Human Factors in Design, Engineering, and Computing, Vol. 159, 2024, 1406–1415
https://doi.org/10.54941/ahfe1005708

User Trust in a Depression Screening App When Outcomes Are Labelled as Either AI-Generated or Doctor-Generated: A Pilot Study

Yeganeh Shahsavar[1] and Avishek Choudhury[1,2]

[1]Industrial and Management Systems Engineering, West Virginia University, Morgantown, WV 26506, USA
[2]Occupational Safety and Health, West Virginia University, Morgantown, WV 26506, USA

ABSTRACT

This study aimed to quantify and compare user trust when interacting with a web-based depression screening app, where outcomes were labeled as either AI-generated or doctor-generated. The app calculated a depression score based on user input and presented two screening outcomes, one labeled as "doctor-generated" and the other as "AI-generated." Participants were then asked to select the outcome they trusted the most. Seventeen individuals participated in the study. Despite identical outcomes, 11 participants chose the AI-generated outcome (group-AI), while 6 selected the doctor-generated outcome (group-DR). To assess user trust (also attention), electroencephalogram (EEG) signals were recorded during the task, focusing on Alpha (Pz) and Beta (Fc1, Fc2) channels. Attention was measured through Alpha activity at Pz, while trust was assessed through Beta activity at Fc1 and Fc2. Post-intervention, participants' perceived trust in the outcomes was measured using a survey. The mean normalized power spectral density (PSD) values were calculated and correlated with the survey-based trust scores. Comparisons of PSDs and trust scores were made both between and within the AI and DR groups. Results showed that the mean PSD value for attention (Pz) was 0.116 $\mu V^2/Hz$, while the values for trust (Fc1 and Fc2) were 0.648 $\mu V^2/Hz$ and 0.646 $\mu V^2/Hz$, respectively. The mean trust score for the AI-based outcome was 3.118, compared to 3.235 for the doctor-based outcome. A weak to moderate correlation was observed between survey trust scores and PSD values in Fc1 and Fc2. Group-AI exhibited lower Alpha power at Pz (0.108 $\mu V^2/Hz$) and higher Beta power at Fc1 (0.660 $\mu V^2/Hz$) and Fc2 (0.659 $\mu V^2/Hz$) compared to group-DR, which showed higher Alpha power at Pz (0.131 $\mu V^2/Hz$) and lower Beta power at Fc1 (0.626 $\mu V^2/Hz$) and Fc2 (0.621 $\mu V^2/Hz$). In conclusion, our findings suggest that while participants may express marginal preference for doctor-generated outcomes in self-reported trust, their EEG data reveals a nuanced picture where those choosing AI-based outcomes may exhibit higher levels of trust on a cognitive level.

Keywords: User trust, Artificial intelligent, Depression, Digital health, Human factors

INTRODUCTION

Depression has emerged as a significant public health challenge, with particularly severe impacts on younger populations (Zhdanava et al., 2021, Bitsko, 2022). As traditional mental health services struggle with accessibility, digital mental health technologies have shown great potential to provide scalable solutions (Bakker et al., 2016, Torous et al., 2018, Chung et al., 2018, Ramos et al., 2019). Among these technologies, AI-based mental health applications have seen significant growth. Powered by large language models (LLMs), these tools are increasingly used for mental health screening, therapeutic interventions, and conversational support. However, as these AI-driven tools become more prevalent, it is crucial to understand how users perceive, and trust outcomes generated by AI compared to those generated by human doctors.

Trust in AI presents a double-edged sword. On one hand, users might develop a blind trust in AI-generated outcomes simply because of the perceived authority of AI, without critically evaluating the quality and reliability of the information provided. This could lead to over-reliance on AI, potentially neglecting the nuanced judgment that human clinicians can offer. On the other hand, under-trust in AI-generated outcomes solely because they are AI-based could prevent users from benefiting from innovative and effective solutions, especially if these tools have been rigorously validated and offer high-quality care (Choudhury and Chaudhry, 2024). Both blind trust and under-trust are undesirable, as they can respectively lead to misuse or underutilization of AI in mental health care (Choudhury, 2022). While there has been growing interest in assessing user attention and trust in digital health applications, research has largely relied on self-reported data to measure these constructs. Such perception based self-reported measures, while useful, are prone to biases such as social desirability and inaccurate self-assessment, which can distort the true relationship between perceived trust and actual behaviour. Thus, objective measures of trust, such as neurophysiological assessments, are needed to gain a deeper understanding of how users engage with AI-based mental health interventions.

