.2328910

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2024.2328910

Investigating Trust in Human-AI Collaboration for a Speech-Based Data Analytics Task

Abdullah Aman Tutul[a], Ehsanul Haque Nirjhar[a], and Theodora Chaspari[b]

[a]Texas A&M University, College Station, TX, USA; [b]University of Colorado Boulder, Boulder, CO, USA

ABSTRACT

Complex real-world problems can benefit from the collaboration between humans and artificial intelligence (AI) to achieve reliable decision-making. We investigate trust in a human-in-the-loop decision-making task, in which participants with background on psychological sciences collaborate with an explainable AI system for estimating one's anxiety level from speech. The AI system relies on the explainable boosting machine (EBM) model which takes prosodic features as the input and estimates the anxiety level. Trust in AI is quantified via self-reported (i.e., administered via a questionnaire) and behavioral (i.e., computed as user-AI agreement) measures, which are positively correlated with each other. Results indicate that humans and AI depict differences in performance depending on the characteristics of the specific case under review. Overall, human annotators' trust in the AI increases over time, with momentary decreases after the AI partner makes an error. Annotators further differ in terms of appropriate trust calibration in the AI system, with some annotators over-trusting and some under-trusting the system. Personality characteristics (i.e., agreeableness, conscientiousness) and overall propensity to trust machines further affect the level of trust in the AI system, with these findings approaching statistical significance. Results from this work will lead to a better understanding of human-AI collaboration and will guide the design of AI algorithms toward supporting better calibration of user trust.

KEYWORDS

Explainable AI; transparency; human trust; trust calibration

1. Introduction

Artificial intelligence (AI) algorithms have been heralded as promising tools for supporting decision-making due to their ability to process large data samples and capture fine-grain patterns in data that are not easily discernible by a human observer (Ezer et al., 2019). Recently, AI algorithms have become more prevalent in decision-making tasks that are complex, sensitive, and carry significant consequences, such as the ones pertaining to health, education, command and control, and commerce (Phillips-Wren, 2012). Certain complex tasks within these domains require a collaborative approach, as neither humans nor AI agents can achieve success independently. AI is capable of finding patterns from vast amounts of data beyond human capacity, but struggles with cases deviating from learned patterns (D'Amour et al., 2022). Conversely, humans possess unique skills such as intuition, inventiveness, and common sense, which are inherently more challenging for current AI systems (Hemmer et al., 2021). In this context, collaborative decision-making between humans and AI involves the two leveraging their complementary expertise and working side-by-side to solve complex decision-making tasks that cannot be perfectly solved by either party.

In order for users to understand when they should trust the AI and when they should rely on their judgment for particular decisions, proper trust calibration in human-AI teaming is crucial. Trust of a human agent in an automated agent can be defined as the human agent's attitude that the automated agent will help them achieve their goal in a situation characterized by uncertainty and vulnerability (Lee & See, 2004). The notion of trust in a human-AI environment differs from that of automation. Since automation is characterized by static rules, trust in automation is often associated with the clarity and predictability of its actions (e.g., users may understand the mechanisms and logic behind a robot's operations) (Kaplan et al., 2023). On the contrary, trust in AI often hinges on explainability and involves high uncertainty, especially as AI systems become more complex. In the context of human-AI teaming, trust can be defined as the "human agent's willingness to rely on the AI system's output driven upon positive expectations that the AI system is accurate and beneficial to the focal task" (Gillespie et al., 2023). In order for the AI agent to become a trusted teammate, it needs to be flexible and adaptive to the human partner and the environment in which it operates. At the same time, humans should be able to understand the capacity of the AI agent and calibrate their trust to the abilities and performance of the system (Bansal et al., 2019; Lee & See, 2004). Miscalibrated trust may lead to wrongful decisions with severe consequences (Kaindl & Svetinovic, 2019; Okamura & Yamada, 2018; Parasuraman & Riley, 1997). Human teammates who over-trust the AI tend to

CONTACT Abdullah Aman Tutul abdullahaman633@tamu.edu Texas A&M University, College Station, TX, USA

overestimate the ability of the AI agent to solve the problem, therefore they agree with the decision of the AI system even when it is wrong (Payne et al., 2008). On the contrary, human teammates who under-trust the AI agent tend to underestimate its capacity, thus disagreeing with the AI decision even though it is correct. Trust calibration is particularly important in high-stake complex domains which tend to be bounded by legal or ethical constraints, thus, highly benefiting by the complementary skills of the human and the AI agent (Bansal et al., 2021; Jarrahi, 2018).

Trust is a multifaceted concept that encompasses both cognitive and affective dimensions, making its measurement a challenging yet critical aspect of collaborative decision-making systems. Measures of trust vary across disciplines and are highly dependent on the context of the application. Early studies have measured trust via self-reports, which have been administered once at the end of the collaborative task, pre/post-task, multiple times at the end of each trial, or over pre-specified intervals (Alarcon et al., 2018; Schaefer, 2013). Given the subjectivity of self-reported measures, other work has used behavioral measures to infer trust, such as the extent to which a human user agrees with the automated system, depicts over-confidence to the automation, or underuses the automation (Drnec et al., 2016). With the advancement of sensor capabilities that continuously collect multimodal data, recent work has further introduced signal-based measures of trust. These include neural measures of action monitoring and error awareness that are captured via electroencephalogram signals (de Visser et al., 2018; Dong et al., 2015), as well as acoustic and linguistic markers that capture characteristics of trusted speech (Chen et al., 2020; Levitan et al., 2015).

Prior work on human trust in automation has demonstrated that trust depends on individual differences, contextual factors, and system characteristics (Lee & See, 2004; Siau & Wang, 2018). Particularly, individual differences, such as one's overall trust propensity, personality, and task expertise, can influence the initial levels of trust as well as the way trust evolves over time while the user is interacting with the autonomous system (Böckle et al., 2021; Hoff & Bashir, 2015; Müller et al., 2019). Contextual factors that impact trust in automation include social norms and expectations regarding the system, as well as affective and cognitive variables that describe the state of the user, such as fatigue, mood, and perceived cognitive demand (Merritt, 2011). In terms of system characteristics, prior work has explored the competence of a system and its ability to explain its decision as additional factors of trust (Cheng et al., 2019; Lai & Tan, 2019; Okamura & Yamada, 2018; Yang et al., 2020). Trust in AI may vary over time based on the system's performance on specific tasks. Positive experiences and successful outcomes can bolster trust, while errors or suboptimal results may lead to fluctuations in trust levels (Schaefer et al., 2014). Explainable AI (XAI) can play a crucial role in assisting human users to properly calibrate their trust in AI since it explains the predictions in a way that is comprehensible by humans. Despite many state-of-the-art AI models depicting equivalent or even better performance than humans in complex tasks, a significant number of these operate as blackbox models, leaving users unable to understand the reason why the AI model makes a particular decision. XAI focuses on opening these blackbox models and unveiling their reasoning. This can be often achieved via providing global explanations that offer a broad comprehension of a model's learned concepts (Guyon & Elisseeff, 2003; Kim et al., 2018) and local explanations that seek to explain the logic behind a specific AI decision (Baehrens et al., 2010; Ribeiro et al., 2016). Leveraging XAI with appropriate interface design in collaborative human-AI decision-making tasks can potentially contribute to trust calibration (Naiseh et al., 2021, 2023). While there has been an extensive research on how user characteristics and system factors affect trust in automated systems such as robotic agents and autonomous vehicles (Bawack et al., 2021; Pop et al., 2015), the effect of such factors on human trust in AI is still under-explored (Tutul et al., 2021).

Here, we investigate a human-in-the-loop decision making task in which human annotators and AI work side-by-side to estimate one's anxiety levels from speech. Human annotators with a background in psychology collaborated with an explainable AI algorithm to provide a final decision on a speaker's level of anxiety. We measure trust in AI using self-reported (i.e., administered via a questionnaire) and behavioral (i.e., the extent to which the annotator agrees with the AI) measures. We aim to answer the following research questions: *RQ1: Do humans and AI depict differences in performance in the considered anxiety estimation task?* As humans and AI collaborate to estimate anxiety levels from speech, it raises intriguing questions about the unique decision-making process of each party in the considered task. Human annotators may perform better than the AI in cases where they leverage contextual understanding and consider behavioral nuances, while they might perform worse when they need to process low-level acoustic measures. Answering this question can help us to unravel the unique contributions and limitations of both human and AI agents in a task of significant societal impact. *RQ2: What is the association between behavioral and self-reported measures of trust in the considered human-AI collaboration task?* Given the inherent complexity of trust dynamics in human-AI interactions, understanding the relationship between behavioral and self-reported measures of trust is crucial for gaining insights into the alignment or potential disparities between subjective perceptions and objective behaviors. *RQ3: To what extent is human trust calibrated with the capacity of the AI system?* Trust is foundational in fostering user acceptance and effective collaboration. However, for trust to be effective, it needs to be calibrated accurately with the AI system's capabilities. Understanding the extent to which trust aligns with the actual capacity of the AI system is crucial for developing effective collaborative environments for trustworthy decision-making. *RQ4: To what extent does trust in AI vary over time?* Understanding how trust evolves over time is crucial for designing adaptive AI systems that can respond to changing user perceptions and requirements. Users may experience changes in trust as they interact more

with the AI system. *RQ5: To what extent is trust in AI affected by the characteristics of the system and the traits of the human annotator?* The motivation behind this research question arises from the recognition that trust in AI is a nuanced construct influenced by various factors. Understanding the multifaceted relationship between trust, system attributes, and human factors in the context of human-AI collaboration provides valuable insights for designing AI systems that are not only technically proficient but also align with user expectations and preferences.

Results indicate that humans and AI depict differences in performance depending on the nature of the case that is being reviewed. Behavioral and self-reported measures of trust in AI are further positively correlated. The levels of trust in AI, as well as the relation between trust in AI and AI capacity broadly differ among people with most users over-trusting the AI. Human annotators overall depict increased trust in the AI system over time, with momentary decreases in trust after the AI makes an error. The annotators' characteristics further moderate this association; participants with high propensity to trust machines and more agreeableness characteristics depict high trust in AI, whereas conscientious annotators depict low trust in the AI system, which has an increasing trend over time. However, these associations only approached statistical significance. Implications of these findings on ways to achieve effective human-AI collaboration for decision-making are discussed.

2. Prior work

With the rise of AI, both academics and practitioners have shown a growing interest in human-AI interaction, especially for addressing inherently challenging tasks that cannot be adequately solved by either of the two parties. Human trust in human-AI teaming is multifaceted and shaped by various factors such as the perceived reliability and performance of the AI, the transparency of its decision-making process, and the user's familiarity and experience with the technology (Glikson & Woolley, 2020). Moreover, trust is not a static attribute and can evolve over time based on user interactions, feedback, and the system's ability to adapt to different situations (Ezer et al., 2019). Recent studies have delved into human trust in AI during cognitively demanding tasks, utilizing both the user's adherence to AI decisions and self-reported measures from questionnaires. These investigations have explored the impact of diverse explanations, interfaces, visual representations, and spatial layouts on user trust in AI across tasks such as age estimation, medical diagnosis, university admissions, and identifying deceptive hotel reviews. Below we outline some of these studies and the main findings.

