Understanding Linguistic and Visual Factors that Affect Human Trust Perception of Virtual Agents

Natalia Tyulina
CUNY Graduate Center
New York, USA
ntyulina@gradcenter.cuny.edu

Tatiana Aloi Emmanouil
Baruch College
New York, USA
tatiana.emmanouil@baruch.cuny.edu

Sarah Ita Levitan
Hunter College
New York, USA
sarah.levitan@hunter.cuny.edu

ABSTRACT

This work investigates how visual and spoken cues of virtual agents interact to affect user perception of agent trustworthiness. It is directly motivated by practical applications, such as an assistive robot companion for the elderly or homebound, or a virtual agent that can provide psychological assessment and treatment for individuals with mental health challenges. Such technologies have the capacity to assist human users in impactful ways, but without human trust in these systems, adoption and usage will remain severely limited. Our findings reveal strong correlations between both visual and auditory features and perceived trustworthiness. This underscores the importance of incorporating a comprehensive range of nonverbal cues and auditory signals into interface design.

CCS CONCEPTS

• **Human-centered computing** → **User centered design**; **Human computer interaction (HCI)**; **Empirical studies in HCI**;

KEYWORDS

Conversational Agents, Auditory, Visual, Trust Cues, Human Perception, Multimodal

ACM Reference Format:
Natalia Tyulina, Tatiana Aloi Emmanouil, and Sarah Ita Levitan. 2024. Understanding Linguistic and Visual Factors that Affect Human Trust Perception of Virtual Agents. In *ACM Conversational User Interfaces 2024 (CUI '24), July 08–10, 2024, Luxembourg, Luxembourg*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3640794.3665581

1 INTRODUCTION

We are rapidly approaching a future in which conversational agents, including chatbots, virtual assistants, and robots with dialogue capabilities, are becoming increasingly integrated into our daily lives. Advances in machine learning and speech technologies are enabling human-like conversations with virtual agents, gaining traction across various domains such as customer service, healthcare, and education [2, 12, 15]. As these agents aim to simulate human-like conversation, a crucial factor that profoundly influences user engagement and satisfaction is trust [6, 7]. Users must feel confident in the reliability, credibility, and trustworthiness of the information and responses provided by these agents. Researchers across various disciplines have explored identifying specific signals of trust, examining both nonverbal and verbal cues to determine trustworthiness [11]. Recent studies have also highlighted the role of vocal dynamics, such as vocal pitch and temporality, in shaping trust perception[4]. However, there is little interdisciplinary work that systematically investigates how combinations of factors from different modalities affect user trust. As systems are increasingly multimodal, it is critical to understand how multimodal factors interact with each other to affect user perception of agent trustworthiness. The main objective of this work is to understand linguistic and visual factors that affect human trust perception of virtual agents.

CUI '24, July 08–10, 2024, Luxembourg, Luxembourg

ACM ISBN 979-8-4007-0511-3/24/07
https://doi.org/10.1145/3640794.3665581

2 METHODOLOGY

Our approach focuses on assessing how individuals formulate rapid intuitive judgments in contexts where information is limited. This deliberate choice allows us to explore initial impressions made solely based on voice clips and static images. To achieve our research objective, we conducted an IRB approved perception study using multimodal stimuli of combined visual and auditory modalities. Amazon Mechanical Turk (MTurk) was used to employ 150 participants for the first set of experiments. MTurk has been regarded as a reliable source capable of producing quality data representative of various U.S. subpopulations [5, 8]. The participants were presented with static faces and corresponding speech recordings and were asked to provide their judgments on the perceived trustworthiness of the face and voice using a Likert scale ranging from 1 (not at all trustworthy) to 5 (extremely trustworthy).

The faces were sampled from the 10k US Adult Faces Database [1], containing natural face photographs rated for a number of attributes including trustworthiness. The speech samples were drawn from a large corpus of deceptive and truthful dialogues that have also been rated for trustworthiness in a large-scale perception study [3]. Although the faces and speech samples have been previously independently rated for trustworthiness, they have not been used to study the perception of trustworthiness in a multimodal setting. Stimuli were created with different combinations of trust levels between the visual and speech modalities, both congruent and incongruent.