The aim of this study is to quantify and compare user trust when interacting with a web-based depression screening app when outcomes are labelled as either AI-generated or doctor-generated. By using electroencephalography (EEG) to measure neurophysiological responses, this study seeks to provide a more objective assessment of trust, bridging the gap between perceived and actual user engagement with AI in mental health care.

METHOD

The study received ethical approval from the West Virginia University Institutional Review Board under protocol number 230782097. To conduct the research, we developed a web-based mental health application called “My Friendly Mind,” designed to screen for depression using self-reported data. Participants were instructed to use the app in a controlled environment, ensuring consistency across the experiment. During the session, participants answered nine questions derived from the patient health questionnaire, a widely used tool for assessing depression severity (Kroenke et al., 2001).

Based on their responses, the app generated a depression score for each participant. To assess trust and attention, the app presented participants with two identical explanations of their depression score. However, these explanations were randomly labeled as either an "AI-outcome" or a "doctor-outcome." Participants were then required to choose which result they trusted more. By keeping the outcomes identical, the study eliminates confounding factors related to the quality or accuracy of the diagnosis or recommendation.

Materials

Each participant used the app for approximately 30 minutes, during which they signed up, answered depression screening questions, and eventually reached the screening outcome page. Throughout the entire session, EEG signals were captured using a 32-channel water-based Bitbrain device configured according to the international 10–20 system. Electrodes were placed at the following positions: Fp1, Fp2, Fpz, AF3, AF4, F7, F3, Fz, F4, F8, FC5, FC1, FC2, FC6, T3-T7, C3, Cz, C4, T4-T8, CP5, CP1, CP2, CP6, P3, Pz, P4, T5-P7, Poz, T6-P8, O1, Oz, and O2. The reference was set to the average of the linked mastoid electrodes, and the ground electrode was positioned at AFz.

In this study, we specifically extracted and analysed EEG data only from the period when participants were on the screening outcome page, where they read their results and made a choice between the AI-generated and doctor-generated outcomes. Immediately following the experiment, participants completed a survey designed to assess their perceived trust in the screening results. The survey consisted of two questions, each rated on a Likert scale ranging from 0 (Not at all) to 4 (A lot), as detailed in Table 1.

Table 1. The trust survey questions.

Variable	Questions
Trust in AI outcome	How much do you trust the assessment outcome of your depression according to a validated and trained Artificial Intelligence System?
Trust in doctor outcome	How much do you trust the assessment outcome according to a Patient Health Questionnaire, a common tool used by doctors?

Data Analysis

Participants who selected AI generated outcome were termed as Group -AI and those who selected doctor-generated outcome were termed as Group-DR throughout this study. We used MATLAB to calculate the absolute Power Spectral Density (PSD) values for alpha (8-13 Hz) and beta (13-30 Hz) band frequencies from the 3 channels of Fc1, Fc2, and Pz. The EEG signal was filtered using a digital elliptic bandpass filter with a low cutoff frequency of 0.5 Hz and a high cutoff frequency of 35 Hz, involving an 8th-order low-pass and a 4th-order high-pass filter, both implemented using ellip function with a passband ripple of 0.1 dB and a stopband attenuation of 70 dB. We also applied the filtfilt function for zero-phase filtering, with a sampling rate of 256 Hz. After filtering, artifact removal was performed which excluded data segments exceeding three times the standard deviation or 90% of the maximum signal value, ensuring data clean from artifacts such as eye blinks and muscle movements for analysis. Power spectral density was estimated using the pwelch function (Welch, 1967), where a segment of the filtered signal was divided into overlapping windows. For each window, the PSD was computed using a Fast Fourier Transform. The absolute power values for the alpha and beta bands were then extracted and normalized to obtain relative power, facilitating the comparison of frequency activity across channels. We used JASP software (Love et al., 2019) to conduct correlation, Mann-Whitney U independent t-test and Wilcoxon signed rank paired t-test to compare the survey and PSD results between group-AI and group-DR.