Chu et al. (2020) conducted a human-in-the-loop experiment where users were asked to predict the age of a person based on their image after viewing the decision of an AI model. Trust was quantified as the absolute difference between the user's and the model's estimate of a person's age. Users were also presented different types of explanations by the AI system, which were not found to significantly affect human trust in AI, even after controlling for the quality of the explanation. Alam and Mueller (2021) further examined the effect of different explanations on trust and satisfaction with the AI in a medical diagnosis task. As part of the study, 80 undergraduate students who acted as patients were asked to provide self-reports of perceived trust level to the AI and user satisfaction. Participants were more satisfied and trusted the AI more when the AI system explained the reason for making a particular diagnosis compared to when it explained the general diagnosis process or did not provide any explanation at all. Zhang et al. (2020) explored an AI-assisted decision making framework for an income prediction task, in which the human users and the AI system depicted comparable performance (i.e., 65% and 75% accuracy, respectively). Participants trusted the AI prediction more when the confidence level of the AI was high (i.e., above 80%). However, the confidence interval of the AI system did not appear to affect the user trust. Lundberg and Lee (2017) further assessed the effect of local explanations (i.e., explanations for a specific sample) on human trust and found that such explanations do not significantly affect trust in AI. Yang et al. (2020) explored the effect of visual representations and spatial layouts on users' trust in AI for a leaf classification task, with visual representations having a greater impact on users' trust. Cheng et al. (2019) ran an online experiment in which 199 people used several explanation interfaces provided by the AI algorithm for deciding university admissions. Interactive and white-box interfaces improved users' comprehension of the algorithm more than static and black-box interfaces. Lai and Tan (2019) analyzed how showing different explanations and accuracy statements to the user can impact human trust in AI for the task of identifying deceptive hotel reviews. They found that users trust the instances that were correctly predicted by the AI more than incorrect ones. In addition, they showed that both feature-based and example-based explanations increase trust in AI. de Brito Duarte et al. (2023) showed that AI trust in recommendation system improves when AI explanations along with feature importance and counterfactual explanations are provided to the users. Communicating the AI's accuracy to the user, irrespective of the specific numerical value, also enhanced the user's trust in the AI. Collectively, these studies provide insights into the role of explanations, visual representations, and system confidence across diverse application domains. While the impact of AI explanations on human trust has not been consistently observed, certain studies highlight benefits in trust calibration through the provision of visual representations and interactive interfaces. Interestingly, system confidence did not emerge as a consistent contributor to human trust, but knowledge regarding the system's accuracy appeared to enhance trust among users.

In addition to system-based factors, individual characteristics can also affect trust in automation. Bawack et al. (2021) employed self-reports to quantify user trust in AI via an online survey of 224 U.S. based voice shoppers and found that agreeableness and conscientiousness are positively correlated with trust, while neuroticism and extroversion are

not correlated with trust. Chien et al. (2016) analyzed the effect of personality characteristics on self-reported trust in automation for 120 participants residing in U.S., Turkey, and Taiwan population. Results suggested a significant positive correlation between agreeableness and initial trust in AI and also, between conscientiousness and initial trust in AI. However, no significant correlation was found between trust and other personality traits, such as neuroticism, openness, and extroversion. Kraus et al. (2020) analyzed the relation between the personality factors and trust in automated driving. The results indicate that neurotic people hold lower trust in automated driving, while agreeable and extrovert people depict more trust in automated driving. No association was found between trust and user openness or conscientiousness. In their study, Yang et al. (2020) did not find an association between trust and users' expertise, familiarity with the considered task, and overall propensity to trust. However, other studies indicate that users with high propensity to trust machines show higher initial trust in automation compared to their counterparts. This high initial trust declines to a larger extent when automation errors are found (Ebert et al., 2009; Madhavan et al., 2006; Merritt & Ilgen, 2008).

In terms of exploring the association between self-reported trust and behavioral trust, Sharan and Romano (2020) conducted a study in which 171 volunteers were asked to play a card game assisted by an AI system. Trust ratings were measured via a questionnaire administered at the end of the task and behavioral trust concordance was measured as the total number of responses that were same as the AI suggestion. Findings suggest a significant low-to-moderate positive association between self-reported trust and behavioral trust (i.e., $r = 0.22$, $p < 0.05$). Sofianos (2022) found that self-reported trust measures were significantly related with behavioral trust in a trust game played by two human participants. This association was moderated by one's perception of the partner's intentions. Ahmed and Salas (2009) also identified associations between behavioral and self-reported trust measures. However, these associations were influenced by the cultural background of the participants. Conversely, there is evidence that indicates an incongruent relationship between the two types of trust. Kulms and Kopp (2019) evaluated trust in a cooperative game that requires users to perform a joint activity with a computer. Results indicated that varying the degree of anthropomorphism of the agent from computer-like to human-like did not impact behavioral trust, but increased self-reported trust levels. These studies underscore the intricate dynamics between self-reported and behavioral trust, emphasizing the need for a nuanced understanding of their interplay in human-AI interactions.

This paper presents the following contributions in relation to prior work: (1) While prior studies examining trust in AI have focused on relatively objective vision-based task (Lai & Tan, 2019; Yang et al., 2020), this work considers a more subjective task of anxiety estimation based on speech; (2) This work examines users who possess a foundational understanding of the subject matter and hold domain knowledge in life science, as opposed to employing "naive" annotators recruited from the general population, such as annotators from Amazon Turk. This gives us a better sense on how users with more direct background with life science interpret the XAI decisions for anxiety detection that requires users' perception on human psychology. Particularly when the AI is intended for decision-making in critical areas such as health and education, studying trust in AI for annotators who hold some domain knowledge in life science can be more appropriate compared to utilizing naive annotators, since it can more closely approximate practical settings where AI can be deployed; (3) This study examines how AI errors affect trust in AI over time, which has been studied before in automation (Ebert et al., 2009; Madhavan et al., 2006; Merritt & Ilgen, 2008), but not adequately explored before for XAI and perceptual tasks with a lot of uncertainty. Understanding how trust evolves over time and corresponding factors is crucial for designing systems that can effectively support users across different phases of interaction; and (4) We investigated the extent to which humans and AI systems rely on the same or different acoustic measures of speech when assessing public speech anxiety. Gaining insight into the different expertise and performance of both entities provides essential knowledge that guides the assignment of tasks according to the capabilities of each, ultimately leading to task execution that is more effective and efficient.

3. Explainable AI system for estimating anxiety

3.1. Data description

We used the VerBIO dataset (Yadav et al., 2020), a multimodal bio-behavioral dataset of individuals' anxiety responses while performing public speaking tasks in real-life and virual reality (VR) settings. The data includes 78 audio recordings collected from 55 undergraduate and graduate students (23 female, 32 male) between 18 and 30 years old. Each participant performed 10 different public speaking presentations including PRE (1 session, Day 1), TEST (8 sessions, Day 2–3), and POST (1 session, Day 4) parts. During the PRE and POST sessions, participants gave a speech in a conference room in-front of a real-life audience that included professors and graduate students. The TEST portions took place in various VR environments (i.e., classroom, small theater, seminar room, boardroom) and in front of various types of VR audiences (e.g., positive, neutral, negative). Before each public speaking presentation, participants were randomly assigned to a news article from a list of topics (i.e., history, travel, business, health, nature, culture, science) and were provided 10 minutes to prepare an oral presentation. Following that, they delivered a 5-minute speech to the audience. We only used the PRE and POST sessions of the dataset, since they involved a real-life audience that can better simulate real life settings. We randomly selected four speech files out of the 78 speech files of the VerBIO dataset, which were provided as part of the annotation procedure twice in random order. Therefore, the annotators were asked to rate a total of 82 files. This served as an

additional checkpoint to evaluate the attention of each annotator in the decision making task. In order to obtain the ground truth for the study, a human expert with experience in behavioral coding listened to each audio and provided his perceived anxiety levels of the speaker on a 5-point Likert scale (i.e., 1: No anxiety, 5: Very high anxiety). The expert listened to the audio files as many times as necessary in order to make a reliable decision. We have used these scores as the ground truth in this study.

3.2. Designing an explainable AI algorithm for estimating anxiety from speech

The AI agent relied on an explainable AI algorithm based on the Explainable Boosting Machine (EBM) model (Nori et al., 2019), a glass-box model which produces interpretable explanations of the decision outcome. The EBM model's explainability is rooted in three key factors: (1) The model's input comprises four highly interpretable features that assess speech's prosodic characteristics, directly linked to anxiety; (2) Global explanation graphs generated by the EBM model offer an overarching understanding of the anxiety outcome's dependence on each feature, providing users with an overall view on how each feature is associated with anxiety; and (3) At the audio level, the local explanation graph further elucidates the dependence of anxiety outcomes on individual features, enabling users to better understand specific acoustic patterns within each audio file that influenced the EBM model's decision.

The EBM estimates the speaker's levels of anxiety based on four acoustic features, including the mean pause duration, loudness (i.e., computed as the logarithm of the mean square energy), jitter (i.e., computed as the frame-to-frame pitch period length deviations), and shimmer (i.e., computed as the frame-to-frame amplitude deviations between pitch periods). The average of these measurements was calculated for the spoken segments of each audio clip over an analysis window of 30 miliseconds. These were selected since these are intuitive, easily interpretable, and related to the level of anxiety (Batrinca et al., 2013; Chollet et al., 2016). The EBM model finds the contribution of each feature to the outcome of the model, and has comparable performance to state-of-the-art ML methods, such as bagging and boosting. It is a generalized additive model that follows the following mathematical formulation:

$$g(\mathcal{E}[y]) = \beta_0 + \sum f_j(x_j) \tag{1}$$

Where g is the identity function in our model, β_0 is the intercept, and $\mathcal{E}$ is the expected value. The function f_j indicates how each feature x_j contributes to the model's prediction for estimating the level of anxiety. The pair-wise feature interactions were not considered in our experiment, since they would increase the complexity of the model and would likely be less intuitive for the users (Lou et al., 2012). Training is conducted on one feature at a time in a round-robin fashion using very low learning rate cycling through all features x_j and learning the best feature function f_j for each feature x_j and the outcome of interest y. The contribution of each feature to the final prediction renders the EBM highly interpretable and a good fit for this study given its focus on the collaboration between the AI system and a human annotator, who will rely on this explanation to interpret the system's decision-making process. The contribution of each feature x_j to the final prediction can be understood by plotting f_j. We used leave-one-sample-out cross validation for evaluation due to the small number of samples. As hyper-parameters, we used learning rate of 0.01, early stopping rounds of 100, and the total number of maximum boosting rounds of 10,000. The Spearman's correlation between the estimated and actual levels of anxiety for this model was 0.261 ($p < 0.05$), a result equivalent with prior work on the same (Yadav et al., 2020) and other (Booth et al., 2022) datasets. We could have used more features to improve the model performance but prior research (Poursabzi-Sangdeh et al., 2021) shows that using fewer features in a glassbox model results in users being able to simulate the model predictions more accurately. We used Spearman's correlation, since it captures the monotonic (i.e., rank) relation between two variables and the focal ground truth of anxiety is an ordinal variable (i.e., 1, …5) rather than continuous (Hauke & Kossowski, 2011; Rebekić et al., 2015).

Using the output of the EBM model, we can visualize the correlation between each feature and the state of anxiety based on all the data via the global explanation graph (Figure 1). The feature values are shown on the graph's x-axis, and their contributions to the anxiety outcome are shown on the y-axis. Positive contribution values (i.e., solid blue line; Figure 1) show a positive correlation between the associated acoustic feature and the anxiety outcome, whilst negative contribution values show the reverse relationship. Additionally, the graph shows, through shaded areas, the model's level of confidence for each feature value. Higher levels of decision uncertainty are shown by thicker shaded areas. In Figure 1(a), we observe that anxiety related to public speaking tends to decrease as speakers get more loud during their speech, whereas the relationship between public speech anxiety and speech pause duration is the opposite (SEE Figure 1(b)). Prior studies also show that confident speakers depict high loudness (Monarth & Kase, 2007) and take fewer pauses (Jiang & Pell, 2017). The EBM model also has the ability to explain the decision for each sample via the local explanation graph, which depicts the contribution of each feature in the level of anxiety. Figure 2 presents a local explanation graph for a sample audio, for which pause duration is the most important feature that is associated with increased anxiety level, while loudness is the least important one. As seen in the same figure, the annotators were further provided with a description of the local explanation graph which indicated the acoustic features that influenced the most and the least the decision outcome, as well as the features that were positively and negatively correlated with the outcome.

