Human-AI Collaboration for the Detection of Deceptive Speech

Adullah Aman Tutul
Texas A&M University
abdullahaman633@tamu.edu

Theodora Chaspari
Texas A&M University
chaspari@tamu.edu

Sarah Ita Levitan
Hunter College
sarah.levitan@hunter.cuny.edu

Julia Hirschberg
Columbia University
julia@cs.columbia.edu

***Abstract*—This paper investigates human trust in artificial intelligence (AI) during human-AI collaboration on a speech-based data analytics task. Human users worked together with an explainable AI algorithm that took as an input acoustic and linguistic measures for the detection of deceptive speech. The working performance of the AI was manipulated resulting in a high performing (HP) AI and a low performing (LP) AI. Trust was measured via self-reported and behavioral measures, which were associated with each other. Various personality characteristics, including openness, neuroticism, and extroversion, moderated one's trust in the AI, but results were mixed in terms of the considered self-reported and behavioral trust metrics.**

***Index Terms*—Human-AI collaboration, trust, deceptive speech**

I. Introduction

The detection of deception is a notably challenging task, yet holds sigificant importance across various domains like law enforcement, national security, and medical encounter and diagnostics. Human performance in deception detection is not far above chance [1]. Machine Learning (ML) models can model facial expressions, body gestures, and speech that are indicative of deception with performance ranging between 60-70% [2]. When jointly working together, humans and ML algorithms can potentially leverage their complementary skills for achieving better accuracy on this task. Here, we investigate human trust in a collaborative decision-making task between humans and an explainable artificial intelligence (XAI) algorithm for detecting deception from speech. Via a between-subject study, users were randomly assigned to a high-performing (HP) AI or a low-performing (LP) AI condition. We aim to answer the following research questions: **RQ1:** Does human trust vary across the different AI conditions? **RQ2:** What is the association between self-reported and behavioral measures of trust? and **RQ3:** Do individual factors of personality and trust propensity impact user trust in AI? Results do not indicate significant differences between conditions. Self-reported trust is associated with one's level of agreement with the AI. Findings in terms of personality were not consistent between self-reported and behavioral trust; overall, open users agreed more with the AI, extrovert users agreed less, and neurotic users reported higher trust in the AI.

II. PRIOR WORK

Explainable AI has been used in decision-making to augment human performance in tasks related deception detection for online reviews [3] and spoken dialog [4]. Deception detection is a challenging task for both humans and AI. ML models with lexicon-based features can differentiate between deceptive and truthful views about abortion with 67% accuracy, where the human judges obtain 52% accuracy [5]. In addition, preliminary evidence suggests human performance on this task does not benefit from the AI decision when using linguistic features only [4]. Previous studies found that linguistic features, such as the amount of negative emotion words, language pleasantness, activation, and depicted imagery, can be indicative of deception [5]. However, beyond the speech content, acoustic measures such as pitch, pause behavior, and loudness have been widely studied as acoustic indicators of deception [6], [7] with prior work demonstrating that people rely more on non-verbal cues for this task [8]. Previous research further demonstrated individual differences in human trust in automation. Prior work has found the dependence of human trust on personality characteristics, such as openness and extroversion [9], [10]. One's overall propensity to trust has been examined as a factor of trust with conflicting results [11].

The contributions of this study are: (1) In contrast to the majority of studies that rely on image or text [3], this study examines speech as an alternative, but important modality that can be used in data analytics tasks. XAI systems relying on speech need to capture the unique perceptual characteristics of this modality, which can be less intuitive to the human users compared to images; and (2) To the best of our knowledge, this study is the first to investigate a human-AI collaborative system for identifying speech deception using XAI based on both linguistic and acoustic data; other studies examined the human-AI collaboration for deception detection without using XAI and using linguistic data only [4].

