Investigating Trust in Human-Machine Learning Collaboration: A Pilot Study on Estimating Public Anxiety from Speech

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

Ehsanul Haque Nirjhar
nirjhar71@tamu.edu
Texas A&M Univeristy
College Station, TX, USA

Theodora Chaspari
chaspari@tamu.edu
Texas A&M Univeristy
College Station, TX, USA

ABSTRACT

Trust is a key element in the development of effective collaborative relationships between humans and increasingly complex artificial intelligence (AI) systems. Here, we examine trust in AI in the context of a human-AI partnership that involves a joint decision making task for estimating levels of public speaking anxiety based on speech signals. The AI system is comprised of an explainable machine learning (ML) algorithm, that takes acoustic characteristics as input and outputs the estimate of public speaking anxiety levels, a local explanation about the most important features that contributed to the decision of each speech sample, and a global explanation about the most important features for the data overall. We analyze interactions between AI and human annotators with background in psychological sciences, and measure trust over time via the annotators' agreement with the AI model and the annotators' self-reports. We further examine factors of trust that are related to the characteristics of the human annotator and the ML algorithm. Results indicate that trust in AI depends on the openness level of the annotator and the importance level of input features. Findings from this study can provide guidelines to designing solutions that properly calibrate human trust in AI in collaborative human-AI tasks.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning approaches**.

KEYWORDS

Trustworthy AI; human-AI interaction; public speaking anxiety; speech

ACM Reference Format:
Abdullah Aman Tutul, Ehsanul Haque Nirjhar, and Theodora Chaspari. 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 (ICMI '21), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3462244.3479926

ICMI '21, October 18–22, 2021, Montréal, QC, Canada

ACM ISBN 978-1-4503-8481-0/21/10…$15.00
https://doi.org/10.1145/3462244.3479926

1 INTRODUCTION

Recent advances in artificial intelligence (AI) have contributed to significant progress in every field imaginable, including health, education, commerce, and entertainment. While the decisions provided by an AI system can be readily used in some domains (e.g., image-based object recognition in industrial applications), they require additional checkpoints in human-centered applications, such as the fields of medicine [10, 26, 27] and justice [12]. Individuals and stakeholders need to calibrate their trust on the algorithms that power the decision making process in such human-centered applications, which tend to be subjective and involve high stakes [14, 20]. Poor trust calibration results into individuals overly or insufficiently relying on the AI with serious implications in the final outcome [22, 29].

Trust in automation has been investigated in the light of various human factors, including user traits (e.g., demographics, personality, trust propensity, attitudes) and states (e.g., attentional control, stress, fatigue) [23]. While human-related characteristics have been examined in prior work as potential factors of trust to automation [13, 32], to the best of the authors' knowledge, human factors of trust in AI are to date left unexplored. Prior work has further investigated various system-related factors that affect human trust in AI, in particular. The performance and characteristics of the machine learning (ML) algorithm that powers the AI system, including the transparency, explainability, privacy preservation, and fairness, have been posited as important system-based factors [4, 18, 31]. In contrast to black-box ML models, explainable (or glass-box) models can contribute to trustworthy AI by providing a justification of the output decisions[4, 13, 32].

This study investigates trust in AI during a human-in-the-loop decision making task, that aims to estimate one's anxiety levels from speech signals. Annotators with background on psychological sciences are asked to collaborate with an explainable AI algorithm to provide a final decision about a speaker's anxiety levels during a public speaking task. Annotators listen to a presentation and are also being presented the decision of the AI algorithm and corresponding explanations. Trust is quantified via the annotators' agreement with the ML model, as well as self-reports. Statistical analysis of the collected data is conducted through correlation analysis and linear mixed effects (LME) models, the latter used to account for the repeated measures within each annotator, and aims to answer the following research questions:

RQ1: How is trust in AI manifested in a speech-based decision making task jointly conducted between a human stakeholder and an AI system?

RQ2: To what extent is trust in AI affected by the annotator's expertise and personality characteristics?

RQ3: To what extent is trust in AI affected by the ML characteristics?
RQ4: Does trust in AI change over time?
Results indicate that trust in AI varies across annotators. Annotators with open personality traits depict higher trust to AI compared to their counterparts. System characteristics also affect trust in AI with annotators depicting higher trust in cases where speech pause duration has been considered important by the explainable ML system. Finally, trust in AI appears to be a dynamically changing concept with some participants depicting decreasing trust in the AI over the course of their interaction with the system, while others present opposite trends.

2 PRIOR WORK & STUDY CONTRIBUTIONS

Trust in AI has been recently investigated in a limited number of studies. Vivian *et al.* conducted an experiment with Amazon Mechanical Turkers, who were asked to classify deceptive reviews from TripAdvisor with the assistance of an AI model [13]. The authors examined the annotators decision without the help of the AI, as well as across three human-AI collaboration conditions (i.e., presenting the ML decision; presenting the feature explanation provided by the ML; presenting both the ML decision and feature explanation). Providing both the ML decisions and explanations depicts similar levels of human performance compared to providing the AI decisions alone. Zhang *et al.* designed an AI-powered healthcare system for the diagnosis of radiology reports, and examined how user trust is affected by the extent of explanation provided by the system and the system performance [33]. The same group also examined elements of trustworthiness in a prediction context where participants were asked to predict whether one's annual income would exceed $50k based on their demographic and job information [32]. Wang & Yin conducted experiments on Amazon Mechanical Turk to understand the extent to which various explainable AI methods and models can improve people's understanding of the decision making task and help calibrate their trust to the AI [28]. The decision making tasks were inspired by domains related to criminal justice and environmental sustainability. Ayoub *et al.* applied an explainable natural language processing model to COVID-19 claims. Trust in model prediction was evaluated with self-reports across three different conditions that involved presenting different components of the model [1].