3.3. User interface to enable human-AI collaboration

We created a web interface (Figure 3) through which annotators interacted with the EBM model. The main elements

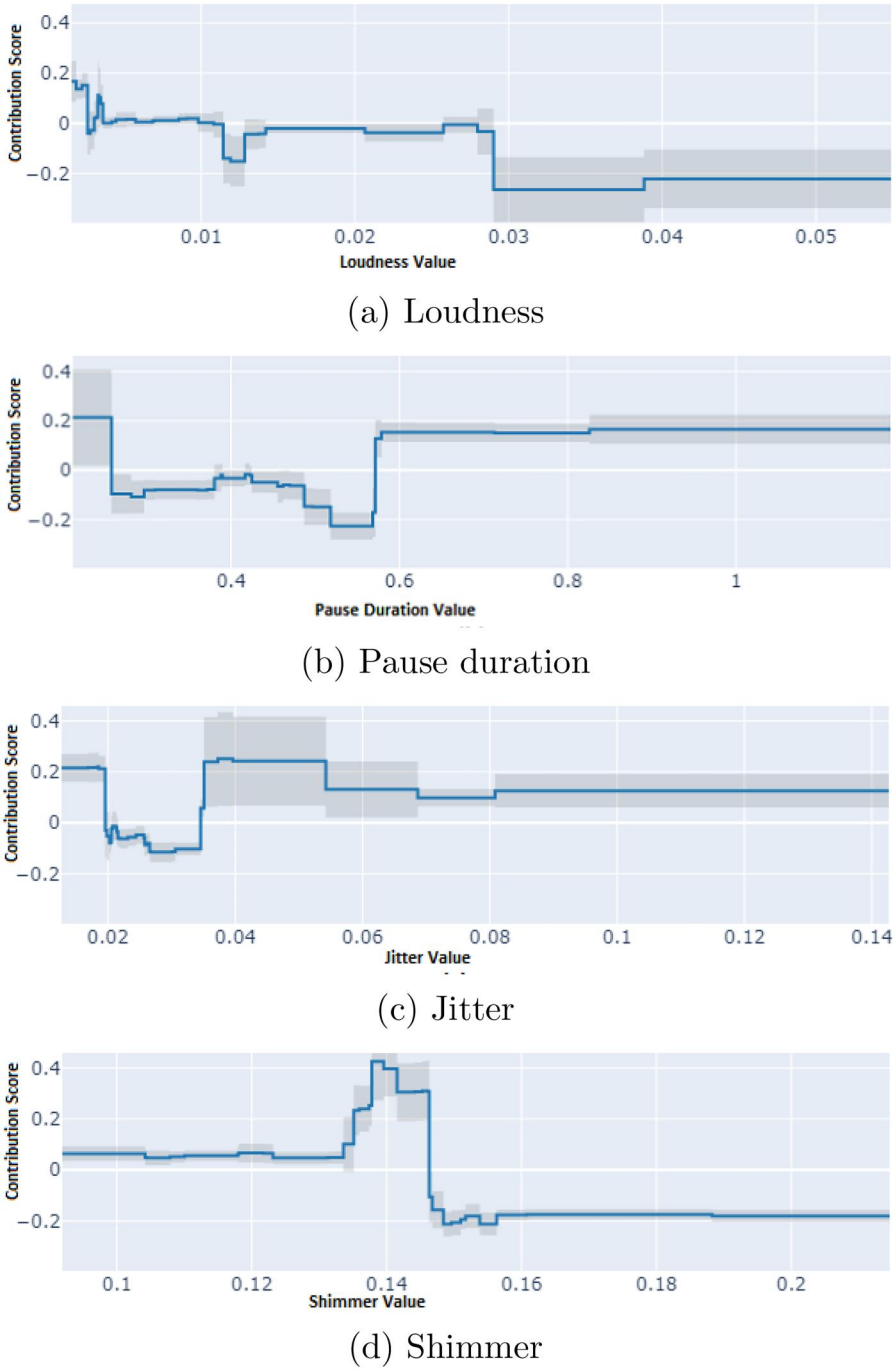


Figure 1. Global explanation graphs provided by the explainable boosting machine (EBM) that capture the effect of different features in estimating the anxiety levels. The feature values are shown on the graph's *x*-axis, and their contributions to the anxiety outcome are shown on the *y*-axis.

of the interface are: (1) An audio player for listening to the audio files; (2) The global explanation graphs (Section 3.2) explaining the contribution of each feature to the estimated anxiety level; (3) The local explanation graph (Section 3.2) explaining the relative importance of each feature for each audio file; (4) A comment box where participants can comment on the reasons behind their decision for each audio file; and (5) Help buttons so that annotators can quickly refer to explanations about the AI model as needed. We piloted the interface with five annotators, who tested the initial design. Annotators were provided with a mini-tutorial that explained the goal of the project, the role of each acoustic feature, as well as the functionality and output of the EBM model. Based on the comments of these five annotators, we introduced various "help" buttons to the interface that directed the user to a brief description of each component, such as the global and local explanation graphs and the features.

4. User study design

4.1. Participant recruitment

The main goal of the study is to analyze how people who possess a foundational understanding of the subject matter

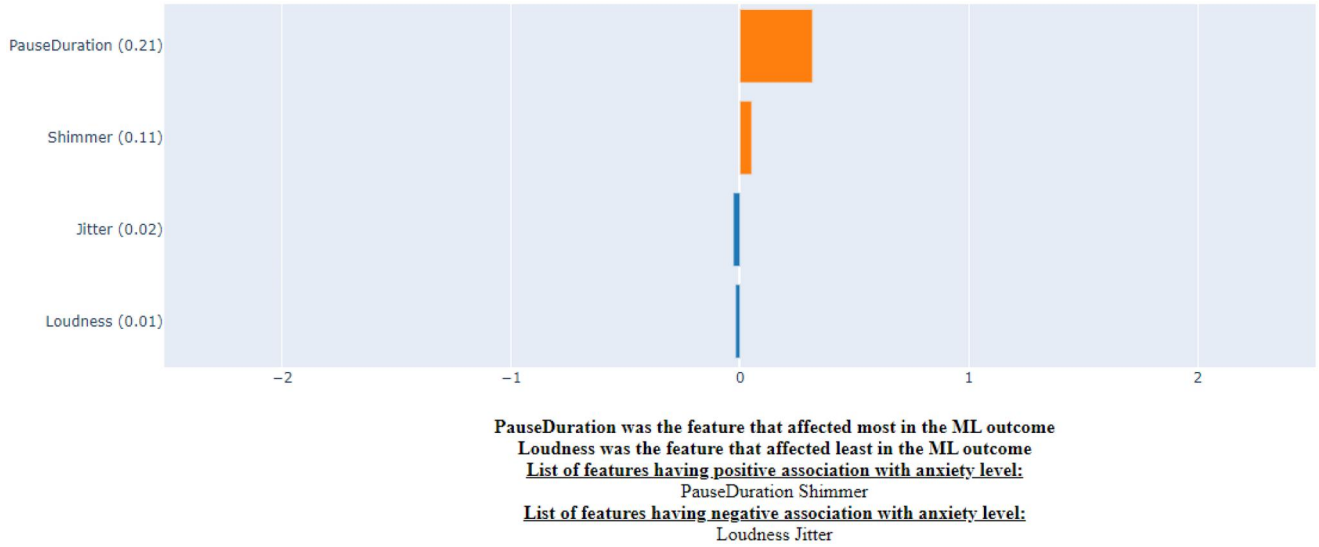


Figure 2. Example of a local explanation graph provided by the explainable boosting machine (EBM) that indicates the importance of each feature in making the final decision for a given sample. Larger absolute values indicate higher importance of the corresponding feature in estimating anxiety levels for the corresponding audio sample. Positive values contribute to the anxiety outcome, while the opposite holds for the negative values. A brief verbal explanation of the graph is also provided to the annotators.

and hold domain knowledge in life science interact with an explainable AI as part of a speech-based data analytics task, and also examine their trust in the AI model. Therefore, eligibility criteria for the human annotators were: (1) being older than 18 years; and (2) being enrolled as an undergraduate student in the department of Psychological & Brain Sciences at Texas A&M University (TAMU). The second criterion was included so that annotators are familiar with basic concepts related to human behavior and they can better perceive and interpret the anxiety in speech. The fact that annotators had not yet obtained their degree allowed us to recruit participants from various academic levels. Recruitment was conducted via bulk emails. Our study includes 13 annotators (10 female, 3 male; 19.84 (M) ± 1.23 (SD) years; 3 Asian, 6 White/Caucasian, 3 Hispanic/Latino, 1 Black/African American). From these annotators, two were freshmen, three were in their sophomore year, five were juniors, and three were seniors.

4.2. Study protocol

The overall workflow of the study protocol is summarized in Figure 4. First, participants completed two questionnaires that recorded individual differences. These included the Big Five Inventory (John & Srivastava, 1999) that captures personality traits and the Propensity to Trust Machines questionnaire (Merritt et al., 2013) reflecting one's general tendency to trust machines. The distribution of these scores for the annotators is shown in Table 1. Following that, the first author then conducted a one-to-one meeting with each participant, in which he explained the task, the EBM model, and the web interface, and answered their questions. The first author was also available throughout the duration of the experiment for any additional questions. After explaining the task and the interface to the annotators, the annotators were provided with access to the interface and were instructed to annotate the 82 files with the help of the AI model.

The annotation procedure is a cognitively demanding task, thus the annotators were told to annotate 8 batches of files spread across different points in time. Each batch included 10 files except from the last batch which included 12 files. Annotators were instructed to devote roughly two hours to each batch. For each file, the annotator was asked to listen to the corresponding audio, view the state of anxiety decision, local explanation graph, and global explanation graph of the EBM model, and provide their final decision in terms of the perceived anxiety level on a 5-point Likert scale (i.e., same scale as the ground truth; Section 3.1). Note that the annotators had been told that their final annotation score does not need to be aligned with the AI decision, and that they may agree or disagree with the AI. In addition, they were strongly advised to provide a comment for each audio file explaining their thought process and a concise reason on why they agree or disagree with AI decision. After completing each batch of files, annotators were further asked to rate the extent to which they trusted AI in making their decision on a 5-point Likert scale (i.e., 1: Not at all; 5: Extremely) which was used as the self-reported trust measure in our analysis. Since the self-reported trust measure was based on each batch of files, annotators were instructed to listen and annotate a batch of files at a time, which can potentially mitigate recall error. Each annotator was compensated with $180 at the end of the study.

5. Data analysis and results

Here, we describe the data analysis methodology and corresponding results. We first provide the definitions of the main variables considered in our analysis (Section 5.1). Then, we examine the reliability of collected data by investigating the discrepancy of each annotator in rating the