2.1 Data

The lowest and the highest-rated images and audio fragments were extracted from the two datasets. For the images, the means across 15 raters' scores were used to select 80 female and 80 male images

Table 1: Survey Conditions

Condition	Number of Trials
high trust image & high trust audio	40
high trust image & low trust audio	40
low trust image & high trust audio	40
low trust image & low trust audio	40

deemed most and least trustworthy. We followed a similar procedure for the audio selection, where the highest and lowest trust ratings were represented by a vector of three trust labels, one per annotator. The highest trust audio samples received a score of 1 from all three annotators, whereas the lowest trust ones received all 0 scores.

Furthermore, we applied a number of additional filtering criteria. Audio fragments were filtered based on all utterances being spoken in Standard American English, with the total number of content words equal to or greater than 2, and each fragment's duration limited to a maximum of 8 seconds. Regarding the images, we excluded the highest age group to account for potential perception mismatches with audio fragments, as they were drawn from a corpus not representative of this particular age group. Finally, a total of 160 unique pairs were obtained. All images and audio fragments were then reshuffled to form new pairs. The reshuffling was performed to avoid possible image-audio matching biases. Both versions of unique pairings were broken down into two parts with 80 pairs each, and all of them were used in the perception experiments.

2.2 Survey Design

The survey implementation utilized Qualtrics, integrated into MTukr, for efficient distribution to a participant pool. Additional measures included incorporating qualifications for the workers, such as near-native understanding of the English language, an overall approval rating higher than 95%, and not having participated in any of the previous batches of this study.

The survey design was structured around four distinct conditions, with each condition comprising 40 trials. These trials were equally divided between sexes associated with each image-audio pair to ensure balance and mitigate bias. The trials were organized into two batches to facilitate manageable participant engagement and prevent fatigue. Each trial presented participants with a unique combination of an image and a corresponding audio fragment. Table 1 summarizes all conditions employed in this survey.

Figure 1 illustrates the survey instructions, and Figure 2 depicts one of the survey questions, demonstrating our user interface decisions.

We opted to display only one pair of image and audio stimuli per page, with the "back" button disabled. This deliberate choice aimed to optimize participants' focus on each multimodal sample. Additionally, we programmed audio playback to start upon image loading, minimizing any perceptual delay between visual and auditory stimuli. Participants were instructed to complete the survey within a thirty minute timeframe and to utilize a laptop or desktop computer, as opposed to a mobile device, for optimal survey experience.

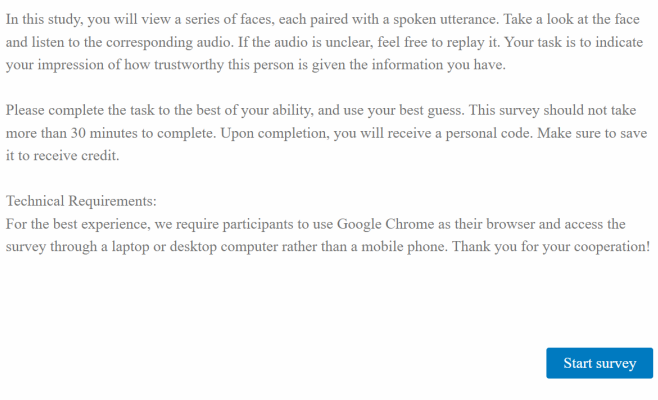

Figure 1: Survey Instructions

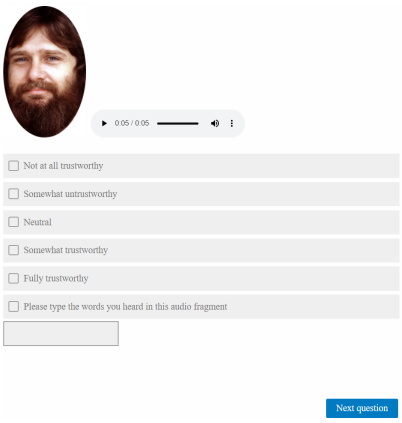

Figure 2: Sample Trial

Participants were presented with 80 image-audio pairs per experiment and were asked to provide their judgments about how trustworthy each individual is. An informed consent form and a set of post-survey questions were also included. An open-ended text entry question was randomly added to two trials for validation purposes, prompting participants to transcribe the content of a given audio fragment.

3 PRELIMINARY RESULTS

The preliminary analysis involved several steps to ensure data integrity and accuracy in assessing the impact of image and audio conditions, as well as sex, on perceived trustworthiness scores. Below is a detailed presentation of the results, followed by their implications.