Prior work on trust in automation has explored various human-related factors that affect the interaction between a human and an autonomous agent (e.g., vehicle, machinery). Previous studies have posited user expertise as a factor of trust and trust repair toward automation [9, 19]. Findings suggest that novice users are more prone to automation bias, meaning that they are more likely to overtrust an automated system. Personality also affects user trust. Drivers with open personality traits depicted less trust in an autonomous vehicle, potentially due to their curiosity in response to novel stimuli [15]. People with high openness were further faster in their response to an automated system, which serves as a measure of trust [25]. Higher agreeableness and conscientiousness result in higher initial trust in automation [5]. Participants who score higher in agreeableness and openness and low on conscientiousness are found to have higher trust in the AI, as manifested via slower reaction time in the decision making process. Openness and agreeableness served as a factor of trust to a conversational agent, according to which open and agreeable individuals were more likely to confide in and listen to the agent [34]. Grounded on these findings, we will explore four main human factors of trust in AI, namely user expertise, agreeableness, conscientiousness and openness.

The contributions of this paper in relation to prior work are as follows: (1) While prior work has examined trust in AI mostly using text and images as inputs [1, 13, 32], in this study annotators were asked to listen and perceive speech signals, which are more complex in nature compared to other modalities; (2) Instead of using Amazon Turkers, this study investigates interactions between AI and human stakeholders with prior training in the domain of interest, which are more difficult to recruit, but can more reliably simulate real-life human-centered applications; and (3) The characteristics of annotators are investigated as factors of trust to the AI. Such factors have been investigated in relation to trust in automation [9, 15, 19, 25], but are still unexplored in relation to trust in AI.

3 SPEECH-BASED AI SYSTEM FOR ESTIMATING PUBLIC SPEAKING ANXIETY

3.1 Data Description

The data used to train the explainable ML for estimating public speaking anxiety come from the VerBIO dataset and include 78 speech files from 55 speakers [30]. We extracted seven acoustic features, including the mean pause duration, loudness (i.e., computed as the logarithm of the mean square energy), fundamental frequency (F0), zero crossing rate, jitter, shimmer, and voicing probability, since these are intuitive, easily interpretable, and related to the public speaking anxiety [2, 6]. A human expert with expertise in behavioral coding listened to each audio file and provided the perceived anxiety levels of the speaker on a 5-point Likert scale (i.e., 1: No anxiety, 5: Very high anxiety). The human expert listened to the audio files as many times as necessary in order to make a reliable decision. The scores from the human expert are used as the ground truth in this study.

3.2 Explainable ML Algorithm

Grounded on work that supports the significance of explainable ML in facilitating human-AI partnerships[13, 28, 32], we used the Explainable Boosting Machine (EBM) [21], a glass-box model with comparable performance to state-of-the-art ML methods. The EBM is trained to estimate the speaker's levels of public speaking anxiety based on the acoustic measures (Section 3.1). Based on the EBM, training is conducted on one feature at a time in a round-robin fashion cycling through all features x_j where j is the index of feature. In this way, EBM can mitigate the effects of co-linearity and learn the best feature function f_j for each feature x_j and the outcome of interest y. Mathematical formulation of the model is as follows:

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

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 shows how each feature contributes to the model's prediction for estimating public speaking anxiety. While the original representation of the EBM [21] includes

Figure 1: Global explanation graph capturing the effect of loudness feature in estimating public speaking anxiety levels, as provided by the Explainable Boosting Machine (EBM).

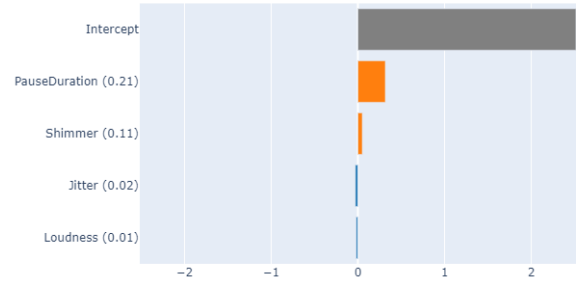


Figure 2: Example of local explanation graph. Larger absolute value indicates higher importance of the corresponding feature in estimating anxiety for the audio sample.

feature interactions, there were not considered in our experiment, since they would increase the complexity of the model and would likely be less intuitive for the users [16].

We first trained the EBM using all acoustic features and evaluated the EBM in a leave-one-participant-out cross-validation framework that yielded 0.0772 ($p = 0.5$) Spearman's correlation between the estimated and actual levels of anxiety. We then performed a feature selection method that filtered out the acoustic measures that depicted low Spearman's correlation (i.e., < 0.1) with the outcome of interest. This resulted in a reduced feature set (i.e., pause duration, loudness, jitter, shimmer) that yielded an improved Spearman's correlation value of 0.2609 ($p < 0.05$), a result equivalent with prior work on the same dataset [30].

The EBM model provides a global explanation graph, which shows the correlation between each feature and output label based on all the data (Fig. 1). The x-axis of the graph denotes the feature values and the y-axis denotes the effect of feature on the outcome. Positive values (i.e., solid blue line; Fig. 1) indicate positive association between the acoustic feature and the anxiety outcome, while the opposite holds for negative values. The graph also presents the confidence of the model for each feature value with thick shaded areas indicating higher uncertainty about the decision. Another output of the EBM model is the importance of each feature for a particular speech file in estimating public speaking anxiety, which is referred to as local explanation. Fig 2 shows a local explanation graph for a sample audio, for which pause duration is the most important feature, while loudness is the least important. In Fig 2, the intercept corresponds to the average public speaking anxiety, represented by variable β_0 in (1).

Characteristic	Mean ± Stand. Dev.	Range
Agreeableness	35.72 ± 3.22	[9 - 45]
Conscientiousness	35.36 ± 3.67	[9 - 45]
Openness	35 ± 5.13	[10 - 50]

Table 1: Personality characteristics of the 11 annotators who participated in the study.