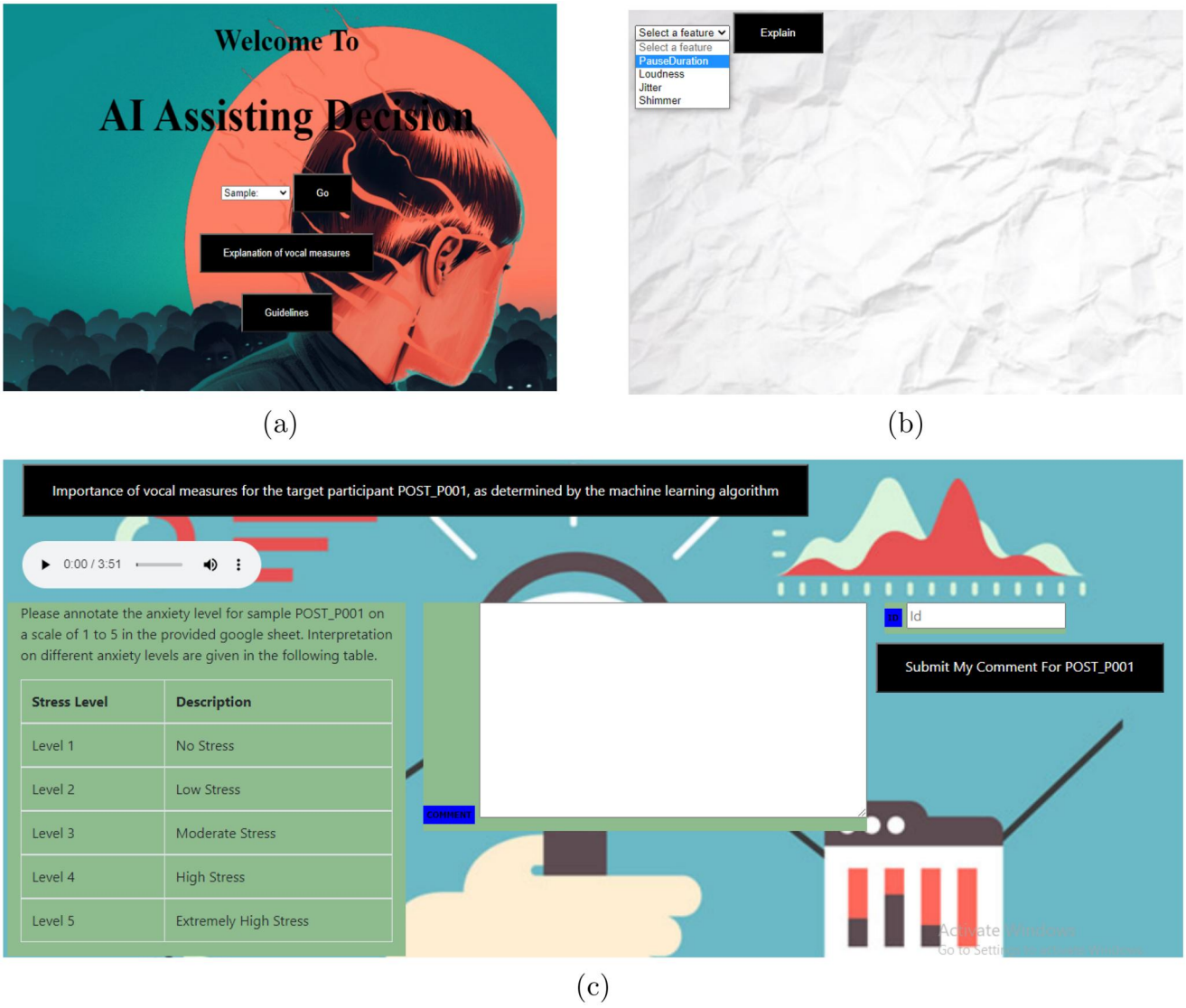


Figure 3. Custom-made web interface used by the annotators to interact with the AI system. (a) Home page, including links to the audio samples, explanations of the acoustic measures, and guidelines about the annotation process; (b) drop down list, containing the four features used as an input of the AI model. After selecting a feature, the user was able to inspect the global explanation graph for that feature; (c) Main web page through which the annotator can listen to the audio file, provide their rating (via a separate sheet), and add a comment explaining their rating (via the white text field). The annotator can also click on the button located at the top left of the page in order to see the anxiety score provided by the AI model and examine the local explanation graph (e.g., "importance of vocal measures for the target participant POST_P001, as determined by the machine learning algorithm").

duplicate samples (Section 5.2). Following that, we study the errors performed by the human annotators and the AI system and conduct a quantitative analysis to better understand cases in which the human and the AI partner might depict differences in performance (Section 5.3). Next, we examine the association between behavioral and self-reported measures of trust (Section 5.4), as well as the association between AI capacity and trust in AI to better understand the extent to which human annotators properly calibrate their trust in the capabilities of AI system (Section 5.5). We further explore how trust in AI evolves over time and in association to any errors conducted by the AI system (Section 5.6). Finally, we investigate individual factors of trust pertaining to overall trust propensity and personality characteristics and their interaction with trust evolving over time (Section 5.7).

5.1. Notation

In the following, we provide the definition of the basic variables that are considered in our analysis, along with the mathematical notation (Table 2). For the sake of consistency, variables referring to the sample level are denoted with small letters, while variables referring to the batch level (i.e., containing many samples) are denoted with capital letters. We define the human error $human\_error_{i,k}$ of annotator i at sample k as the absolute difference between the ground truth and the annotation, while the AI error ai_error_k at sample k is measured via the absolute difference between ground truth and the AI decision. The capacity ai_cap_k of the AI system at sample k is the inverse of the AI Error ai_error_k of sample k, i.e., $ai\_cap_k = \frac{1}{ai\_error_k}$. We further define $S_{i,j}$ as the self-reported trust of annotator i after batch

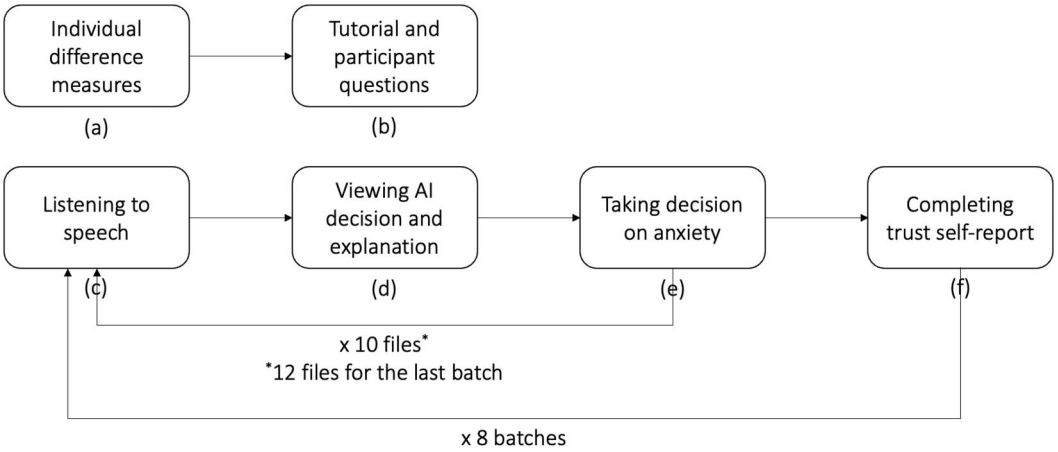


Figure 4. Schematic workflow of the study procedure. (a) Annotators completed the Big Five Inventory, the propensity to trust machines questionnaire, and a general survey about their prior research experience, which served as a proxy of annotator task expertise; (b) annotators were presented with a mini-tutorial about the goal of the experiment, description of acoustic features, and overview of the explainable boosting machine (EBM) model; (c) annotators listened to an audio file; (d) annotators viewed the explanation and decision of the EBM model for the audio file; (e) annotators rated the anxiety level of the audio file. (f) Annotators provided their self-reported trust level after rating each batch of samples. Steps (c-e) were repeated per sample.

Table 1. Distribution of annotator characteristics.

ID Range	Extroversion [8–40]	Agreeableness [9–45]	Conscientiousness [9–45]	Neuroticism [8–40]	Openness [10–50]	Propensity to trust machines [6–30]	General Propensity to trust [12–60]
P1	21	35	38	34	40	25	28
P2	17	35	31	33	34	12	37
P3	18	39	36	33	36	21	37
P4	14	30	34	33	41	28	45
P5	20	42	44	14	41	30	34
P6	19	35	29	24	30	22	43
P7	22	35	35	23	28	21	36
P8	27	34	37	17	36	26	26
P9	27	32	34	31	31	29	55
P10	35	39	35	24	41	15	35
P11	33	37	36	25	27	25	38
P12	35	36	42	26	40	28	30
P13	36	38	27	38	45	22	39
	23 ± 6.38	35.72 ± 3.22	35.36 ± 3.67	26.45 ± 6.58	35 ± 5.13	23.1 ± 5.4	37.63 ± 7.61

Table 2. Notation of variables considered in the analysis.

Variable	Notation
Human error of annotator i at sample k	$human\_error_{i,k}$
AI error at sample k	ai_error_k
AI capacity at sample k	ai_cap_k
Self-reported trust of annotator i at batch j	$S_{i,j}$
Behavioral trust of annotator i at batch j	$B_{i,j}$
Behavioral distrust of annotator i at batch j	$D_{i,j}$
Behavioral trust of annotator i at sample k	$b_{i,k}$

j has been completed, and $D_{i,j}$ as the behavioral distrust measure of annotator i for batch j, quantified as the average absolute discrepancy between the AI decision and the annotator's decision (Chu et al., 2020). Following that, the behavioral trust measure $B_{i,j}$ of annotator i for batch j is defined as the inverse of the behavioral distrust measure, i.e., $B_{i,j} = \frac{1}{D_{i,j}}$. Finally, the behavioral measure of trust $b_{i,k}$ of annotator i at sample k is defined as the inverse of the absolute discrepancy between the human annotator i and ML prediction for sample k.

5.2. Reliability of collected data

To evaluate the attention of the annotators during the decision making task, we compute the average difference in the annotation scores for the four duplicate samples that were provided in the data. This score is low (i.e., 0.424 (M) ± 0.277 (SD)) for the majority of the annotators, which indicates that overall annotators were attentive to the considered task. However, this error is relatively large (i.e., > 0.7) for Annotator 12 and Annotator 13, suggesting reduced attention by these two annotators to the decision-making task. Thus, the data from these annotators were excluded from the following analysis. Of note, we get a lower average difference in annotation scores for the duplicate samples after excluding these annotators' data (i.e., 0.338 (M) ± 0.195 (SD)).

5.3. Error analysis of human annotators and AI model

We first obtain an overall understanding of cases in which the human annotators and the AI system depict differences in performance. For this purpose, we create a scatter plot in which the x-axis represents the human error $human\_error_{i,k}$ of each annotator i at sample k, while the y-axis represents the AI error ai_error_k at sample k (Figure 5). We further identify the following four regions in this plot: (R_1): AI performs well and annotator performs well ($N = 527$); (R_2): AI performs well and annotator performs poorly ($N = 221$);

(R_3): AI performs poorly and annotator performs well ($N = 46$); (R_4): AI performs poorly and annotator performs poorly ($N = 64$). The boundaries of the above regions are defined by the average of the maximum and minimum values of the human error (i.e., for the x-axis) and the AI error (i.e., for the y-axis). It is evident that the human annotators and the AI system depict different performance over samples that belong to regions R_2 and R_3. Via a paired t-test, we further observe significant differences between the error distributions of the two partners in these regions (i.e., R_2: $t(220) = 29.2$, $p < 0.000$, $N = 221$; R_3: $t(45) = -12.53$, $p < 0.000$, $N = 46$), as depicted in Figure 6.

Via qualitative analysis, we further attempt to better understand these observed discrepancies between the human annotators and the AI system. To accomplish this, we coded the 204 comments provided by the annotators after reviewing each audio sample (Section 4.2) and assigned each comment to one of nine categories (Table 3). Since our objective is to study different capabilities between human and AI decision-making, we did not include any comments to these categories that entirely focus on AI features. Additionally, a comment may fall under more than one category. We compute the average AI prediction score and the average human annotation scores for the samples belonging to each category (Table 3). Via a paired t-test, we find significant differences between the AI prediction score and the human annotation score for each comment category (e.g., $p < 0.05$) except for the categories C_2, C_7, and C_8. Therefore, we did not consider these comment categories (e.g., C_2, C_7, and C_8) in the following analysis. Results indicate that human annotators tend to make fewer errors than the AI on average in cases in which they take into account the speech loudness (C_3). Although speech loudness is included in the features of the AI model, it was computed as the mean loudness over the whole utterance. In contrast, it appears annotators considered loudness in association to certain words, which was referred in the comments as "*speech emphasis.*" Besides, the annotators perform better than AI on average when they consider stuttering of the speaker (C_4). A potential reason is that stuttering is a reliable marker of anxiety (Blood & Blood, 2007; Ollendick & Hirshfeld-Becker, 2002) which was not considered by AI. In addition, human annotators depict reduced error compared to the AI system on average when they consider the natural pauses of the speaker (C_9). Even though the mean pause interval over the speech was included in the AI model, the AI model did not consider whether any pause was natural or the pause happened due to the nervousness of the speaker. The annotators depict these natural pauses not related to anxiety and provide lower anxiety score than AI and perform better than AI on average. On the contrary, the AI system makes less errors than the human annotators, when the latter take into account the speaker's general impression (C_1), speaking rate (C_5), as well as accent (C_6). A potential explanation of these