III. Explainable AI for deceptive speech detection

A. Dataset Description

We used the Columbia X-Cultural Deception (CXD) Corpus, that contains dialogues between 340 participants. Each participant answered 24 biographical questions, and they provided the true answer to half of the questions and a false answer to the rest. These responses were used as ground truth. We used only the responses from native speakers of Standard American English to mitigate any bias in the perception experiments due to the nationality [12]. This produced 4373 question-answer pairs (i.e., 2493 and 1880 responses from female and male speakers, respectively).

This research was supported by the Air Force Office of Scientific Research (# FA9550-22-1-0010, PI: Chaspari).

We extracted linguistic and acoustic features that were deemed as important indicators of deception by prior studies. The linguistic features included the percentage of negations [13] and total number of function words [14] extracted by the Linguistic Inquiry and Word Count (LIWC) toolbox [15], mean pleasantness in the response extracted by the Dictionary of Affect in Language (DAL) [16], creativity (i.e., the degree of similarity of a response to other responses of the same topic, where similarity was measured based on the frequency-inverse document frequency (TF-IDF) vectors on unigrams and bigrams between question/response pairs) [17], and number of filled pauses as defined in [17]. Acoustic features included the maximum speech amplitude and maximum pitch [18] within the response extracted using Praat [19].

B. Explainable AI (XAI) system for detecting deceptive speech

We used the explainable boosting machine (EBM) [20] to classify deceptive speech. The EBM estimates the deception outcome y based on a set of features $\{x_j\}$ as follows:

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

where $\mathcal{E}$ is the expected value of the outcome, β_0 is the intercept, and f_j is a feature function that shows the contribution of each feature x_j to the model's output. We did not include any pairwise feature interactions in our model so that it is more intuitive to the users. EBM was trained on 75% of the CXD data using acoustic and linguistic features (Section III-A). The remaining 25% data was employed as the test set. We used stratified sampling based on biological sex and deception outcome for the train-test split. Via varying the hyper-parameters of the EBM, we built the HP AI and LP AI. The HP AI depicted high sensitivity (i.e., true positive rate of deception detection), F1-score, and accuracy in deception detection, and higher accuracy compared to the human baseline accuracy (i.e., 56.75%), while the LP AI showed low sensitivity and lower accuracy compared to the human baseline (Table I). Given that deception is the focal outcome and bears a higher miss cost compared to non deception, we anticipate the user perception about the AI performance will be associated with the sensitivity metric. We randomly selected 25 deceptive and 15 non-deceptive samples, balanced by gender, from the test set for our user study. Out of the total 25 deceptive samples, the HP AI can correctly detect 16 samples, while the LP AI can correctly detect the 9 samples. The performance of HP and LP AI on non-deceptive samples is same with both systems that are able to correctly detect 9 non-deceptive samples out of 15 total non-deceptive samples. The EBM model provides a global explanation graph, which shows the correlation between each feature and output label based on all the data, and a local explanation graph which explains the contribution of each feature in estimating the deception for each audio sample (see Supplementary Material).

IV. User Study Design

A. Participant Recruitment

Eligibility criteria for participants were being: (1) older than 18 years; (2) fluent in English; and (3) a current student or holding a diploma from a field in social sciences. These criteria rendered participants likely to be familiar with basic concepts related to human behavior. Participants were recruited online via mailing lists at the university of the research team. We recruited 37 participants (23.4 ± 7.67 years; 28 female, 8 male, 1 other; 75.6% undergraduate, 24.4% graduate students).

TABLE I
Performance of high-performing (HP) and low-performing (LP) models in automatically detecting deception from speech.