4 USER STUDY DESIGN

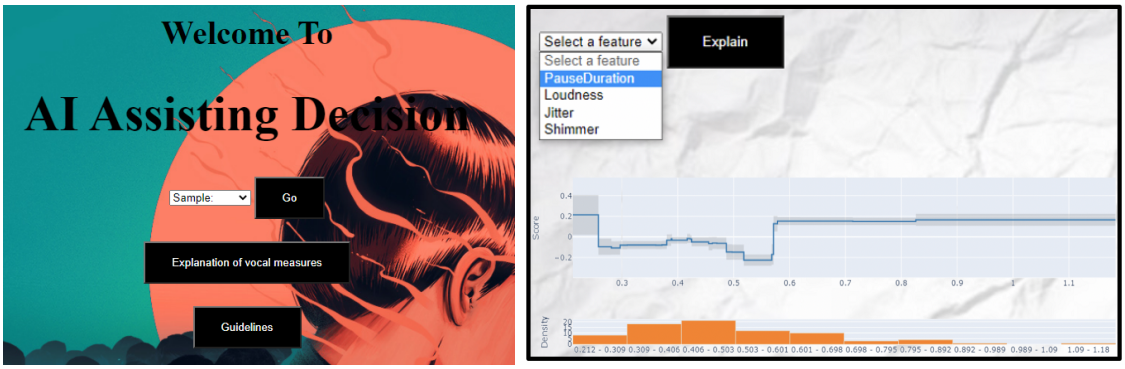
The purpose of our study is to investigate how human stakeholders interact with and trust the explainable AI model (Section 3) in estimating anxiety levels based on speech. Our study included 11 annotators (8 female, 3 male; 19.6 ±0.97 years), who are students from the department of Psychological & Brain Sciences. Recruited annotators were familiar with basic concepts related to human behavior and felt sense. Two annotators were Asian, five were White/Caucasian, three were Hispanic/Latino, and one was Black/African American. Each annotators was compensated with $180.

We created a web interface (Fig 3, Supplementary material) through which annotators interacted with the AI model. The interface contains:

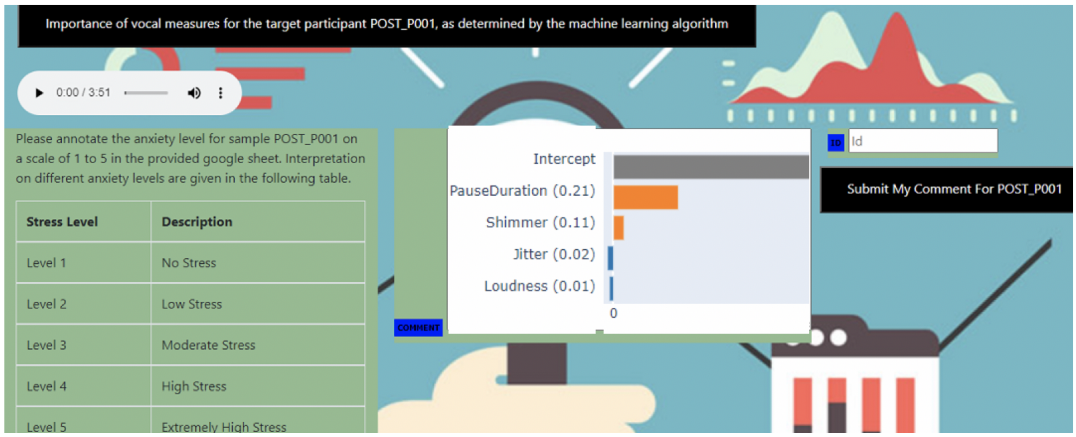
- An audio player to listen to the audio files.
- The global explanation graphs (Section 3) explaining the association between each feature and the estimated anxiety.
- The local explanation graph (Section 3) explaining the relative importance of each feature for each audio file.
- A comment box so that participants can provide their observations for each audio file.
- Help buttons so that annotators can quickly refer to these as needed.

We first tested our website with five annotators, who were compensated with an additional amount of $40. Since our initial version of the algorithm contained 7 features, these were deemed as too many features by the five annotators. Therefore, per annotators' request, we reduced the number of features to 4 and added explanations of the local explanation graphs. Prior to the study, the annotators completed the Big Five Inventory [11] to capture personality traits, and in particular, agreeableness, conscientiousness, and openness, which are the focus of this study (Section 2). The distribution of these scores for all annotators is shown in Table 1. Based on the annotators' resume, we also measured their expertise in behavioral coding. This comprised a binary variable with value of 1 if the annotator had conducted behavioral coding before, and 0 otherwise. Three out of the eleven annotators in our data had expertise with behavioral coding. Annotators were also provided a mini-tutorial that introduced the basics of speech measures through a presentation that explained the intuition and interpretation of the four features used in the EBM model (i.e., pause duration, jitter, shimmer, loudness), along with examples of audio samples with high, medium and low value of each of the features. The first author further conducted a one-to-one meeting with each annotator, in which he explained the task, EBM model, and web interface, and answered their questions. The first author was also available throughout the duration of the experiment for any additional questions.

As part of the study procedure, annotators were first instructed to see the global explanation graph for each feature so that they can understand how the AI model interprets the association between



(a) (b)



(c)

Figure 3: Custom-made web interface used by the annotators. (a) Home page, including links to the audio sample to be selected for annotation, explanations of vocal measures, and guidelines of the annotation process; (b) Drop down list which contains all the four features used in our model. After selecting a feature, the user will be able to inspect the global explanation graph for the selected feature; (c) Web page which gives the ability to play the audio file, inspect the local explanation graph, provide the estimated public speaking anxiety rating, and write comments.

each feature and the anxiety level. After that, they listened to each audio file, observed the local explanation graph provided by the AI (i.e., EBM model), and were shown the anxiety level estimated by the AI. Based on these, they were instructed to report the anxiety level of the speaker in the corresponding audio file on a 5-point Likert scale (i.e., same as the ground truth; Section 3.1), which serves as the final annotation score for each annotator. The annotators were explained that their final annotation score does not need to be aligned with AI's decision, and that they may agree or disagree with the AI. They were further highly encouraged to provide a comment for each audio file explaining their thought process and the reason why they agree or disagree with AI decision.