3.1 Data Pre-Processing

Initially, the Shapiro-Wilk test conducted on the raw data indicated a violation of normality assumptions. To address this, standardized z-scores were adopted for subsequent analyses. These z-scores represent the disparity of individual observations from the mean, measured in standard deviation units, thus providing a standardized

Table 2: Summary Statistics

Image	Audio	Sex	Mean	SD
high	high	female	.235	.325
high	high	male	.202	.349
high	low	female	.068	.333
high	low	male	.013	.370
low	high	female	.003	.362
low	high	male	−.157	.397
low	low	female	−.096	.319
low	low	male	−.268	.434

metric of participants' responses across experimental conditions. The obtained z-score values ranged from $[-1.180, 1.038]$.

Figure 3 demonstrates that no extreme outliers were detected. The boxplot reveals a consistent trend: females generally receive higher scores than males across all conditions, with the upper limit of scores being higher for females. Additionally, males tend to exhibit lower scores, particularly in low image conditions. This suggests that low trust images have a stronger impact on perceived lack of trust compared to low trust audio.

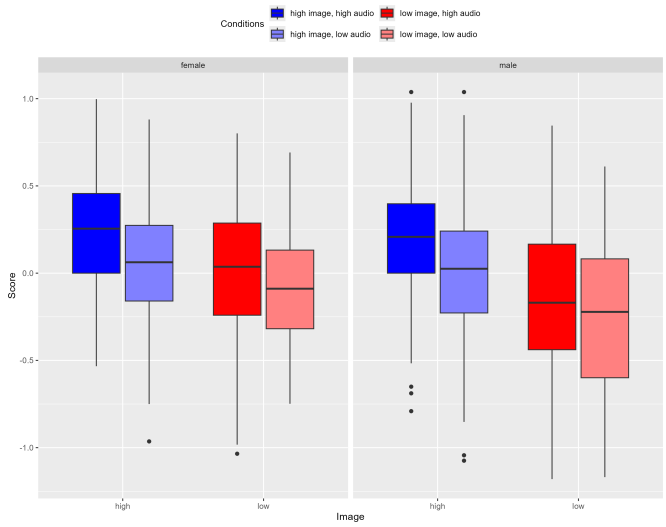

Figure 3: Representation of Scores by Image and Audio Conditions, Faceted by Sex

Table 2 provides detailed statistics regarding the distribution of z-scores. The wider disparity between male and female scores in low-trust image conditions suggests that the depicted sex may influence perceived trustworthiness differently across conditions.

Figure 4 further explores the mean distribution per condition, focusing on understanding differences between pairs with varying trust levels between image and audio modalities. Pairs characterized by high image but low audio trust demonstrate higher mean scores compared to pairs with the opposite configuration, suggesting a significant interaction effect between image and audio trust levels on perceived trustworthiness.

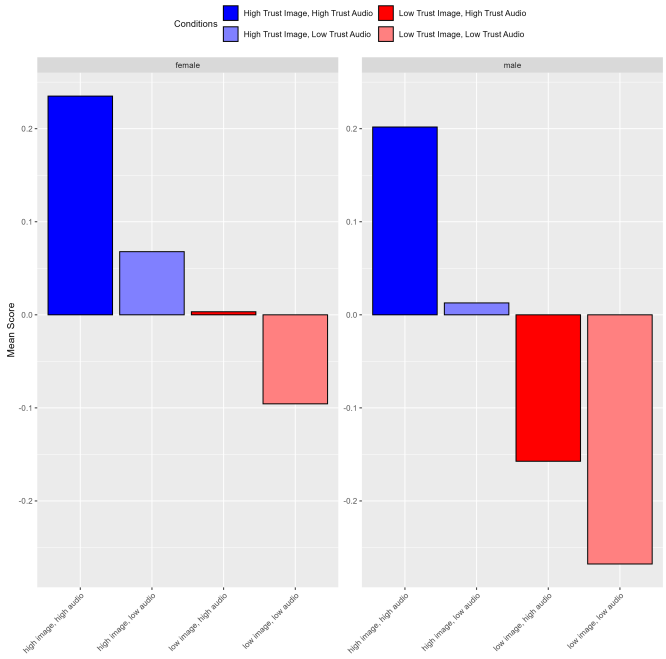

Figure 4: Mean Distribution per Condition

Table 3: ANOVA Results

Effect	*DFn*	*DFd*	*F*	*p*	$p < .05$	*ges*
image	1	149	57.383	< .001	*	.113
audio	1	149	40.788	< .001	*	.037
sex	1	149	21.965	< .001	*	.021
image:audio	1	149	4.216	.04	*	.003
image:sex	1	149	10.966	.001	*	.007
audio:sex	1	149	.192	.662		.000
image:audio:sex	1	149	.021	.886		.000