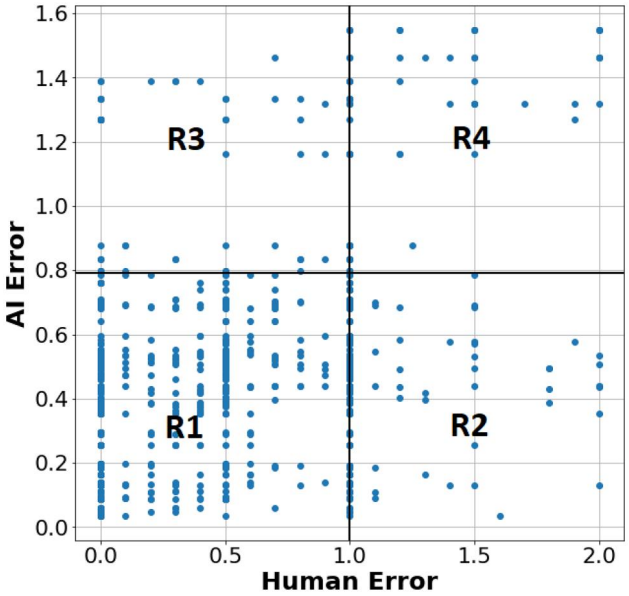
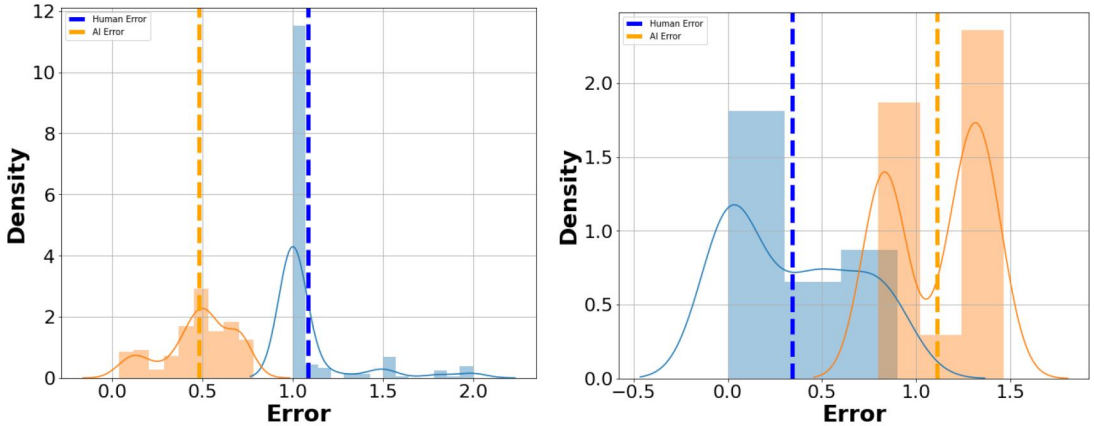


Figure 5. Visualizing the association between the errors conducted by the humans and the AI model, across four different regions. (R_1): AI performs well and annotator performs well; (R_2): AI performs well and annotator performs poorly; (R_3): AI performs poorly and annotator performs well; (R_4): AI performs poorly and annotator performs poorly.



(a) (b)

Figure 6. The distribution of human error and AI error for samples in which the human and the AI depict different performance. (a) $\mathbf{R_2}$: the AI performs well and the human annotator performs poorly; and (b) $\mathbf{R_3}$: the AI performs poorly and the human annotator performs well. Blue plots refer to human error and orange plots refer to AI error. The dotted vertical lines refer to the mean of these distributions.

Table 3. Coded categories of comments provided by the human annotators.

Comment type	Average human score	Average AI score	Paired t-test results between AI score and human score	Average ground truth	Example
C_1: General impression	2.1 ± 0.78	2.47 ± 0.23	$t(85) = -4.66,\ p < 0.01$	2.32 ± 0.75	Sounded pretty stressed. (P4)
C_2 : Filler words	2.63 ± 0.79	2.55 ± 0.22	$t(47) = 0.68, p = 0.50$	2.72 ± 0.57	[…] Lack of filler words signal little to no stress. (P5)
C_3 : Speech loudness	1.51 ± 0.38	2.49 ± 0.18	$t(8) = -9.65, p < 0.01$	1.77 ± 0.42	[…] only thing having a positive effect on anxiety score was loudness, but changes in loudness seemed to be to make her point/ emphasize things rather than a sign of nervousness. (P11)
C_4 : Stuttering/ stumbling	3.04 ± 0.54	2.52 ± 0.25	$t(43) = 6.88,\ p < 0.01$	2.86 ± 0.46	She was very nervous and you could hear her voice tremble. (P8)
C_5: Speaking rate	2.84 ± 0.61	2.49 ± 0.26	$t(19) = 2.14,\ p < 0.05$	2.60 ± 0.49	Subject spoke incredibly fast in some places which is indicative of anxiety. (P9)
C_6: Accent/Way of talking	2.16 ± 0.60	2.70 ± 0.29	$t(14) = -3.36,\ p < 0.01$	2.80 ± 0.54	I decided to go a little lower than the AI model because it sounds like this individual is struggling with choosing the right wording. (sounds like English is not the first language) so certain pauses might not be due to stress. I do not think the AI model accounts for these issues which is why I went a little lower when scoring. (P1)
C_7: Audio artifact/noise	2.32 ± 0.66	2.45 ± 0.31	$t(9) = -0.70,\ p = 0.50$	2.50 ± 0.50	[…] I think the microphone was just terrible for this speech. (P4)
C_8 : Perceived level of preparation	2.72 ± 0.97	2.81 ± 0.20	$t(5) = -0.24, p = 0.82$	2.67 ± 0.47	Speaker was unprepared and it showed […]. (P15)
C_9 : Natural pause	1.95 ± 0.58	2.51 ± 0.11	$t(5) = -2.63, p < 0.05$	2.17 ± 0.37	[. .] the pause duration felt very natural considering the speed at which they were speaking, so I put a lower anxiety score. (P9)

For each category, the average human annotation anxiety score, average AI prediction anxiety score, average ground truth anxiety score with standard deviation in $M \pm SD$ format, the paired t-test results between AI score and human anxiety score are reported, and examples of comments are provided.

findings might be that the general impression of a speaker can be quite subjective, therefore human annotators perform worse than the AI system when they take this factor into account. Also, the AI system can quantify speaking rate more objectively compared to the human annotator, therefore these features tend to be erroneously perceived by the human annotators. Human annotators depict more errors than the AI when they consider the speaker's accent, potentially due to the fact that this affect the voice quality, but is not necessarily associated with the state of anxiety.

5.4. Association between behavioral and self-reported measures of trust

We use a linear-mixed effects (LME) model with random intercept to find the association between self-reported trust measures and behavioral trust measures. The use of the LME model stems from its ability to handle complex and multifaceted data structures, particularly repeated measures and nested data (i.e., multiple trust measurements per annotator in our case). The LME models are an extension of linear regression that incorporates both fixed effects, representing population-level relationships, and random effects, capturing variability at the individual level. The LME model for this analysis is defined as follows:

$$S_{i,j} = \beta + a_1 \times B_{i,j} + x_i \tag{2}$$

Where $S_{i,j}$ is the self-reported trust of annotator i after batch j has been completed, and $B_{i,j}$ is the behavioral measure of trust of annotator i for batch j. In (2), a_1 serves as a fixed-effect coefficient, which is constant for all observations, and x_i serves as a random-effect coefficient, which is different for each participant i. The random effect coefficient x_i incorporates the random variability in trust from person to person. The coefficient a_1 quantifies the association between self-reported and behavioral trust measures. The results suggest significant positive association between the two (i.e., $a_1 = 1.56,\ p = 0.002,\ N = 88$), a finding which is also supported by prior studies (Sharan & Romano, 2020). Thus, in the following analysis, we will be using these two measures interchangeably.

5.5. Association between AI capacity and trust in AI

In this section, we examine the extent to which human annotators can calibrate their trust in AI according to the capacity of the AI system. Drawing from Lee and See (2004), we inspect the 2D scatter plots of AI capacity ai_cap_k against the behavioral trust $b_{i,k}$ of annotator i at sample k for all samples k per annotator (Figure 7). Calibration pertains to the alignment between an individual's trust in automation and the actual capabilities of the automation. Therefore, when the annotator's trust is fully calibrated with the AI capacity, the points of the 2D scatter plots should follow the identity line (i.e., $y = x$). Lee and See (2004) also represented good calibration of trust by the diagonal line in trust vs automation capability plot, where the level of trust matches automation capabilities. On the contrary, annotators over-trust the AI for the samples lying above the identity line, and under-trust the AI for the samples lying below the identity line. The larger the distance of a sample from the identity line, the higher is the degree of

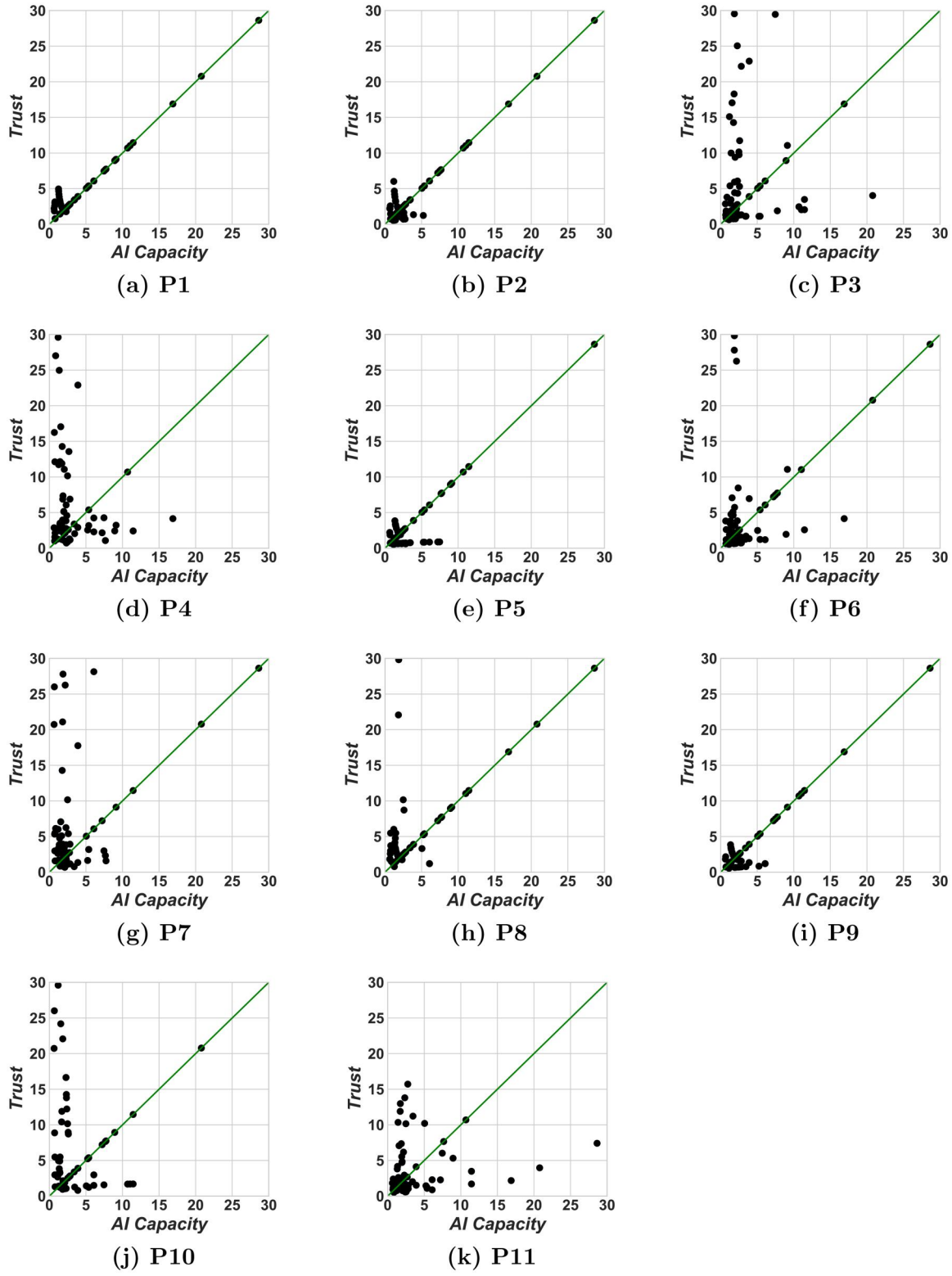
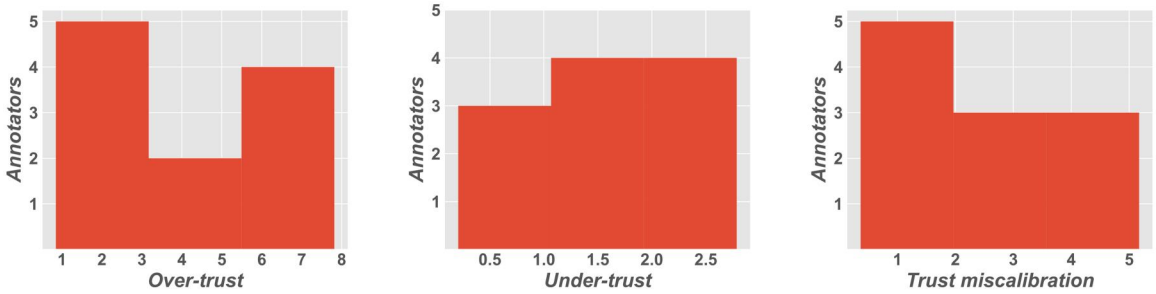