Four randomly selected audio files were provided twice to each annotator in order to check their consistency. This resulted in a total of 82 audio files. Since the annotation process was a cognitively demanding task, annotators were instructed to annotate 8 batches of files. Each batch included 10 files except the last batch which included 12 files and annotators were advised to spend approximately 2 hours for each batch. 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 (1: Not at all; 5: Extremely).

At the end of the study, annotators were asked to report the most and least useful acoustic features in regards to the decision making task. The first author also conducted an exit discussion with each annotator and asked them about their experience with the AI, trust or mistrust issues with the AI, and interpretability and usability of each feature for estimating a speaker's anxiety level.

5 DATA ANALYSIS AND RESULTS

We will show our results in five parts. First, we will demonstrate the within-person consistency of each annotator (Section 5.1). Then we will present the annotators' agreement with the ground truth and the AI (Section 5.2), and demonstrate how the annotators' and AI characteristics moderate the annotators' agreement with the ground truth and the AI (Sections 5.3, 5.4). Finally, we will explain how the annotators trust in AI evolves over time (Section 5.5).

Annotator ID	Average difference
1	0.25
2	0.375
3	0.325
4	0.2
5	0.5
6	0.6
7	0.2
8	0.125
9	0
10	0.625
11	0.525
12	0.75

Table 2: Average difference in the annotation score for duplicate samples by each annotator.

5.1 Consistency of annotators

To quantify the consistency of annotators, we compute the average difference in the annotation scores for the four duplicate samples that were provided during the annotation process (Table 2). This score is close to zero for the majority of the annotators, while it becomes larger than zero for Annotator 12, whom we excluded from the following analysis.

5.2 Annotators' agreement with ground truth and AI - Measures of trust in AI

We compute the Spearman's correlation $r_{GT}^{(i)}$ between the ground truth labels and the scores provided by annotator i. The median of $r_{GT}^{(i)}$ values across annotators is 0.414 and the distribution of the corresponding values is shown in Fig 4, which indicates that most of the annotators depict moderate to high positive correlation with the ground truth. This suggests that the annotators are doing a reasonable job in providing these scores. We further compute the Spearman's correlation $r_{AI}^{(i)}$ between the estimates provided by the AI and the annotation scores provided by annotator i. In this case, we observe a wider range of Spearman's correlation values with the corresponding distribution skewed toward the left. This indicates that the majority of the annotators depict low to moderate agreement with the AI, which can be justified by the moderate performance of the EBM model in estimating anxiety levels (i.e., $r = 0.26$; Section 3). It also appears that there are some annotators (i.e., three annotators with $r_{AI}^{(i)} > 0.5$) who are in high agreement with the AI system. The distribution of $r_{AI}^{(i)}$ in Fig 5, the average $r_{AI}^{(i)}$ value from 11 annotators is 0.427 and the median of $r_{AI}^{(i)}$ is 0.417.

Based on these two correlations and grounded on prior work [24], we also compute a measure of trust in AI for each annotator i as:

$$Ratio^{(i)} = \frac{r_{AI}^{(i)}}{r_{GT}^{(i)}} \tag{2}$$

If annotator i agrees more with the ground truth than the AI decision, then the $Ratio_i$ will be less than 1, which potentially suggests mistrust to the AI system. On the contrary, if annotator i agrees more with the AI decision compared to the ground truth, then $Ratio_i$ will be greater than 1. In the latter case, this would suggest

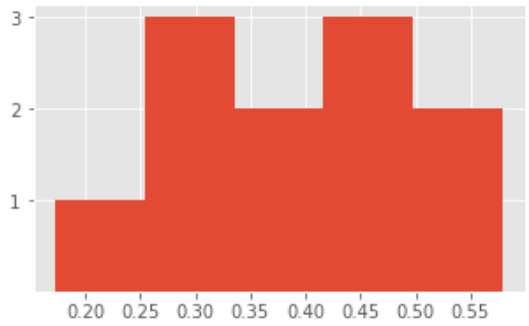


Figure 4: Distribution of Spearman's correlation $r_{GT}^{(i)}$ between the ground truth and annotation scores provided by each annotator.

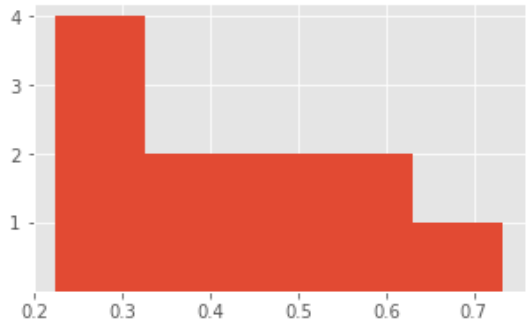


Figure 5: Distribution of Spearman's correlation $r_{AI}^{(i)}$ between the AI decision and annotation scores provided by each annotator.

that the annotator is biased over the decision of the AI system, which can potentially lead to overtrusting the AI. The median value of $Ratio_{(i)}$ computed over all annotators is 0.9, while the distribution of this measure is shown in Fig 6. We notice that the $Ratio_{(i)}$ values are centered around one for 4 annotators, which indicates that the annotators' judgement matches equally with the true label and the AI decision, therefore suggesting good calibration of trust to the AI. However, there appear to be 3 annotators who slightly overtrust the AI (i.e., $Ratio_{(i)} > 1$) and 1 annotator who tends to extremely overtrust the AI system (i.e., $Ratio_{(i)} >> 1$), as well as 3 annotators who mistrust the AI (i.e., $Ratio_{(i)} < 1$).