3.2 ANOVA Results and Interaction Effects

Table 3 summarizes the results of repeated measures 3-factor ANOVA. The analysis revealed statistically significant two-way interactions between image and sex ($F(1, 149) = 10.996, p = .001$), and image and audio ($F(1, 149) = 4.216, p = .04$). Significant main effects were observed for image ($F(1, 149) = 57.383, p < .001$), audio ($F(1, 149) = 40.788, p < .001$), and sex ($F(1, 149) = 21.965, p < .001$). The absence of a significant three-way interaction implies that the joint effects of image, audio, and sex on perceived trustworthiness do not operate multiplicatively. The significant two-way interactions underscore nuanced relationships between specific pairs of factors, with the image main effect exhibiting the highest effect size, highlighting the critical role of image trustworthiness in shaping overall perceptions.

Figure 5 illustrates the interaction effects between three pairs of factors: image and audio, image and sex, and audio and sex.

The left plot shows significant main effects of image and audio conditions on mean scores, with non-parallel lines indicating an

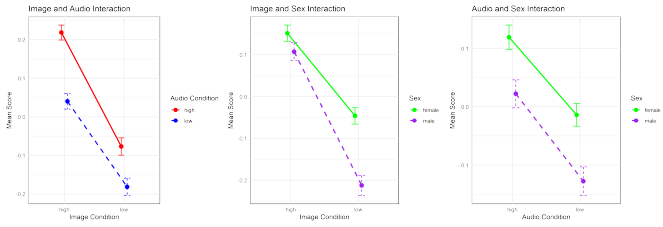

Figure 5: Interaction Line Plots

interaction effect. The difference in mean scores between high and low image conditions is greater for the high audio condition, suggesting a more pronounced impact of image level when audio level is high. The middle plot depicts significant main effects of image and sex, with females generally having higher mean scores than males. The non-parallel lines indicate an interaction effect, with a steeper slope for females suggesting a stronger impact of image condition on their mean scores. The right plot shows significant main effects of audio and sex, with similar slopes but slight differences in intercepts indicating a minor, non-significant interaction. The small distances between the lines suggest minor variations in the effect of audio condition between sexes.

3.3 Feature Correlations with Trust Scores

We also computed correlations between features extracted from original image-only and audio-only corpora and the mean trustworthiness scores associated with each image-audio pair. We identified a number of features that were statistically significantly correlated with the obtained trustworthiness scores. Figure 6 illustrates both positively and negatively correlated features. Specifically, pairs characterized by high image trust but low audio trust demonstrate markedly higher mean scores compared to pairs with the opposite configuration. This suggests a significant interaction effect between image and audio trust levels on perceived trustworthiness.

Features such as *kind*, *sociable*, *friendly*, *caring*, and *calm* exhibited strong positive correlations with perceived trustworthiness. These attributes were previously identified by human subjects in a separate study[1], rather than directly extracted from the images themselves. Similarly, attributes like *irresponsible*, *unintelligent*, *unattractive*, *unfriendly*, and *emotionally unstable* showed strong negative correlations with trustworthiness scores. Interestingly, features related to emotional stability and warmth, as well as perceived intelligence and attractiveness, emerged as key predictors of perceived trustworthiness in visual cues, highlighting the importance of nonverbal cues and personality traits in shaping trust perceptions in combined visual and auditory domains.