Figure 7. 2D Scatter plots of AI capacity against the behavioral trust for each annotator. Each point of a plot represents one audio sample. The x-axis represents the AI capacity calculated as the inverse of AI error and the y-axis represents the behavioral trust of the annotator for the specific samples.

over-trust or under-trust in AI, depending on whether the point is above or below the line, respectively. We employ the degree of under/overtrust rather than the frequency, since the first gives emphasis to the intensity or magnitude of the focal phenomenon and is more sensitive to changes. Based on this rationale, we define three metrics of trust per participant. *Over-trust* is quantified as the average absolute distance of samples above the identity line. *Under-trust* is measured as the average absolute distance of all points below the identity line. *Trust miscalibration* is finally defined as the average absolute distance of all the points from the identity line, which is close to zero in the ideal case where annotators' trust is fully calibrated with the capacity of the AI system.



(a) Over-trust (b) Under-trust (c) Trust miscalibration

Figure 8. Distribution of over-trust, under-trust, and trust miscalibration metrics for all annotators.

The average of these measures across all samples is computed for each annotator and their corresponding distribution is provided in Figure 8. These measures vary across annotators with some depicting relatively appropriate trust calibration to the capacity of the AI system (e.g., P1, P9; Figure 7; i.e., lower *trust miscalibration* metric), others over-trusting the AI system (e.g., P4, P7; Figure 7; i.e., higher *over-trust* metric), and others under-trusting the AI system (e.g., P11; Figure 7; i.e., higher *under-trust* metric). The average over-trust value per user (i.e., 4.12 (M) ± 2.73 (SD)) is higher than the average under-trust value per user (i.e., 1.59 (M) ± 0.73 (SD)) and the paired-t test results between average over-trust value per user and average under-trust value per user is statistically significant (i.e., $t(10) = 3.42$, $p < 0.01$). Therefore, human annotators mostly over-trust the AI system, which is consistent with prior findings suggesting that having a favorable experience with automation leads to confidence that extends beyond the system capabilities (Ullrich et al., 2021).

5.6. Evolution of trust over time

We further explore how annotators change their trust in AI over time, depending on the error of the AI system. We build a LME model with random intercept that estimates the behavioral measure of trust in AI as a function of time and AI error, as follows:

$$b_{i,k} = \beta + a_2 \times ai\_error_k + b_2 \times k + x_i \tag{3}$$

Where $b_{i,k}$ denotes the behavioral measure of trust in AI of annotator i for sample k and ai_error_k denotes the AI error for the kth sample. In (3), a_2 and b_2 are the fixed-effect coefficients, which are constant for all observations, and x_i is a random-effect coefficient, which is different for each participant i. When estimating the coefficients of the LME model in (3), we exclude the outlier samples for which the AI system is highly accurate, but the annotators depict very low trust in the system, and vice-versa, the samples for which the system performs very poorly, but the annotators depict overly increased trust. For this purpose, for each annotator i, we exclude the outlier samples for which the ratio of behavioral trust $b_{i,j}$ to AI capacity ai_cap_k is too low or too highi.e., $\frac{b_{i,k}}{ai\_cap_k}$ is either lower than the 2.5% quantile or larger than the 97.5% quantile value of all samples from the corresponding annotator. This resulted in excluding 66 samples out of the total 902 samples (i.e., 7.3%) from the LME model. Results indicate that annotators increase their trust in AI over time (i.e., $b_2 = 6.92$, $p = 0.009$, $N = 836$), but momentarily decrease their trust in AI when the AI system makes error (i.e., $a_2 = -12.184$, $p = 0.001$, $N = 836$). The observed overall increase of annotators' trust in AI over time is also reflected in their comments. For instance, P1 mentioned, "*I was not sure how I felt about this one, so I relied heavily on the opinion of the AI model. This is also where I am realizing that I am trusting the opinion of the model much much more compared to when I started.*". Similarly, P10 commented that, "*As stated before, as I continue to use the AI's observations and review them after I have reviewed the audio files, I see that the AI is able to make relatively accurate observations in my opinion.*". There was no comment from the annotators which indicates a momentary decline of trust in AI when the latter makes an error. However, the discovered negative association between AI error and human trust in AI is also consistent with prior work regarding trust in automation (Hancock et al., 2011).

5.7. Effect of personality characteristics and prior research experience on human trust in AI

Grounded in prior work that has found significant personality effects on trust in automation (Böckle et al., 2021; Müller et al., 2019), we investigate the extent to which trust in the AI system depends on the characteristics of the annotator, such as their overall propensity to trust machines, their personality, and their prior research experience. We build the following LME model with random intercept to analyze the effect of annotators' characteristics on self-reported trust in AI over time:

$$\begin{aligned} S_{i,j} = \beta + a_3 \times j + b_3 \times D_{i,j} + c_3 \times M_i + d_3(j \times M_i) \\ + e_3 \times A_i + f_3(j \times A_i) + g_3 \times C_i \\ + h_3(j \times Ci) + i_3 \times E_i + j_3(j \times E_i) + k_3 \times R_i + l_3(j \times Ri) + x_i \end{aligned} \tag{4}$$

In (4), $S_{i,j}$ denotes the self reported trust of annotator i for batch j, $D_{i,j}$ is the average behavioral distrust of annotator i for batch j, M_i is annotator's i overall propensity to trust machines, A_i is annotator's i agreeableness, C_i is annotator's i conscientiousness, E_i is annotator's i extroversion, and R_i is annotator's i research experience encoded as a Boolean variable (i.e., 0 or 1 for the absence or presence of

Table 4. Linear Mixed Effect (LME) model estimates of fixed and interaction effects of individual factors and time for estimating self-reported trust in AI.

Individual factors	Fixed-effect	Interaction effect
Time j	$a_3 = 0.74$	
Behavioral distrust $D_{i,j}$	$b_3 = -2.04^{*}$	
Propensity to trust machines $M_{i,j}$	$c_3 = 2.20^{\dagger}$	$d_3 = -4.07^{*}$
Agreeableness $A_{i,j}$	$e_3 = 2.71^{\dagger}$	$f_3 = -4.76^{*}$
Consciousness $C_{i,j}$	$g_3 = -2.72^{\dagger}$	$h_3 = 6.64^{*}$
Extroversion $E_{i,j}$	$i_3 = -0.41$	$j_3 = 1.37^{\dagger}$
Prior research experience $R_{i,j}$	$k_3 = -0.26$	$l_3 = 0.86$

$^{*},^{\dagger}: p < 0.05, p < 0.1$

prior experience with behavioral coding, respectively). In the same equation, a_3, b_3, c_3, d_3, e_3, f_3, g_3, h_3, i_3, j_3, k_3, l_3 are fixed-effect coefficients which remain constant for all observations, and x_i is a random-effect coefficient which is different for each participant i. The model also considers the interaction between each of the annotator characteristics and time, so that we can understand whether the evolution of trust over time varies between different people. Results are provided in Table 4 ($N = 88$), where it is important to note that the threshold of $p < 0.1$ is not employed as a criterion for statistical significance but is utilized as an indicator that the corresponding coefficients may be approaching statistical significance given the constraints of the small sample size. As expected, similar to findings in Section 5.4, results indicate a negative association between self-reported trust and behavioral distrust ($b_3 = -2.04^{*}$), or else, a positive association between self-reported and behavioral trust. Agreeable annotators depict increased trust in the AI ($e_3 = 2.71^{\dagger}$), although this association does not reach statistical significance. Agreeable annotators' trust decreases over time ($f_3 = -4.76^{*}$), potentially due to the fact that these annotators initially start with higher levels of trust compared to their counterparts. Similar findings hold for annotators who have inherently high propensity to trust machines ($c_3 = 2.20^{\dagger}, d_3 = -4.07^{*}$). On the contrary, conscientious annotators depict overall lower trust to AI compared to their counterparts ($g_3 = -2.72^{\dagger}$), an association which is not statistically significant, but their trust increases over time ($h_3 = 6.64^{*}$). Extroversion and prior research experience do not appear to significantly affect trust in AI.

6. Discussion

This paper examines human trust in AI in a collaborative data analytics task, in which humans and AI worked together to estimate a speaker's levels of anxiety from speech. We have explored five research questions via the conducted analysis. This section includes a summary of the research questions and corresponding findings, as well as a discussion of their implications.

In response to *RQ1*, we observe differences in performance between the human annotators and the AI system. Human annotators achieve better performance than the AI when they consider speech speaker stuttering and the speaker's emphasis on certain words (Section 5.3), both of which are not explicitly modeled in the AI system. In addition, human annotators do better than the AI when they take into account the natural pauses of the speaker, a feature which is difficult to capture via the speech signal, since it requires additional context information. On the contrary, the AI performs better than the human annotators in cases when the latter considered the speaker's accent and speaking rate as an indicator of anxiety. It further appears that the annotators perform worse than the AI when they take into account the general impression of the speaker for their decision which is quite subjective. Human annotators' superior performance in aspects such as recognizing speech speaker stuttering, emphasis on certain words, and natural pauses, reflects the unique ability of human cognition to consider context-dependent features that are not explicitly modeled in the AI model highlighting the potential of leveraging human expertise in collaborative tasks where such nuanced understanding is crucial. However, the AI's better performance in cases where human annotators relied on indicators like the speaker's accent and speaking rate suggests that the AI system might be able to grasp dimensions that can be subjective and challenging for human annotators to assess. This suggests the complementary nature of human and AI capabilities, emphasizing the potential for effective collaboration where each entity contributes its strengths. Implications of these findings into future research could encompass the development of frameworks that offer explicit guidelines or training for human annotators, aiming to address subjective elements in the decision-making process. Additionally, future work can involve the design of systems that facilitate effective communication and the seamless integration of human insights with machine capabilities. This includes the creation of interfaces that enable humans to comprehend and interpret AI decisions more easily, while also allowing AI systems to better understand human capabilities by incorporating human-like perceptual abilities in a personalized manner.

In exploring *RQ2*, we found significant positive correlation between self-reported and behavioral trust in AI (Sections 5.4 and 5.7). This aligns with similar findings in prior studies, such as the work by Sharan and Romano (2020), which also identified a moderate positive correlation between self-reported and behavioral measures of trust, the latter captured via reaction time and user agreement with the AI. It further suggests that behavioral measures can serve as a viable proxy for assessing user trust in AI, which can have important implications in the design of adaptive AI systems. Behavioral measures provide a continuous stream of data, allowing AI systems to adapt in real-time. For instance, if a sudden decrease in user trust is detected through behavioral indicators, the AI system can adjust its behavior or provide additional explanations to repair trust. This can further contribute to tailoring the system's responses to each user. However, the fact that these results are found in a small sample size, combined with previous evidence that suggests an incongruence between behavioral and self-reported trust (Kulms & Kopp, 2019), also underscore the importance of further exploration into additional measures of trust that potentially can be more objectively quantified via brain activity, speech, and language (Chen

et al., 2020; de Visser et al., 2018; Dong et al., 2015; Levitan et al., 2015).