RESULTS

Seventeen individuals participated in the study. Participants had a mean age of 24 years and 12 females. Out of 17 participants in the study, 11 were group-AI and 6 group-DR.

Quantitative Measures

User attention during the task was indicated by alpha power activity in the parietal region, with a mean normalized PSD value of 0.116 $\mu V^2/Hz$ at Pz, while trust was represented by frontocentral beta power activity, with mean normalized PSD values of 0.648 $\mu V^2/Hz$ at Fc1 and 0.646 $\mu V^2/Hz$ at Fc2. Figure 1 presents the topographic map of all 17 participants across 32 channels during the task.

The mean trust score for the AI-based outcome was 3.118, compared to 3.235 for the doctor-based outcome. A weak to moderate correlation was observed between trust in the AI-based outcome and the beta normalized PSD values at Fc1 ($r = 0.167, 0.217$) and Fc2 ($r = 0.118, 0.223$). Similarly, trust in the doctor-based outcome showed a weak to moderate correlation with Beta PSD values at Fc1 and Fc2, reflecting the complex relationship between perceived trust and neurophysiological indicators of trust.

Comparing User Trust and Attention

The mean normalized PSD values for Group AI were Pz = 0.108 $\mu V^2/Hz$, Fc1 = 0.660 $\mu V^2/Hz$, and Fc2 = 0.659 $\mu V^2/Hz$, while for Group DR, the mean PSD values were Pz = 0.131 $\mu V^2/Hz$, Fc1 = 0.626 $\mu V^2/Hz$, and Fc2 = 0.621 $\mu V^2/Hz$. Figure 2 illustrates the comparison of these PSD mean values between Group AI and Group DR. Statistical analysis revealed no significant differences between the groups for Pz ($P = 0.180$), Fc1 ($P = 0.884$), and Fc2 ($P = 0.808$).

Figure 3 shows the difference between group-AI and group-DR normalized PSD values. The differences indicate lower alpha power and higher beta power activities in group-AI compared to group-DR. However, these different were not statistically significant.

Table 2 summarizes the descriptive statistics of the participants' perceived trust levels in AI-based and doctor-based outcomes for both group-AI and group-DR. Participants in group-DR (those who selected the doctor-generated outcome) showed higher trust in doctor-based outcomes

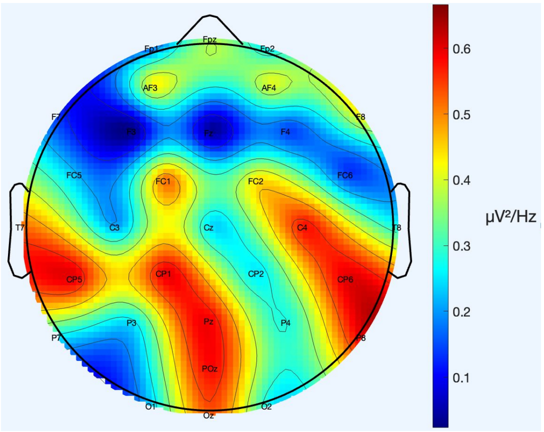

Figure 1: Mean power spectral density (PSD) across 32 EEG channels during the task. The colour scale represents PSD values in microvolts squared per hertz (μV^2/Hz), with warmer colours (red) indicating higher PSD values and cooler colours (blue) indicating lower PSD values. The map illustrates the distribution of Alpha and Beta power across the scalp, with notable activity observed in frontocentral and parietal regions, reflecting user attention and trust during interaction with the depression screening app.