5.3 Effect of annotators' characteristics on their agreement with ground truth and AI

We further explore the extent to which the association between the annotations and the ground truth, as well as between the annotations and the AI decision is moderated by the annotators characteristics. Grounded on prior work [5, 7, 9, 15, 19, 25, 34] (Section 2), we examine the annotators' expertise with behavioral coding, agreeableness, conscientiousness, and openness. In the following models, we used 78 annotation scores provided by each of the 11 annotators, resulting in a total 858 samples, and all input data were normalized. We build a LME model with random intercept to estimate the annotation score of audio sample j provided by annotator i as follows:

$$Y_{i,j} = \beta + a_1 \times T_j + b_1 \times A_i + c_1 \times (T_j \times A_i) + x_i \tag{3}$$

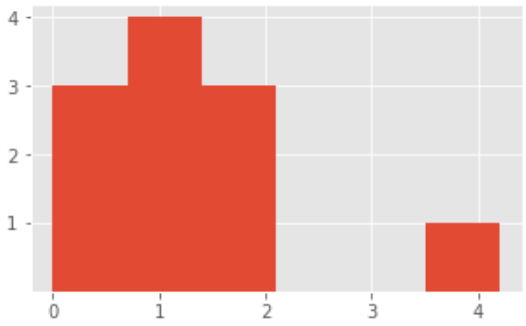


Figure 6: Distribution of $Ratio^{(i)} = \frac{r_{AI}^{(i)}}{r_{GT}^{(i)}}$, serving as a measure of annotators' trust to the AI system. $Ratio^{(i)} \approx 1$ suggests calibration of trust, $Ratio^{(i)} > 1$ overtrust, and $Ratio^{(i)} < 1$ mistrust.

where $Y_{i,j}$ denotes the annotation score of sample j by annotator i, T_j denotes the ground truth for sample j, A_i is a characteristic (i.e., expertise in behavioral coding, agreeableness, conscientiousness, openness) related to annotator i, β represents the fixed intercept for all observations, and x_i denotes the random intercept for each annotator i. In (3), a, b, and c are fixed effect coefficients which are constant for all observations, and x_i is a random effect coefficient which is different for each participant i. Coefficients a and b quantify the association between annotation score with ground truth, as well as annotation score with the annotators' characteristics, respectively. Parameter c quantifies the moderation effect of annotator's characteristic on the association between annotation score and ground truth (i.e., higher absolute values of c indicating stronger moderation). The estimated model coefficients and the corresponding p–values are reported in Table 3. Results suggest a significant positive association between annotation score and true label for all annotator characteristics. They also suggest that agreeableness moderates the association between annotation score and ground truth in a positive and significant way (i.e., $c_1 = 0.86$, $p < 0.05$), indicating that annotators with agreeable personality characteristics tend to agree more with the ground truth. Similarly, annotators with expertise in behavioral coding appear to agree more with the ground truth (i.e., $c_1 = 0.54$, $p < 0.05$). Conscientious annotators tend to also agree more with the ground truth compared to their counterparts, although this moderation is approaching significance ($c_1 = 0.74$, $p = 0.06$). Finally, no significant moderation effect was found for openness.

Similarly, we analyzed the extent to which the annotators' characteristics moderate the association between the annotation score and the AI decision via the following LME model:

$$Y_{i,j} = \beta + a_2 \times M_j + b_2 \times A_i + c_2 \times (M_j \times A_i) + x_i \quad (4)$$

All the variables in (4) are the same as in (3) except M_j, which denotes the anxiety score estimated by the AI model for the j-th audio sample (Section 3). Results demonstrate a significant positive association between the annotation score and the AI decision (i.e., positive values of a; Table 4). Annotators with highly open personality tend to agree more with the AI decision compared to their counterparts with low openness (i.e., $c_2 = 0.76$, $p < 0.01$).

Annotator Characteristic	a_1	b_1	c_1
Agreeableness	0.88(p=0)	-0.69(p=0.009)	0.86(p=0.020)
Expertise in behavioral coding	1.15(p=0)	-0.50(p=0.001)	0.54(p=0.015)
Conscientiousness	0.98(p=0)	-0.67(p=0.019)	0.74(p=0.068)
Openness	1.40(p=0)	-0.10(p=0.625)	-0.181(p=0.50)

Table 3: Linear mixed effects (LME) model estimates of fixed and interaction effects of annotator's characteristics on the association between annotation score and ground truth.

Annotator Characteristic	a_2	b_2	c_2
Agreeableness	1.05(p=0)	-0.47(p=0.099)	0.41(p=0.284)
Expertise in behavioral coding	1.23(p=0)	-0.25(p=0.13)	0.06(p=0.804)
Conscientiousness	1.06(p=0)	-0.53(p=0.090)	0.42(p=0.312)
Openness	0.80(p=0)	-0.64(p=0.002)	0.76(p=0.006)

Table 4: Linear mixed effects (LME) model estimates of fixed and interaction effects of annotator's characteristics on the association between annotation score and AI decision.

Feature	a_3	b_3	c_3
Pause	0.54(p=0.008)	-0.82(p=0.068)	1.85(p=0.002)
Loudness	1.47(p=0)	0.17(p=0.417)	-1.20(p=0.002)
Jitter	0.76(p=0)	-2.97(p=0.010)	5.49(p=0.004)
Shimmer	1.30 (p=0)	0.13(p=0.726)	-0.10(p=0.868)

Table 5: Linear mixed effects (LME) model estimates of fixed and interaction effects of acoustic feature value on the association between annotation score and AI decision.

The other characteristics did not yield any significant moderation effects.