In addressing *RQ3*, our results indicate that annotators depict different levels of trust calibration with respect to the capacity of the AI system. While some annotators demonstrate effective trust calibration based on the AI's proficiency, others exhibit trust miscalibrion (Section 5.5). The predominant pattern in trust miscalibration involves overtrust in the AI. One possible explanation for this finding could be that, despite possessing domain knowledge in life science, the human annotators, being undergraduate students with limited practical experience in the field, may lack a high level of confidence in their abilities. Prior studies also show that when individuals are unable to rely on their own judgment, reliance in automation is especially evident (Fan et al., 2008; Sanchez et al., 2014). These findings underscore the importance of addressing overtrust tendencies among annotators, especially when considering the practical deployment of AI in collaborative decision-making tasks within specific sensitive domains. Strategies to enhance annotators' awareness about the limitations and capabilities of the AI system, coupled with effective training on when to rely on AI predictions and when to exercise human judgment, could contribute to more balanced trust calibration (Aroyo et al., 2021). From a system design perspective, incorporating feedback mechanisms that highlight uncertainties and potential pitfalls in the AI's decision-making process can be also beneficial (Buçinca et al., 2021).

In answering *RQ4*, trust in AI overall increases over time, but momentarily decreases when the AI makes more errors (Section 5.6). Since the AI system depicts moderate performance, this increasing trend of trust in AI over time could be explained by the fact that human annotators might project their potentially positive initial experience with the AI to the audio samples that are being observed at the latter batches of the data collection. However, it appears that annotators can also momentarily differentiate between successful situations and instances in which the AI makes an error, which is inline with prior work on human trust in robotic errors (Geiskkovitch et al., 2019; Ragni et al., 2016). This observed pattern is noteworthy in the context of trust in studies that involve human-AI collaboration. The possibility of more uncertainty in human-AI decision-making, compared to human-robot interaction, could have suggested that AI errors might irreversibly impact trust. However, this study's results suggest otherwise. Further exploration of this finding, especially in studies involving subject matter experts with substantial experience in behavioral coding tasks, could offer valuable insights into the generalizability of these observations.

Finally, in response to *RQ5*, we found that overall propensity to trust machines, agreeableness, and consciousness affect human trust in AI, while extroversion and prior research experience did not emerge as significant factors in this context. However, these results only approach statistical significance in our study. The majority of these findings coincide with prior work on human trust in general automation (e.g., robotic agents), while it's crucial to emphasize that this study contributes novel insights by providing preliminary evidence on how these individual factors distinctly influence human trust in AI systems. People with high propensity to trust depict high trust, which decreases more over time compared to their counterparts (Section 5.7). This is in accordance to prior work that indicates that people with high propensity to trust depict higher initial trust in automation, which decreases in the presence of an error (Ebert et al., 2009; Madhavan et al., 2006; Merritt & Ilgen, 2008). Our results suggest similar trends for agreeable people with approaching statistical significance. Also prior work has found that agreeable people hold high initial trust (Chien et al., 2016) and high overall trust (Bawack et al., 2021; Kraus et al., 2020) in the AI systems. On the contrary, conscientious annotators depicted less trust in AI than their counterparts, although this result did not reach statistical significance, with increasing trends over time (Section 5.7). Previous results regarding the effect of the conscientiousness in trust in automation are inconclusive. Some studies do not suggest a correlation between conscientiousness and trust in AI (Kraus et al., 2020), while others demonstrate positive correlation between the two (Bawack et al., 2021) irrespective of time. Other studies indicate that people with high conscientiousness have higher initial trust in AI (Chien et al., 2016). Collecting data from annotators with more extreme conscientiousness characteristics or explicitly manipulating the experimental conditions in terms of the quality and quantity of explanation provided by the AI system might help better answering this question. Our study does not indicate a significant association between extroversion and trust in AI, which is in line with prior work that did not find significant correlation between trust in AI and extroversion (Chien et al., 2016). Finally, prior research experience does not appear to be a moderating factor of trust in our experiment. Prior work indicates that user expertise is loosely related to dimensions of trust in automation (K. E. Schaefer et al., 2014). For example, users with limited task expertise tend to over-trust automated systems (Nourani et al., 2020). Despite our participants possessing knowledge in life science, their limited prior research experience in behavioral annotation (i.e., only 3 out of 11 participants had prior experience) may explain the lack of significance in the research experience variable. The minimal variability in this variable could be a potential contributing factor to its non-significant role in influencing trust. The implications of these findings for system design carry significance for creating AI systems that effectively engage and interact with human users. For example, designing interfaces that cater to individuals with high propensity to trust and high conscientiousness might involve incorporating elements that foster a sense of reliability and transparency. For agreeable individuals, interfaces can prioritize user-friendly features and clear communication to maintain their trust. Finally, recognizing the decreasing trend of trust over time for individuals with high propensity to trust suggests the need for dynamic adaptation in system behavior. This could involve periodic reinforcement of reliability or adjusting the level of explanation to sustain trust.

Despite the encouraging results, our study presents the following limitations. First, this paper relies on data from users with direct background from life science, which is inherently more difficult to acquire compared to crowdsourced data. For this reason, our analysis contains data from a small number of participants, which raises considerations about the generalizability and statistical power of the results. With a limited number of participants, the findings may be more susceptible to the influence of individual variations and outliers, potentially affecting the overall generalizability of the study. Furthermore, the limited diversity within the small sample, as demonstrated via the low variation in terms of personality characteristics and prior research experience, may impact the external validity of the results. Collecting data from a larger participant pool might allow us to better explain the impact of personality and prior research experience on the user trust and trust calibration. Second, we measured self-reported trust via a one-item measure, which may oversimplify the complexity of the focal psychological construct raising reliability concerns. While acknowledging this as a limitation in our study, it is important to note that participants were asked to self-report their trust eight times throughout the study protocol, since our research aimed to capture the evolving nature of user trust over time. Given the study design, employing a validated questionnaire with multiple items might have introduced user fatigue and potentially heightened subjectivity in the measurement. As part of our future work, we will supplement the existing self-reported and behavioral measures of trust with neural measures that have been empirically validated in human-automation settings (de Visser et al., 2018). Finally, as part of the study protocol, we did not explicitly control for the performance of the AI system (e.g., via manipulating this variable), which would have allowed us to better understand its effect on human trust.

7. Conclusion

In conclusion, this study delved into the dynamics of human trust in AI within a human-AI collaborative task that focused on estimating anxiety levels from speech. An explainable AI system, the EBM model, interacted with human annotators with background on psychological sciences and provided explanations about local and global feature importance, along with the AI decision. Trust in AI was captured via self-reports and behavioral measures. Human trust in the AI system increases over time with errors conducted by the AI being associated with momentary decrease in user trust. The study revealed nuanced differences in performance between human and AI partners, influenced by the characteristics of the cases under consideration. The findings from the study further underscore the importance of proper trust calibration, highlighting individual variations in overall trust in AI, where factors such as general propensity to trust, agreeableness, and conscientiousness emerged as influential determinants of trust in this collaborative human-AI setting. Overall, this work contributes to deepening our knowledge for the development and deployment of trustworthy AI applications in real-world collaborative scenarios. Moving forward, future work in this domain could explore additional techniques for enhancing the collaboration between humans and AI in tasks related to human state detection. This may involve refining the XAI system to further improve the interpretability and clarity of its explanations. Additionally, investigating the impact of different types of errors on user trust and developing strategies for effective error mitigation could be a valuable avenue. Incorporating professionals in psychology as study participants offers the opportunity to broaden the research to diverse user populations, enhancing the practical relevance of the findings. Further investigation into neural measures could be pursued as a promising avenue for quantifying human trust at a moment-to-moment level.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Science Foundation [CAREER: Enabling Trustworthy Speech Technologies for Mental Health Care: From Speech Anonymization to Fair Human-centered Machine Intelligence, #2046118, PI: Chaspari] and the Air Force Office of Scientific Research [Trust & Influence Program, #FA9550-22-1-0010, PI: Chaspari].

References

Ahmed, A. M., & Salas, O. (2009). The relationship between behavioral and attitudinal trust: A cross-cultural study. *Review of Social Economy*, *67*(4), 457–482. https://doi.org/10.1080/00346760902908625

Alam, L., & Mueller, S. (2021). Examining the effect of explanation on satisfaction and trust in ai diagnostic systems. *BMC Medical Informatics and Decision Making*, *21*(1), 178. https://doi.org/10.1186/s12911-021-01542-6

Alarcon, G. M., Lyons, J. B., Christensen, J. C., Klosterman, S. L., Bowers, M. A., Ryan, T. J., Jessup, S. A., & Wynne, K. T. (2018). The effect of propensity to trust and perceptions of trustworthiness on trust behaviors in dyads. *Behavior Research Methods*, *50*(5), 1906–1920. https://doi.org/10.3758/s13428-017-0959-6

Aroyo, A. M., de Bruyne, J., Dheu, O., Fosch-Villaronga, E., Gudkov, A., Hoch, H., Jones, S., Lutz, C., Sætra, H., Solberg, M., … Tamò-Larrieux, A. others (2021). Overtrusting robots: Setting a research agenda to mitigate overtrust in automation. *Paladyn, Journal of Behavioral Robotics*, *12*(1), 423–436. https://doi.org/10.1515/pjbr-2021-0029

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., & Müller, K.-R. (2010). How to explain individual classification decisions. *The Journal of Machine Learning Research*, *11*, 1803–1831. https://dl.acm.org/doi/epdf/10.5555/1756006.1859912

Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., & Horvitz, E. (2019). Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 2429–2437. https://doi.org/10.1609/aaai.v33i01.33012429

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI conference On Human Factors in*

Computing Systems (pp. 1–16). ACM. https://doi.org/10.1145/3411764.3445717

Batrinca, L., Stratou, G., Shapiro, A., Morency, L.-P., & Scherer, S. (2013). Cicero-towards a multimodal virtual audience platform for public speaking training. In *International Workshop on Intelligent Virtual Agents* (pp. 116–128). Springer.

Bawack, R. E., Wamba, S. F., & Carillo, K. D. A. (2021). Exploring the role of personality, trust, and privacy in customer experience performance during voice shopping: Evidence from SEM and fuzzy set qualitative comparative analysis. *International Journal of Information Management*, *58*, 102309. https://doi.org/10.1016/j.ijinfomgt.2021.102309

Blood, G. W., & Blood, I. M. (2007). Preliminary study of self-reported experience of physical aggression and bullying of boys who stutter: Relation to increased anxiety. *Perceptual and Motor Skills*, *104*(3 Pt 2), 1060–1066. https://doi.org/10.2466/pms.104.4.1060-1066

Böckle, M., Yeboah-Antwi, K., & Kouris, I. (2021). Can you trust the black box? The effect of personality traits on trust in AI-enabled user interfaces. In *International Conference on Human-Computer Interaction* (pp. 3–20). Springer.

Booth, B. M., Vrzakova, H., Mattingly, S. M., Martinez, G. J., Faust, L., & D'Mello, S. K. (2022). Toward robust stress prediction in the age of wearables: Modeling perceived stress in a longitudinal study with information workers. *IEEE Transactions on Affective Computing*, *13*(4), 2201–2217. https://doi.org/10.1109/TAFFC.2022.3188006

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–21. https://doi.org/10.1145/3449287

Chen, X. L., Ita Levitan, S., Levine, M., Mandic, M., & Hirschberg, J. (2020). Acoustic-prosodic and lexical cues to deception and trust: Deciphering how people detect lies. *Transactions of the Association for Computational Linguistics*, *8*, 199–214. https://doi.org/10.1162/tacl_a_00311

Cheng, H.-F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., & Zhu, H. (2019). Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 Chi Conference on Human Factors in Computing Systems* (pp. 1–12). ACM.