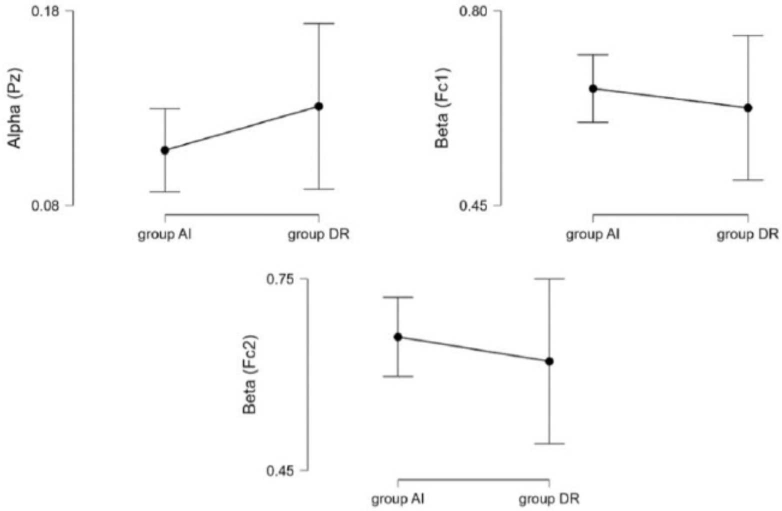

Figure 2: The mean of normalized PSD values for Alpha (Pz), Beta (FC1, FC2) activity between participants in Group AI and Group DR during the task.

(Mean = 3.333) compared to AI-based outcomes (Mean = 2.667). Participants in group-AI (those who selected the AI-generated outcome) displayed relatively balanced trust levels between AI-based outcomes (Mean = 3.364) and doctor-based outcomes (Mean = 3.182).

Table 3 presents the results of paired t-tests comparing trust in AI-based outcomes versus doctor-based outcomes within each group (group-AI and group-DR). For group-DR, there was a marginally non-significant difference in trust between AI-based and doctor-based outcomes, suggesting a trend towards higher trust in doctor-based outcomes. Similarly, for group-AI, no

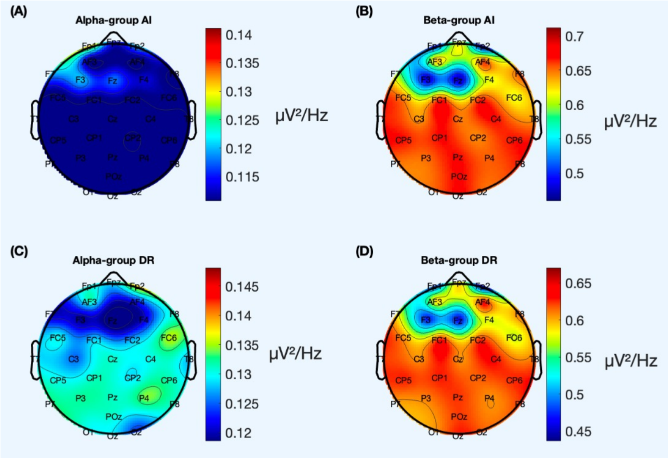

Figure 3: Mean normalized power spectral density (PSD) in microvolts squared per hertz (μV^2/Hz) for alpha and beta frequency bands across two participant groups: Group AI and Group DR. **(A)** Alpha-band activity for Group AI; **(B)** Beta-band activity for Group AI; **(C)** Alpha-band activity for Group DR; **(D)** Beta-band activity for Group DR. The colour scale represents the PSD values, with warmer colours indicating higher power and cooler colours indicating lower power.

Table 2. The question's descriptive for group-AI and group-DR.

Measures		Descriptive	
Group	Question	Mean (SD)	SE
DR	Trust in AI-based outcome	2.667 (0.516)	0.211
	Trust in doctor-based outcome	3.333 (0.516)	0.211
AI	Trust in AI-based outcome	3.364 (0.809)	0.244
	Trust in doctor-based outcome	3.182 (0.751)	0.226

SD: standard deviation; SE: standard error

significant difference was found between trust in AI-based and doctor-based outcomes, indicating that participants in this group viewed both outcomes similarly in terms of trust.