5.4 Effect of AI input to the annotator's agreement with AI

We also investigate the extent to which the value of the acoustic features, as well as the importance of the acoustic features, as estimated by the EBM model, affects the annotators' trust to the AI. We build the following LME model:

$$Y_{i,j} = \beta + a_3 \times M_j + b_3 \times f_{k,j} + c_3(M_j \times f_{k,j}) + x_i \quad (5)$$

All the parameters in (5) are the same as in (4) except $f_{k,j}$, which denotes the value of feature k (i.e., jitter, shimmer, pause, loudness) for the j-th audio sample. As expected, we observe a significant positive association between the annotation score and AI decision (Table 5). Annotators agree more with the AI model for samples with high pause duration ($c_3 = 1.85$, $p < 0.01$) and jitter ($c_3 = 5.49$, $p < 0.01$). In contrast, annotators agree less with the AI model for samples with high loudness ($c_3 = -1.20$, $p < 0.01$). Non-significant moderation effects were found for shimmer.

We further study the extent to which the importance of each feature, as estimated by the AI through the EBM model, affects the

Feature	a_4	b_4	c_4
Pause	0.62(p=0.003)	-0.92(p=0.017)	1.92(p=0.001)
Loudness	1.48(p=0)	0.43(p=0.03)	-2.12(p=0)
Jitter	1.30(p=0)	0.64(p=0.12)	-0.45(p=0.368)
Shimmer	1.49(p=0)	0.86(p=0.012)	-0.99(p=0.026)

Table 6: Linear mixed effects (LME) model estimates of fixed and interaction effects of acoustic feature importance on the association between annotation score and AI decision.

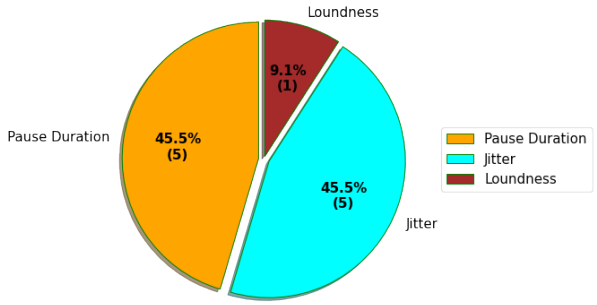


Figure 7: Annotators' responses indicating the most important feature for AI-enabled decision making.

annotators' trust to the AI, via the following LME equation:

$$Y_{i,j} = \beta + a_4 \times M_j + b_4 \times IM_{k,j} + c_4 \times (M_j \times IM_{k,j}) + x_i \quad (6)$$

All the parameters in (6) are the same as in (5) except $IM_{k,j}$ which denotes the absolute feature importance of feature k for the j audio sample. We used the absolute value of feature importance, rather than the actual one, since this is easier to interpret (i.e., large absolute value indicates high importance, irrespective to positive or negative association between the feature and the anxiety score). Annotators tend to agree more with the AI when pause duration is deemed as an important feature (i.e., $c_4 = 1.92$, $p < 0.01$), while they agree less with the AI when loudness (i.e., $c_4 = -2.12$, $p = 0$) and shimmer (i.e., $c_4 = -0.99$, $p < 0.05$) are considered important features (Table 6). This indicates that annotators might perceive pause duration as an important feature for estimating anxiety, while there is potentially a confusion in regards to loudness and shimmer. No significant results were found for jitter. These findings are consistent with the answers to the surveys provided by the annotators at the end of the user study. The majority of annotators reported that pause duration and jitter were the most useful features, followed by loudness and shimmer (Fig. 7). Loudness and shimmer were further found the least useful features (Fig. 8).

5.5 Evolution of human trust in AI over time

The self-reported score of trust to AI across time for each annotator is provided in Fig 9. We observe that 7 out of the 11 total annotators show higher or equal trust in AI at the end, compared to the beginning of the annotation task (i.e., solid lines; Fig 9). The remaining four annotators depict a decrease in their trust in the AI throughout the annotation procedure (dotted lines; Fig 9). We further empirically examine the characteristics of the annotators in the two groups. Annotators who depict increasing trust over time are more open and agreeable (i.e., openness = 35.85 ± 4.76,

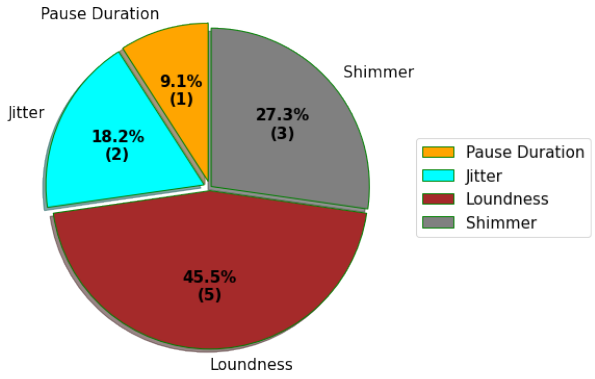


Figure 8: Annotators' responses indicating the least important feature for AI-enabled decision making.

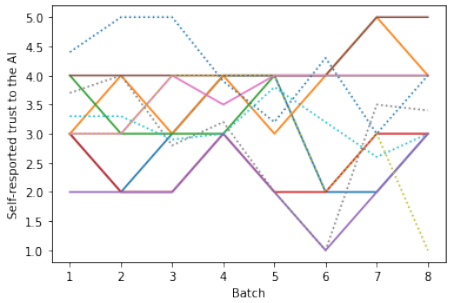


Figure 9: Evolution of trust in AI over time for each annotator.The y-axis denotes the AI-trust score for each batch of 10 audio samples and x-axis denotes batch index in chronological order. The solid lines correspond to the 7 annotators who show higher or equal trust in AI at the end than the beginning and the dotted lines correspond to rest of the annotators whose trust in AI tends to decrease at the end of the task.

agreeableness = 36 ± 3.11) compared to the annotators who have depicted decreasing trust in AI (i.e., openness = 33.5 ± 5.4, agreeableness = 35.25 ± 3.34). Although additional quantitative analysis is necessary, this can potentially suggest that open and agreeable individuals are more tolerant to the errors of the AI: even if the system makes an error due to its overall moderate performance, they still recognize its value over time.