Model	Sensitivity	Specificity	F1-Score	Accuracy
HP	0.61	0.55	0.59	0.58
LP	0.28	0.73	0.36	0.49

B. Study Protocol

Initially, each participant completed a set of questionnaires that included the Big Five Inventory [21] (i.e., extroversion, agreeableness, conscientiousness, openness, neuroticism), Propensity to Trust Machines questionnaire [22], and Trusting Scenarios with AI [23] (see Supplementary Material for the distribution and range of the above). After that, each participant viewed a mini tutorial which explained the goal of the task, the basic intuition of the linguistic and acoustic features, and the EBM model. Half of the participants (i.e., 19) were randomly assigned to HP AI condition and the rest (i.e., 18) were assigned to LP AI condition. The participants interacted with the assigned AI via a user interface (see Supplementary Material for snapshots), listened to the audio samples in the same order, and were told that their own decision does not necessarily need to be aligned with the AI decision. First, they reviewed the global explanation graphs to understand the association between each feature and the probability of perception provided by the EBM model. Then, they read each interview question and listened to the audio response. Following that, the participants viewed the local explanation graph and the AI decision for that sample. Finally, they provided their decision about whether the audio sample response was deceptive or not. This was repeated 40 times (i.e., for each audio sample; Section III-B). After providing decisions on 8 consecutive audio samples (i.e., one batch), participants were asked to rate their perceived trust in AI on a 1-5 Likert scale (1: Low Trust, 5: High Trust), which served as a measure of self-reported trust. Each participant was compensated with a $50 Amazon gift card. The study was approved by the Institutional Review Board (IRB).

V. Methods & Results

A. Measures of trust

Self-reported trust measures include the 1-item trust rating obtained after performing decisions for 8 consecutive audio files. Behavioral measure of trust is encoded as 1/0 depending on whether the annotator decision matches with the AI decision or not for each sample. In order to obtain similar time resolution between the two measures, the behavioral trust was averaged per batch.

B. Comparing HP AI and LP AI conditions

We conducted independent t-test to investigate significant differences between the HP and LP AI conditions in terms of self-reported and behavioral trust. We did not find any significant difference for any of these metrics between the AI conditions. This potentially indicates a weak manipulation of the AI performance. So, for the remaining of the analysis we will be examining data from the two AI conditions altogether.

C. Self-reported and behavioral measures of trust

Here, we examine the association between self-reported and behavioral measures of trust via the following linear mixed-effects (LME) model with random intercept:

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

where $S_{i,j}$ is the self-reported trust of annotator i after batch j has been completed and $AG_{i,j}$ is the average agreement between AI and 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 coefficient a_1 quantifies the association between self-reported trust and agreement with AI. The results suggest significant positive association between the two (i.e., $a_1 = 0.107$, $p = 0.058$, $N = 184$), a finding which is also supported by prior studies [24].

D. Effect of user characteristics on human trust in AI

We investigate individual factors of trust that are related to the user's personality and overall trust propensity and perceptions. In terms of individual factors, we focused on the five personality traits, one's overall propensity to trust the automation, and one's willingness to trust the AI (Section IV-B). In addition, we explored how individual factors moderate the evolution of human trust in AI. In (3)-(4), we employed the LME models with random intercept to analyze the effect of annotators' characteristics on self-reported trust in AI and agreement with the AI over time.

$$\begin{aligned} S_{i,j} &= \beta + a_2 \times j + b_2 \times A_i + c_2(j \times A_i) + d_2 \times C_i \\ &+ e_2(j \times Ci) + f_2 \times E_i + g_2(j \times E_i) + h_2 \times O_i \\ &+ k_2(j \times O_i) + l_2 \times N_i + m_2(j \times N_i) + n_2 \times M_i \\ &+ o_2(j \times M_i) + p_2 \times TS_i + q_2(j \times TS_i) + x_i \end{aligned} \tag{3}$$

$$\begin{aligned} AG_{i,j} &= \beta + a_3 \times j + b_3 \times A_i + c_3(j \times A_i) \\ &+ d_3 \times C_i + e_3(j \times Ci) + f_3 \times E_i \\ &+ g_3(j \times E_i) + h_3 \times O_i + k_3(j \times O_i) \\ &+ l_3 \times N_i + m_3(j \times N_i) + n_3 \times M_i \\ &+ o_3(j \times M_i) + p_3 \times TS_i + q_3(j \times TS_i) + x_i \end{aligned} \tag{4}$$