Table 3. Comparing trust in AI-based outcome, doctor-based outcome and the app for group-AI and group-DR separately.

Measures		Paired t-test
Group	Questions	(P value)
DR	Trust in AI-based outcome ~ Trust in doctor-based outcome	0.072
AI	Trust in AI-based outcome ~ Trust in doctor-based outcome	0.530

Table 4 shows the results of independent t-tests comparing trust levels between group-AI and group-DR for each trust measure. The comparison between groups for trust in AI-based outcomes approached significance, suggesting a potential difference in trust levels between group-AI and group-DR, with group-AI tending to trust AI-based outcomes more. No significant difference was found between the groups regarding trust in doctor-based outcomes, indicating that both groups had similar levels of trust in doctor-generated outcomes.

Table 4. Comparing group-AI and group-DR trust in AI-based outcome, doctor-based outcome and the app.

	Measures	Independent t-test
Group	Questions	(P value)
AI ~ DR	Trust in AI-based outcome	0.076
	Trust in doctor-based outcome	0.780

DISCUSSION

Our study provides insights into the cognitive (trust and attention) involved in selecting between AI-generated and doctor-generated screening outcomes. Despite receiving the same screening outcome, a disparity emerged in their preferences, with 11 participants choosing the AI-generated outcome (group-AI) and 6 opting for the doctor-generated outcome (group-DR). This observation is consistent with findings from (Kerstan et al., 2023), who notes that individuals often rely on implicit associations with AI when making trust-related decisions, particularly in healthcare contexts. Furthermore, the concept of calibrated trust, as discussed by (Naiseh et al., 2021), indicates that users may misjudge their trust in AI systems, which can lead to either over-trust or under-trust in AI-generated recommendations (Choudhury, 2022). This miscalibration could explain the preference divergence observed in our study, where participants may have perceived AI-generated outcomes as more trustworthy despite identical results.

Moreover, our EEG analyses, specifically focusing on the alpha (Pz) and beta (Fc1, Fc2) channels, were recorded during the task to assess attention and trust, respectively. The alpha band at Pz (8-12 Hz) is commonly associated with attention levels (Klimesch et al., 1998, Ray and Cole, 1985), while the Beta band (13-30 Hz) at Fc1 and Fc2 is linked to cognitive processes related to trust (Wang et al., 2021, Wang et al., 2018). The analysis of the mean normalized PSD values in these channels revealed subtle yet noteworthy differences between the groups. For group-AI, the mean PSD values were 0.108 $\mu V^2/Hz$ at Pz, 0.660 $\mu V^2/Hz$ at Fc1, and 0.659 $\mu V^2/Hz$ at Fc2. These values suggest that participants who selected the AI-generated outcome exhibited marginally higher attention levels (as indicated by lower Pz values) and higher trust (as indicated by higher Fc1 and Fc2 values) compared to the group that selected the doctor-generated outcome. In contrast, group-DR showed lower attention (Pz = 0.131 $\mu V^2/Hz$) and slightly lower trust levels (Fc1 = 0.626 $\mu V^2/Hz$, Fc2 = 0.621 $\mu V^2/Hz$). These differences imply that participants in the AI group may have been more attentive or more deliberate in their trust allocation, possibly influenced by the hype and perceived authority of AI-generated outcomes. The findings also resonate with the literature suggesting that individuals may exhibit differential cognitive engagement based on the source of information, whether AI or human (Kerstan et al., 2023, Chen and Park, 2021).