6 DISCUSSION

In answering the first research question (**RQ1:** How is trust in AI manifested in a speech-based decision making task jointly conducted between a human stakeholder and an AI system?), we observed varying levels of trust across annotators with some depicting overtrust (i.e., high agreement with the AI and low agreement with the ground truth), while others portraying mistrust (i.e., low agreement with the AI and high agreement with the ground truth) in the system. Similar observations were found in terms of self-reports. There are a variety of factors that can explain this variance, which will be discussed in the following.

6.1 Human expertise as a factor of trust

In response to the second research question (**RQ2:** To what extent is trust in AI affected by the annotator's expertise and personality characteristics?), our results indicate that annotators with expertise in behavioral coding tend to agree more with the ground truth compared to their counterparts (Section 5.3). This can be justified by the fact that annotators with more expertise are likely to more reliably recognize anxiety levels in speech, therefore agree more with the ground truth. However, expertise in behavioral coding was not found to be a moderating factor of the annotator's trust in AI (Sections 5.3, 5.5). This finding is partially in agreement with prior work that has found that novice and experienced pilots self-report similar level of trust to an automatic system [17]. A potential reason might be the fact that all annotators are still pursuing their undergraduate degree, therefore the expertise in behavioral coding between the two groups is not substantially different to yield significant results. In addition to this, all annotators had no prior experience in working with AI systems, therefore the novelty of the AI-assisted decision making task might potentially dominate the user experience.

6.2 Human personality as a factor of trust

We found that open individuals trust the AI decision more compared to non-open individuals (Section 5.3). This is in accordance to the majority of work that has explored human openness as a factor of trust in automation [25, 34]. Since open individuals are curious and tend to seek new experiences, it might be the case that they tend to be more trusting during their new encounter with the AI. In contrast to prior work [5, 25, 34], we did not find a significant effect on trust for agreeableness and conscientiousness. Previous work has found that individuals who are more agreeable, therefore more cooperative than their less agreeable counterparts, depict a higher trust in automation [5, 25, 34]. Although our results are in the right direction (i.e., indicating positive association between trust and agreeableness, Table 4), a potential reason for the non-significant finding might be the relative small variance in annotators' agreeableness in our data, which might prevent the LME model from adequately fitting the data. Prior findings are mixed with respect to conscientiousness. Some research suggests that conscientious users are more likely to trust the AI [5], while other work has hypothesized the opposite, but with no significant findings [25].

6.3 AI elements as a factor of trust

In answering the third research question (**RQ3:** To what extent is trust in AI affected by the ML characteristics?), our analysis indicates that feature importance serves as a factor of trust in the AI (Table 6). Specifically, annotators depicted increased trust in the AI when the ML model deemed that pause duration was important for estimating anxiety. This is consistent with the small number of prior findings, which indicate that individuals trust explainable AI systems more when the feature contribution explanation is shown [28] improving the overall accuracy of the decision [13]. Results on loudness were slightly conflicting to what was expected, since annotators depicted decreased trust in the system in cases where the importance of the corresponding feature was high. A potential reason for this might be that loudness is highly dependent on the position of the microphone. During the exit discussion, seven out of the eleven annotators perceived that high loudness was associated with close position of the microphone, rather than the nervousness of the speaker. Our analysis also suggests that annotators do not agree with the AI estimate for samples with high loudness (Table 5). The EBM model also indicates that high loudness has high impact on the decision, though this effect is uncertain (i.e., sharp decline of the blue curve and thick grey area for high loudness values in Fig. 1). As part of our future work, we will take advantage of the modularity of the EBM model and its user interface (Fig. 1) [3], which can be re-adjusted by human experts.

6.4 Trust over time

Our analysis suggests that trust in the AI changes over time (Section 5.5), therefore providing a positive response to the fourth research question (**RQ4:** Does trust in AI change over time?). Empirical findings further indicate that agreeable and open annotators might depict increasing trust in AI, while decreasing patterns were observed for their less agreeable or less open counterparts. While this needs to be formally evaluated via statistical analysis, it provides evidence consistent with prior work that trust is a reactive and volatile property [23], therefore it is important to understand how it evolves over time. It is also necessary to investigate how AI systems can be designed in order to build and actively repair trust over time (e.g., via introducing humanness characteristics) [9].

6.5 Limitations

Our study depicts various limitations, which will be addressed as part of our future work. First, trust has been quantified via the annotator's agreement to the AI, as well as via self-reports. Grounded in findings from previous work [8, 25], we plan to capture behavioral (e.g., reaction time) and neural (e.g., observational error-related negativity/positivity) measures of trust, which we anticipate that they will provide complementary findings. Second, this study explored a user's trust to an explainable AI system (i.e., EBM model) without comparing various ML characteristics (e.g., ML confidence) or a control condition that does not include AI-enabled decision making. Third, the ground truth used in this model was obtained by one expert annotator. Although we have a good reason to believe that this annotator provided reliable ratings of public speaking anxiety, as part of our future work we will construct our ground truth using multiple experts. Finally, as part of our future work we plan to adapt the web interface that was used for the annotation process for colorblind individuals.

7 CONCLUSIONS

We examined trust in a human-AI collaboration paradigm that aims to estimate levels of public speaking anxiety based on speech. We found varying levels of trust in AI among human annotators, which was partially dependent on the annotators' personality and the importance of the acoustic features, as estimated by the explainable ML. Implications from this study can provide guidelines to training and designing ML for retaining proper levels of trust in AI during human-AI collaboration.

ACKNOWLEDGMENTS

This work was partially 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).

REFERENCES

[1] Jackie Ayoub, X Jessie Yang, and Feng Zhou. 2021. Combat COVID-19 infodemic using explainable natural language processing models. *Information Processing & Management* 58, 4 (2021), 102569.