In (3)-(4), $S_{i,j}$ and $AG_{i,j}$ denote the self-reported trust and agreement with AI of participant i at batch j, A_i, C_i, E_i, O_i, and N_i are the agreeableness, conscientiousness, extroversion, openness, and neuroticism characteristics of participant i, M_i is the overall propensity to trust machines of participant i, and TS_i reflects participant's i overall willingness to trust the AI captured via the Trust in AI Scenarios. In (3)-(4), $\{a_2, a_3\}$ are the fixed-effect coefficients which measure the evolution of trust measures over time, $\{b_2, b_3\}$, $\{d_2, d_3\}$, $\{f_2, f_3\}$, $\{h_2, h_3\}$, and $\{l_2, l_3\}$ are the fixed-effect coefficients that capture association between trust measures and personality traits, $\{n_2, n_3\}$ and $\{p_2, p_3\}$ are the fixed-effect coefficients that represent the association between trust measures and participant's overall trust propensity and perceived AI functionality, $\{c_2, c_3\}$, $\{e_2, e_3\}$, $\{g_2, g_3\}$, $\{k_2, k_3\}$, $\{m_2, m_3\}$, $\{o_2, o_3\}$, and $\{q_2, q_3\}$ are the fixed-effect coefficients that capture the interaction between time and participants' traits allowing us to examine whether the evolution of trust over time varies between people with different traits, and x_i is a random-effect coefficient. The results of the LME models defined in (3)-(4) are reported in Table II ($N = 184$). Results indicate that agreeableness and conscientiousness did not significantly moderate trust in AI. Extrovert participants agreed less with the AI compared to their counterparts ($f_3 = -0.042$, $p = 0.010$, $N = 184$). Participants who scored high in openness agreed more with the AI compared to their counterparts ($h_3 = 0.031$, $p = 0.048$, $N = 184$). Participants high in neuroticism depicted higher self-reported trust in AI ($l_2 = 0.33$, $p = 0.027$, $N = 184$), which decreased less with time ($m_2 = 0.12$, $p = 0.023$, $N = 184$), compared to their counterparts. Participants who perceived high functionality of the AI had higher self-reported trust compared to their counterparts ($p_2 = 0.37$, $p = 0.007$, $N = 184$).

VI. Discussion

In response to **RQ1**, self-reported trust and user agreement with AI do not vary across HP AI and LP AI conditions. A potential reason might be that the manipulation of AI performance was not strong enough. Due to the inherently complex nature of the data, the accuracy of the HP AI is around 60%, which perceptually might elicit low reliability. We anticipated that participants would perceive the large difference in deception sensitivity between the LP and HP conditions (i.e., 28% compared to 61%), but this did not appear to be an important deciding factor. In response to **RQ2**, self-reported trust and agreement with AI were correlated, similar to prior work [24]. In response to **RQ3**, our findings agree with prior work that indicates that one's overall predisposition to trust is associated with their momentary trust in and agreement with the automation. Similar to previous studies [11], agreeableness did not serve as a moderating factor of trust, potentially because this personality trait is related to participants' emotional and social functioning, which are not directly associated with the focal cognitive task. Openness was positively associated with trust; participants who score high in openness are open to new experiences, thus they are more likely to agree with the AI decision compared to their counterparts [9]. Our study depicted negative association between extroversion and agreement with the AI. Prior work stipulates a complex association between the two, since extrovert individuals might depict high levels of initial trust, which can decrease dramatically when the AI does not perform well [25]. Finally, prior work provides conflicting findings about conscientiousness [10], which was not a significant moderator of trust in our case.

This study presented the following limitations. The EBM model did not consider spectrotemporal speech variations that could be indicative of deception. Design factors pertaining to cognitive, emotional, and anthropometric characteristics of the AI have not been examined. Finally, the participants did not receive any feedback on their own performance which might have contributed to performance improvement over time.