The post-intervention survey further elucidates EEG findings, with trust scores providing a subjective measure of the participants' trust in the selected outcomes. The mean trust score for the AI-generated outcome was 3.118, while for the doctor-generated outcome, it was 3.235. Although the difference in trust scores is modest, it is indicative of a slightly higher trust in doctor-generated outcomes. However, this is juxtaposed with the EEG data, where group-AI exhibited higher trust-related PSD values, suggesting that those who opted for the AI outcome may have had a more cognitively robust trust in the AI, even if their self-reported trust was marginally lower. The discrepancy between subjective trust scores and objective EEG measures aligns with findings from (Montag et al., 2023), who noted that trust in AI and trust in human agents are not necessarily correlated, indicating that individuals may cognitively process their trust in AI differently than in human interactions. This suggests that while participants may consciously express a preference for human-generated outcomes, their cognitive engagement with AI may reflect a deeper, albeit more complex, trust relationship.

The correlation between survey trust scores and PSD values in the beta channels (Fc1 and Fc2) was found to be weak to moderate, further supporting the complex and multifaceted nature of trust in AI versus human experts. This correlation indicates that while EEG-derived trust measures and self-reported trust are related, they are not perfectly aligned, highlighting the importance of using both objective and subjective measures in understanding trust dynamics. Discrepancies between self-reported trust measures and objective behavioural indicators has also been acknowledged by (Hancock et al., 2011) who discussed the inconsistencies between individuals' perceptions of trust and their observable reactions. Similarly, (Rieger et al., 2023)'s findings indicate that trust attitudes and behaviours do not always correlate perfectly.

Furthermore, the relationship between EEG-derived trust measures and self-reported trust is indicative of the broader discourse surrounding trust calibration in AI contexts. (Kaplan et al., 2021) argue that trust in AI is influenced by various factors, including prior experiences and contextual cues, which can lead to discrepancies between perceived and actual trust levels. This is echoed by (Meimandi et al., 2024), who notes that traditional self-reported measures of trust often fail to account for the evolving and context-dependent nature of trust in AI systems.

Limitations

This study has some limitations that should be considered when interpreting the findings. First, the sample size was relatively small, with only 17 participants, all of whom were students with a mean age of 24 years. This homogeneity in age and background may limit the generalizability of the results to broader populations, particularly those with diverse demographics and clinical backgrounds. Second, the study relied on a controlled experimental setting, which may not fully replicate real-world conditions under which users typically interact with AI-based mental health applications. Finally, the study used identical outcomes labelled as either AI-generated or doctor-generated, which might not entirely reflect the variability and complexity of AI or clinician-generated diagnoses. While this approach was necessary for the experimental design, it may not capture how users would respond to actual differences in AI and doctor-generated outcomes. The design focuses solely on labelling effects and does not provide insight into how participants would weigh the quality, reasoning, or explanations behind AI-generated and doctor-generated outcomes.

CONCLUSION

As AI continues to play an increasingly prominent role in mental health care, it is essential to foster balanced and informed trust among users. This requires not only ensuring the reliability and transparency of AI systems but also understanding the cognitive and emotional factors that drive user trust. Future research should continue to explore these dynamics in larger, more diverse populations and in real-world settings, to better inform the development of AI-based interventions that are both effective and trusted by those who use them.

ACKNOWLEDGMENT

This study was supported by a seed grant from West Virginia University (#3086). We are grateful for the financial assistance provided, which made this research possible.

REFERENCES

Bakker, D., Kazantzis, N., Rickwood, D. & Rickard, N. 2016. Mental health smartphone apps: review and evidence-based recommendations for future developments. *JMIR mental health,* 3, e4984.

Bitsko, R. H. 2022. Mental health surveillance among children—United States, 2013–2019. *MMWR supplements,* 71.

Chen, Q. Q. & Park, H. J. 2021. How Anthropomorphism Affects Trust in Intelligent Personal Assistants. *Industrial Management & Data Systems,* 121, 2722–2737.

Choudhury, A. 2022. Toward an Ecologically Valid Conceptual Framework for the Use of Artificial Intelligence in Clinical Settings: Need for Systems Thinking, Accountability, Decision-making, Trust, and Patient Safety Considerations in Safeguarding the Technology and Clinicians. *JMIR Hum Factors,* 9, e35421.