Chien, S.-Y., Sycara, K., Liu, J.-S., & Kumru, A. (2016). Relation between trust attitudes toward automation, hofstede's cultural dimensions, and big five personality traits. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), 841–845. https://doi.org/10.1177/1541931213601192

Chollet, M., Wörtwein, T., Morency, L.-P., Scherer, S. (2016). A multimodal corpus for the assessment of public speaking ability and anxiety. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 488–495). ELRA.

Chu, E., Roy, D., & Andreas, J. (2020). Are visual explanations useful? A case study in model-in-the-loop prediction. *arXiv preprint arXiv: 2007.12248*

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., Hormozdiari, F., Houlsby, N., Hou, S., Jerfel, G., Karthikesalingam, A., Lucic, M., Ma, Y., McLean, C., Mincu, D., Mitani, A., … Beutel, A. (2022). Underspecification presents challenges for credibility in modern machine learning. *The Journal of Machine Learning Research*, *23*(1), 10237–10297.

de Brito Duarte, R., Correia, F., Arriaga, P., & Paiva, A. (2023). AI trust: Can explainable AI enhance warranted trust? *Human Behavior and Emerging Technologies*, *2023*, 1–12. https://doi.org/10.1155/2023/4637678

de Visser, 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*, *12*, 309. https://doi.org/10.3389/fnhum.2018.00309

Dong, S.-Y., Kim, B.-K., Lee, K., & Lee, S.-Y. (2015). A preliminary study on human trust measurements by EEG for human-machine interactions. In *Proceedings of the* 3rd International Conference on Human-Agent Interaction (pp. 265–268). ACM. https://doi.org/10.1145/2814940.2814993

Drnec, K., Marathe, A. R., Lukos, J. R., & Metcalfe, J. S. (2016). From trust in automation to decision neuroscience: Applying cognitive neuroscience methods to understand and improve interaction decisions involved in human automation interaction. *Frontiers in Human Neuroscience*, *10*, 290. https://doi.org/10.3389/fnhum.2016.00290

Ebert, I. D., Steffens, M. C., Von Stülpnagel, R., & Jelenec, P. (2009). How to like yourself better, or chocolate less: Changing implicit attitudes with one IAT task. *Journal of Experimental Social Psychology*, *45*(5), 1098–1104. https://doi.org/10.1016/j.jesp.2009.06.008

Ezer, N., Bruni, S., Cai, Y., Hepenstal, S. J., Miller, C. A., & Schmorrow, D. D. (2019). Trust engineering for human-AI teams. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *63*(1), 322–326. https://doi.org/10.1177/1071181319631264

Fan, X., Oh, S., McNeese, M., Yen, J., Cuevas, H., Strater, L., & Endsley, M. R. (2008). *The influence of agent reliability on trust in human-agent collaboration*. In *Proceedings of the* 15th *European Conference on Cognitive Ergonomics: The Ergonomics of Cool Interaction* (pp. 1–8). ACM. https://doi.org/10.1145/1473018.1473028

Geiskkovitch, D. Y., Thiessen, R., Young, J. E., & Glenwright, M. R. (2019). What? That's not a chair!: How robot informational errors affect children's trust towards robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 48–56). ACM. https://doi.org/10.1109/HRI.2019.8673024

Gillespie, N., Lockey, S., Curtis, C., Pool, J., & Akbari, A. (2023). *Trust in artificial intelligence: A global study*. The University of Queensland and KPMG Australia. https://doi.org/10.14264/00d3c94

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(Mar), 1157–1182.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527. https://doi.org/10.1177/0018720811417254

Hauke, J., & Kossowski, T. (2011). Comparison of values of Pearson's and Spearman's correlation coefficient on the same sets of data. *Quaestiones Geographicae*, *30*(2), 87–93. https://doi.org/10.2478/v10117-011-0021-1

Hemmer, P., Schemmer, M., Vössing, M., & Kühl, N. (2021). Human-ai complementarity in hybrid intelligence systems: A structured literature review. *PACIS*, *78*.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons*, *61*(4), 577–586. https://doi.org/10.1016/j.bushor.2018.03.007

Jiang, X., & Pell, M. D. (2017). The sound of confidence and doubt. *Speech Communication*, *88*, 106–126. https://doi.org/10.1016/j.specom.2017.01.011

John, O. P., & Srivastava, S. (1999). *The big-five trait taxonomy: History, measurement, and theoretical perspectives* (Vol. 2). University of California Berkeley.

Kaindl, H., & Svetinovic, D. (2019). Avoiding undertrust and overtrust. *Refsq Workshops*.

Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. (2023). Trust in artificial intelligence: Meta-analytic findings. *Human Factors*, *65*(2), 337–359. https://doi.org/10.1177/00187208211013988

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning* (pp. 2668–2677). ACM.

Kraus, J., Scholz, D., & Baumann, M. (2020). What's driving me? Exploration and validation of a hierarchical personality model for trust in automated driving. *Human Factors*, *63*(6), 1076–1105. https://doi.org/10.1177/0018720820922653

Kulms, P., & Kopp, S. (2019). More human-likeness, more trust? The effect of anthropomorphism on self-reported and behavioral trust in continued and interdependent human-agent cooperation. In *Proceedings of mensch und computer 2019* (pp. 31–42). ACM.

Lai, V., & Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 29–38). ACM.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Levitan, S. I., An, G., Wang, M., Mendels, G., Hirschberg, J., Levine, M., & Rosenberg, A. (2015). Cross-cultural production and detection of deception from speech. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection* (pp. 1–8). ACM. https://doi.org/10.1145/2823465.2823468

Lou, Y., Caruana, R., & Gehrke, J. (2012). Intelligible models for classification and regression. In *Proceedings of the 18th ACM Sigkdd International Conference on Knowledge Discovery and Data Mining* (pp. 150–158). ACM. https://doi.org/10.1145/2339530.2339556

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4768–4777.

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors*, *48*(2), 241–256. https://doi.org/10.1518/001872006777724408

Merritt, S. M. (2011). Affective processes in human–automation interactions. *Human Factors*, *53*(4), 356–370. https://doi.org/10.1177/0018720811411912

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors*, *55*(3), 520–534. https://doi.org/10.1177/0018720812465081

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors*, *50*(2), 194–210. https://doi.org/10.1518/001872008X288574

Monarth, H., & Kase, L. (2007). *Confident speaker*. McGraw-Hill Professional Publishing.

Müller, L., Mattke, J., Maier, C., Weitzel, T., & Graser, H. (2019). Chatbot acceptance: A latent profile analysis on individuals' trust in conversational agents. In *Proceedings of the 2019 on Computers and People Research Conference* (pp. 35–42). ACM.

Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2021). Explainable recommendation: When design meets trust calibration. *World Wide Web*, *24*(5), 1857–1884. https://doi.org/10.1007/s11280-021-00916-0

Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2023). How the different explanation classes impact trust calibration: The case of clinical decision support systems. *International Journal of Human-Computer Studies*, *169*, 102941. https://doi.org/10.1016/j.ijhcs.2022.102941

Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*

Nourani, M., King, J., & Ragan, E. (2020). The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *8*(1), 112–121. https://doi.org/10.1609/hcomp.v8i1.7469

Okamura, K., & Yamada, S. (2018). Adaptive trust calibration for supervised autonomous vehicles. In *Adjunct proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 92–97). ACM. https://doi.org/10.1145/3239092.3265948

Ollendick, T. H., & Hirshfeld-Becker, D. R. (2002). The developmental psychopathology of social anxiety disorder. *Biological Psychiatry*, *51*(1), 44–58. https://doi.org/10.1016/s0006-3223(01)01305-1

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Payne, B. K., Govorun, O., & Arbuckle, N. L. (2008). Automatic attitudes and alcohol: Does implicit liking predict drinking? *Cognition & Emotion*, *22*(2), 238–271. https://doi.org/10.1080/02699930701357394

Phillips-Wren, G. (2012). Ai tools in decision making support systems: A review. *International Journal on Artificial Intelligence Tools*, *21*(02), 1240005. https://doi.org/10.1142/S0218213012400052

Pop, V. L., Shrewsbury, A., & Durso, F. T. (2015). Individual differences in the calibration of trust in automation. *Human Factors*, *57*(4), 545–556. https://doi.org/10.1177/0018720814564422

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. In *Proceedings of the 2021 Chi Conference on Human Factors in Computing Systems* (pp. 1–52). ACM. https://doi.org/10.1145/3411764.3445315

Ragni, M., Rudenko, A., Kuhnert, B., & Arras, K. O. (2016). *Errare humanum est: Erroneous robots in human-robot interaction*. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 501–506). IEEE. https://doi.org/10.1109/ROMAN.2016.7745164

Rebekić, A., Lončarić, Z., Petrović, S., & Marić, S. (2015). Pearson's or spearman's correlation coefficient-which one to use? *Poljoprivreda*, *21*(2), 47–54. https://doi.org/10.18047/poljo.21.2.8

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM.

Sanchez, J., Rogers, W. A., Fisk, A. D., & Rovira, E. (2014). Understanding reliance on automation: Effects of error type, error distribution, age and experience. *Theoretical Issues in Ergonomics Science*, *15*(2), 134–160. https://doi.org/10.1080/1463922X.2011.611269

Schaefer, K. (2013). *The perception and measurement of human-robot trust* [Unpublished doctoral dissertation]. University of Central Florida.

Schaefer, K. E., Billings, D. R., Szalma, J. L., Adams, J. K., Sanders, T. L., Chen, J. Y., & Hancock, P. A. (2014). A meta-analysis of factors influencing the development of trust in automation: Implications for human-robot interaction. Army Research Lab Aberdeen Proving Ground Md Human Research and Engineering.

Sharan, N. N., & Romano, D. M. (2020). The effects of personality and locus of control on trust in humans versus artificial intelligence. *Heliyon*, *6*(8), e04572. https://doi.org/10.1016/j.heliyon.2020.e04572

Siau, K., & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal*, *31*(2), 47–53.

Sofianos, A. (2022). Self-reported & revealed trust: Experimental evidence. *Journal of Economic Psychology*, *88*, 102451. https://doi.org/10.1016/j.joep.2021.102451

Tutul, A. A., Nirjhar, E. H., & Chaspari, T. (2021). Investigating trust in human-machine learning collaboration: A pilot study on estimating public anxiety from speech. In *Proceedings of the 2021 international conference on multimodal interaction* (pp. 288–296). ACM. https://doi.org/10.1145/3462244.3479926

Ullrich, D., Butz, A., & Diefenbach, S. (2021). The development of overtrust: An empirical simulation and psychological analysis in the context of human–robot interaction. *Frontiers in Robotics and AI*, *8*, 554578. https://doi.org/10.3389/frobt.2021.554578

Yadav, M., Sakib, M. N., Nirjhar, E. H., Feng, K., Behzadan, A., & Chaspari, T. (2020). Exploring individual differences of public speaking anxiety in real-life and virtual presentations. *IEEE Transactions on Affective Computing*, *13*(3), 1168–1182. https://doi.org/10.1109/taffc.2020.3048299

Yang, F., Huang, Z., Scholtz, J., Arendt, D. L. (2020). How do visual explanations foster end users' appropriate trust in machine learning? In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (pp. 189–201). ACM.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). *Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In* [Paper presentation]. (p. 295–305). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372852

About the authors

Abdullah Aman Tutul received his B.Sc. in Computer Science and Engineering from Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh in 2019. He is pursuing his Ph.D. in Computer Science at Texas A&M University, under the supervision of Dr. Theodora Chaspari.

Ehsanul Haque Nirjhar received his B.Sc. in Electrical and Electronic Engineering from Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh in 2015. He is pursuing his Ph.D. in Computer Science at Texas A&M University, under the supervision of Dr. Theodora Chaspari.

Theodora Chaspari (S'12, M'17) received her Ph.D (2017) and M.S. (2012) in Electrical Engineering from the University of Southern California, and diploma in Electrical & Computer Engineering from the National Technical University of Athens, Greece (2010). She is an Associate Professor in Computer Science at the University of Colorado Boulder.