[2] Ligia Batrinca, Giota Stratou, Ari Shapiro, Louis-Philippe Morency, and Stefan Scherer. 2013. Cicero-towards a multimodal virtual audience platform for public speaking training. In *International workshop on intelligent virtual agents*. Springer, 116–128.

[3] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1721–1730.

[4] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 8 (2019), 832.

[5] Shih-Yi Chien, Katia Sycara, Jyi-Shane Liu, and Asiye Kumru. 2016. Relation between trust attitudes toward automation, Hofstede's cultural dimensions, and big five personality traits. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 60. SAGE Publications Sage CA: Los Angeles, CA, 841–845.

[6] Mathieu Chollet, Torsten Wörtwein, Louis-Philippe Morency, and Stefan Scherer. 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)*. 488–495.

[7] Bart A De Jong, Kurt T Dirks, and Nicole Gillespie. 2016. Trust and team performance: A meta-analysis of main effects, moderators, and covariates. *Journal of Applied Psychology* 101, 8 (2016), 1134.

[8] Ewart J de Visser, Paul J Beatty, Justin R Estepp, Spencer Kohn, Abdulaziz Abubshait, John R Fedota, and Craig G McDonald. 2018. Learning from the slips of others: Neural correlates of trust in automated agents. *Frontiers in human neuroscience* 12 (2018), 309.

[9] Ewart J De Visser, Richard Pak, and Tyler H Shaw. 2018. From 'automation'to 'autonomy': the importance of trust repair in human–machine interaction. *Ergonomics* 61, 10 (2018), 1409–1427.

[10] Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg, and Andreas Holzinger. 2018. Explainable AI: The new 42?. In *International cross-domain conference for machine learning and knowledge extraction*. Springer, 295–303.

[11] Oliver P John, Sanjay Srivastava, et al. 1999. *The Big-Five trait taxonomy: History, measurement, and theoretical perspectives*. Vol. 2. University of California Berkeley.

[12] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

[13] Vivian Lai and Chenhao Tan. 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*. 29–38.

[14] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[15] Wenmin Li, Nailang Yao, Yanwei Shi, Weiran Nie, Yuhai Zhang, Xiangrong Li, Jiawen Liang, Fang Chen, and Zaifeng Gao. 2020. Personality Openness Predicts Driver Trust in Automated Driving. *Automotive Innovation* 3, 1 (2020), 3–13.

[16] Yin Lou, Rich Caruana, and Johannes Gehrke. 2012. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 150–158.

[17] Joseph B Lyons, Nhut T Ho, Anna Lee Van Abel, Lauren C Hoffmann, Garrett G Sadler, William E Fergueson, Michelle A Grigsby, and Mark Wilkins. 2017. Comparing trust in auto-GCAS between experienced and novice air force pilots. *ergonomics in design* 25, 4 (2017), 4–9.

[18] Joseph E Mercado, Michael A Rupp, Jessie YC Chen, Michael J Barnes, Daniel Barber, and Katelyn Procci. 2016. Intelligent agent transparency in human–agent teaming for Multi-UxV management. *Human factors* 58, 3 (2016), 401–415.

[19] Christine Moorman, Rohit Deshpande, and Gerald Zaltman. 1993. Factors affecting trust in market research relationships. *Journal of marketing* 57, 1 (1993), 81–101.

[20] Bonnie M Muir. 1994. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics* 37, 11 (1994), 1905–1922.

[21] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223* (2019).

[22] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.

[23] Kristin E Schaefer, Jessie YC Chen, James L Szalma, and Peter A Hancock. 2016. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors* 58, 3 (2016), 377–400.

[24] Philipp Schmidt and Felix Biessmann. 2019. Quantifying interpretability and trust in machine learning systems. *arXiv preprint arXiv:1901.08558* (2019).

[25] Navya Nishith Sharan and Daniela Maria Romano. 2020. The effects of personality and locus of control on trust in humans versus artificial intelligence. *Heliyon* 6, 8 (2020), e04572.

[26] Hans-Christian Thorsen-Meyer, Annelaura B Nielsen, Anna P Nielsen, Benjamin Skov Kaas-Hansen, Palle Toft, Jens Schierbeck, Thomas Strøm, Piotr J Chmura, Marc Heimann, Lars Dybdahl, et al. 2020. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *The Lancet Digital Health* 2, 4 (2020), e179–e191.

[27] Haolin Wang, Zhilin Huang, Danfeng Zhang, Johan Arief, Tiewei Lyu, and Jie Tian. 2020. Integrating Co-Clustering and Interpretable Machine Learning for the Prediction of Intravenous Immunoglobulin Resistance in Kawasaki Disease. *IEEE Access* 8 (2020), 97064–97071.

[28] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*. 318–328.

[29] Jin Xu. 2018. Overtrust of Robots in High-Risk Scenarios. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 390–391.

[30] Megha Yadav, Md Nazmus Sakib, Ehsanul Haque Nirjhar, Kexin Feng, Amir Behzadan, and Theodora Chaspari. 2020. Exploring individual differences of public speaking anxiety in real-life and virtual presentations. *IEEE Transactions on Affective Computing* (2020), 1–1. https://doi.org/10.1109/TAFFC.2020.3048299

[31] X Jessie Yang, Vaibhav V Unhelkar, Kevin Li, and Julie A Shah. 2017. Evaluating effects of user experience and system transparency on trust in automation. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*. IEEE, 408–416.

[32] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.

[33] Zhan Zhang, Yegin Genc, Dakuo Wang, Mehmet Eren Ahsen, and Xiangmin Fan. 2021. Effect of AI Explanations on Human Perceptions of Patient-Facing AI-Powered Healthcare Systems. *Journal of Medical Systems* 45, 6 (2021), 1–10.

[34] Michelle X Zhou, Gloria Mark, Jingyi Li, and Huahai Yang. 2019. Trusting virtual agents: The effect of personality. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 9, 2-3 (2019), 1–36.