Choudhury, A. & Chaudhry, Z. 2024. Large Language Models and User Trust: Consequence of Self-Referential Learning Loop and the Deskilling of Health Care Professionals. *J Med Internet Res,* 26, e56764.

Chung, K., Jeon, M.-J., Park, J., Lee, S., Kim, C. O. & Park, J. Y. 2018. Development and evaluation of a mobile-optimized daily self-rating depression screening app: A preliminary study. *PloS one,* 13, e0199118.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., Visser, E. J. D. & Parasuraman, R. 2011. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors the Journal of the Human Factors and Ergonomics Society,* 53, 517–527.

Kaplan, A. D., Kessler, T. T., Brill, J. C. & Hancock, P. A. 2021. Trust in Artificial Intelligence: Meta-Analytic Findings. *Human Factors the Journal of the Human Factors and Ergonomics Society,* 65, 337–359.

Kerstan, S., Bienefeld, N. & Grote, G. 2023. Choosing Human Over AI Doctors? How Comparative Trust Associations and Knowledge Relate to Risk and Benefit Perceptions of AI in Healthcare. *Risk Analysis,* 44, 939–957.

Klimesch, W., Doppelmayr, M., Russegger, H., Pachinger, T. & Schwaiger, J. 1998. Induced alpha band power changes in the human EEG and attention. *Neuroscience letters,* 244, 73–76.

Kroenke, K., Spitzer, R. L. & Williams, J. B. 2001. The PHQ-9: validity of a brief depression severity measure. *Journal of general internal medicine,* 16, 606–613.

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, J., Ly, A., Gronau, Q. F., Šmíra, M. & Epskamp, S. 2019. JASP: Graphical statistical software for common statistical designs. *Journal of Statistical Software,* 88, 1–17.

Meimandi, K. J., Bolton, M. L. & Beling, P. A. 2024. Action Over Words: Predicting Human Trust in AI Partners Through Gameplay Behaviors.

Montag, C., Klugah-Brown, B., Zhou, X., Wernicke, J., Liu, C., Kou, J., Chen, Y., Haas, B. W. & Becker, B. 2023. Trust Toward Humans and Trust Toward Artificial Intelligence Are Not Associated: Initial Insights From Self-Report and Neurostructural Brain Imaging. *Personality Neuroscience,* 6.

Naiseh, M., Cemiloglu, D., Al-Thani, D., Jiang, N. & Ali, R. 2021. Explainable Recommendations and Calibrated Trust: Two Systematic User Errors. *Computer,* 54, 28–37.

Ramos, R. M., Cheng, P. G. F. & Jonas, S. M. 2019. Validation of an mHealth app for depression screening and monitoring (psychologist in a pocket): Correlational study and concurrence analysis. *JMIR mHealth and uHealth,* 7, e12051.

Ray, W. J. & Cole, H. W. 1985. EEG activity during cognitive processing: influence of attentional factors. *International Journal of Psychophysiology,* 3, 43–48.

Rieger, T., Kugler, L., Manzey, D. & Roesler, E. 2023. The (Im) perfect Automation Schema: Who Is Trusted More, Automated or Human Decision Support? *Human Factors the Journal of the Human Factors and Ergonomics Society,* 66, 1995–2007.

Torous, J., Wisniewski, H., Liu, G. & Keshavan, M. 2018. Mental health mobile phone app usage, concerns, and benefits among psychiatric outpatients: comparative survey study. *JMIR mental health,* 5, e11715.

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

Wang, Y., Yang, X., Tang, Z., Xiao, S. & Hewig, J. 2021. Hierarchical neural prediction of interpersonal trust. *Neuroscience Bulletin,* 37, 511–522.

Welch, P. 1967. The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics,* 15, 70–73.

Zhdanava, M., Pilon, D., Ghelerter, I., Chow, W., Joshi, K., Lefebvre, P. & Sheehan, J. J. 2021. The prevalence and national burden of treatment-resistant depression and major depressive disorder in the United States. *The Journal of Clinical Psychiatry,* 82, 29169.