TABLE II
LINEAR MIXED EFFECT (LME) MODEL ESTIMATES OF FIXED EFFECTS OF PERSONALITY AND TRUST TRAITS, AND THEIR INTERACTION WITH TIME, FOR ESTIMATING SELF-REPORTED AND BEHAVIORAL MEASURES OF TRUST IN AI.

Personality & trust traits	**Self-reported trust**		**Agreement with AI**	
	Fixed effect	**Interaction effect**	**Fixed effect**	**Interaction effect**
Time j	$a_2 = -0.27^{**}$		$a_3 = -0.03^{**}$	
Agreeableness A_i	$b_2 = 0.13$	$c_2 = 0.07$	$b_3 = 0.01$	$c_3 = 0.01$
Conscientiousness C_i	$d_2 = 0.18$	$e_2 = 0.05$	$d_3 = 0$	$e_3 = -0.01$
Extroversion E_i	$f_2 = 0.03$	$g_2 = 0.07$	$f_3 = -0.04^{**}$	$g_3 = -0.01$
Openness O_i	$h_2 = -0.06$	$k_2 = -0.01$	$h_3 = 0.03^{*}$	$k_3 = 0.01$
Neuroticism N_i	$l_2 = 0.33^{*}$	$m_2 = 0.12^{*}$	$l_3 = 0$	$m_3 = -0.02$
Propensity to Trust Machines M_i	$n_2 = 0.06$	$o_2 = 0$	$n_3 = -0.02$	$o_3 = 0$
Trust Scenarios TS_i	$p_2 = 0.37^{**}$	$q_2 = 0.04$	$p_3 = 0.01$	$q_3 = -0.01$

$^{**}, ^{*}$: $p <= 0.01$, $p <= 0.05$

VII. CONCLUSION

We examined a human-AI collaboration task for deception detection. Our results suggest a weak manipulation of the performance of the AI system. Self-reported and behavioral measures of trust were associated. A user's overall predisposition to trust the AI served as a significant factor of trust, but findings were conflicting in terms of personality. Openness and extroversion moderated one's trust, but did not depict consistent results across self-reported and behavioral metrics.

VIII. ETHICAL IMPACT STATEMENT

There are moral, legal, and social issues arising from the deception detection task, given its dependency on context and culture. Beyond questions pertaining to acceptable performance thresholds of deception detection technologies, it is important to define the framework within which these technologies operate ensuring that they abide with a society's ethical principles. With the transition of such technologies from the lab to real-life, it is imperative to think of how to balance the cost to the individual against the collective societal benefits (e.g., reduced crime/terrorism, improved health diagnostics). Societal beliefs pertaining to the right of non-self-incrimination, human free-will, and individual privacy and freedom will further shape the use of these technologies moving forward. The generalizability of results might be hindered by the small sample size, the relatively unbalanced female to male participant ratio, and the fact that audio from only Native English speakers was considered.

REFERENCES

[1] C. F. Bond Jr and B. M. DePaulo, "Accuracy of deception judgments," *Personality and social psychology Review*, vol. 10, no. 3, pp. 214–234, 2006.

[2] S. I. Levitan, G. An, M. Wang, G. Mendels, J. Hirschberg, M. Levine, and A. Rosenberg, "Cross-cultural production and detection of deception from speech," in *ACM on Workshop on Multimodal Deception Detection*, 2015, pp. 1–8.

[3] V. Lai and C. Tan, "On human predictions with explanations and predictions of machine learning models: A case study on deception detection," in *Proc. FAccT*, 2019, pp. 29–38.

[4] B. Kleinberg and B. Verschuere, "How humans impair automated deception detection performance," *Acta psychologica*, vol. 213, p. 103250, 2021.

[5] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, "Lying words: Predicting deception from linguistic styles," *Personality and social psychology bulletin*, vol. 29, no. 5, pp. 665–675, 2003.

[6] P. Ekman, M. O'Sullivan, W. V. Friesen, and K. R. Scherer, "Invited article: Face, voice, and body in detecting deceit," *Journal of nonverbal behavior*, vol. 15, no. 2, pp. 125–135, 1991.

[7] D. B. Buller and R. K. Aune, "Nonverbal cues to deception among intimates, friends, and strangers," *Journal of Nonverbal Behavior*, vol. 11, pp. 269–290, 1987.

[8] G. Bogaard, E. H. Meijer, A. Vrij, and H. Merckelbach, "Strong, but wrong: Lay people's and police officers' beliefs about verbal and nonverbal cues to deception," *PloS one*, vol. 11, no. 6, p. e0156615, 2016.

[9] M. Böckle, K. Yeboah-Antwi, and I. Kouris, "Can you trust the black box? The effect of personality traits on trust in ai-enabled user interfaces," in *Proc. AI-HCI*. Springer, 2021, pp. 3–20.

[10] R. Riedl, "Is trust in artificial intelligence systems related to user personality? Review of empirical evidence and future research directions," *Electronic Markets*, pp. 1–31, 2022.

[11] J. L. Szalma and G. S. Taylor, "Individual differences in response to automation: the five factor model of personality," *Journal of Experimental Psychology: Applied*, vol. 17, no. 2, p. 71, 2011.

[12] S. I. Levitan, A. Maredia, and J. Hirschberg, "Linguistic cues to deception and perceived deception in interview dialogues," in *Proc. NAACL*, 2018, pp. 1941–1950.

[13] P. Ekman, "Lying and nonverbal behavior: Theoretical issues and new findings," *Journal of nonverbal behavior*, vol. 12, pp. 163–175, 1988.

[14] S. Afroz, M. Brennan, and R. Greenstadt, "Detecting hoaxes, frauds, and deception in writing style online," in *2012 IEEE Symposium on Security and Privacy*. IEEE, 2012, pp. 461–475.

[15] C. K. Chung and J. W. Pennebaker, "Linguistic inquiry and word count (LIWC): pronounced "Luke,"... and other useful facts," in *Applied natural language processing: Identification, investigation and resolution*. IGI Global, 2012, pp. 206–229.

[16] C. Whissell, "Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language," *Psychological reports*, vol. 105, no. 2, pp. 509–521, 2009.

[17] X. Chen, S. Ita Levitan, M. Levine, M. Mandic, and J. Hirschberg, "Acoustic-prosodic and lexical cues to deception and trust: deciphering how people detect lies," *ACL TACL*, vol. 8, pp. 199–214, 2020.

[18] S. I. Levitan, A. Maredia, and J. Hirschberg, "Acoustic-prosodic indicators of deception and trust in interview dialogues." in *Interspeech*, 2018, pp. 416–420.

[19] P. Boersma and V. Van Heuven, "Speak and unSpeak with PRAAT," *Glot International*, vol. 5, no. 9/10, pp. 341–347, 2001.

[20] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker, "Accurate intelligible models with pairwise interactions," in *Proc. ACM SIGKDD*, 2013, pp. 623–631.

[21] O. P. John and S. Srivastava, "The Big-Five trait taxonomy: History, measurement, and theoretical perspectives," in *Handbook of personality: Theory and research*, L. A. Pervin and O. P. John, Eds. New York: Guilford Press, 1999.

[22] S. M. Merritt, H. Heimbaugh, J. LaChapell, and D. Lee, "I trust it, but i don't know why: Effects of implicit attitudes toward automation on trust in an automated system," *Human factors*, vol. 55, no. 3, pp. 520–534, 2013.

[23] R. Haydon, "Trust in artificial intelligence: How personality and risk experience affect human-ai relationships," Ph.D. dissertation, San Francisco State University, 2020.

[24] N. N. Sharan and D. M. Romano, "The effects of personality and locus of control on trust in humans versus artificial intelligence," *Heliyon*, vol. 6, no. 8, p. e04572, 2020.

[25] J. Elson, D. Derrick, and G. Ligon, "Examining trust and reliance in collaborations between humans and automated agents," in *Hawaii International Conference on System Sciences*, 2018, pp. 430–439.