Regarding statistically significantly correlated auditory features, Table 4 lists them along with their more detailed descriptions. Features such as mean energy and maximum energy, indicative of the overall loudness levels in speech, exhibited positive correlations with trustworthiness scores. Similarly, the speed at which a person speaks, captured by the speaking rate feature, showed a positive correlation with trustworthiness, implying that a moderate speaking pace may contribute to higher levels of perceived trust. Conversely, features related to the presence of filler words and response latency

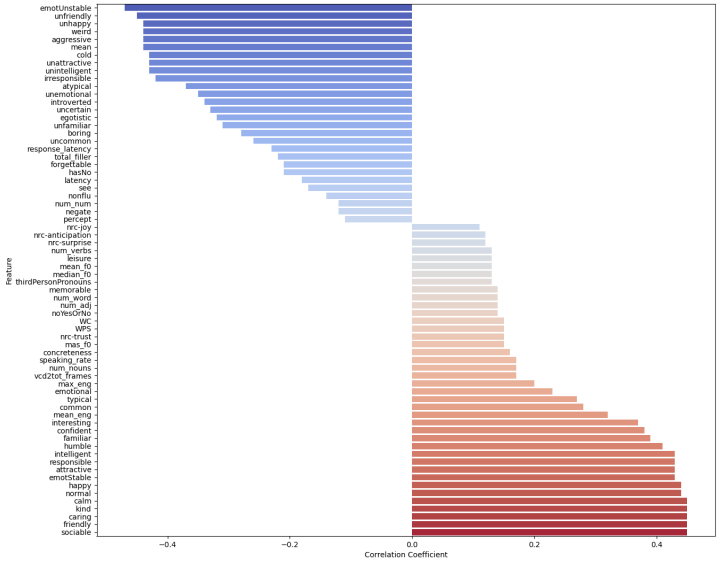

Figure 6: Feature Correlations with Trust Scores

Table 4: Auditory Features Significantly Correlated with Multimodal Trust Ratings

Feature	Description	*r*
mean_eng	mean value of energy (loudness)	.32
max_eng	max value of energy (loudness)	.20
speaking rate	speed at which a person speaks	.17
vcd2tot_frames	proportion of speech versus silence	.17
num_nouns	number of nouns used	.17
mas_f0	mass value of pitch frequency	.15
response_latency	time span before the first word	−.23
total_filler	number of filler words	−.22
hasNo	presence of the word *no*	−.21
latency	delay before a person starts speaking	−.18
nonflu	presence of nonfluencies	−.14
concreteness	a measure of the detail	−.16
negate	presence of negation markers	−.12

demonstrated negative correlations with trustworthiness scores. A higher frequency of filler words and longer response latency may signal hesitancy or lack of confidence, potentially undermining trust perceptions. Additionally, the presence of negation, as indicated by the *hasNo* and *negation* features, was negatively correlated with trustworthiness.

Furthermore, features such as *nonflu* (presence of nonfluencies) and *concreteness* (indicator of the level of details of the speakers' visual and haptic experiences) also exhibited negative correlations with trustworthiness scores. This implies that the presence of

speech disfluencies or lack of specific details in verbal communication may lead to lower perceived trustworthiness. Additionally, the *mas_f0* feature, representing the mass value of pitch frequency, showed a positive correlation with trustworthiness scores, suggesting that individuals with a more varied pitch range may be perceived as more trustworthy. These findings further emphasize the importance of considering a wide range of auditory cues, including linguistic characteristics and speech patterns, in understanding and designing trustworthy conversational interfaces.

4 DISCUSSION AND FUTURE WORK

Building upon these findings, future experiments will explore strategies to mitigate the overwhelming effect of images on perception. One such approach involves conducting experiments where audio stimuli are presented before the corresponding image is displayed, aiming to counterbalance the pronounced impact of images. By manipulating the temporal order of stimulus presentation, we aim to investigate whether priming participants with audio cues can attenuate the strong influence of subsequent visual information. Future work will also involve conducting baseline experiments to better understand the independent effects of each modality. These experiments could include variations in image and audio features to assess their impact on perceived trustworthiness independently.

Further statistical analyses will be conducted to identify more fine-grained features associated with user trust perception. Acoustic-prosodic features will include specific parameters related to pitch, speaking rate, intensity, and voice quality measures, and visual features will involve facial features such as distance between the eyes, and face height to width ratio. These features have been previously found to predict perceived trustworthiness [3, 9, 10, 13]. Additional linguistic features, such as specific syntactic patterns and lexical cues will also be examined more closely.

Other future research avenues may involve extending this work to experiment with video interfaces rather than static face images. By incorporating dynamic visual and auditory cues into experimental stimuli, we will attempt to simulate real-world social interactions more closely, including the effect of multimodal emotional expression that proved to be a significant contributor in prior research [14]. Additionally, we acknowledge the importance of studying trust judgments in live interactions, and we are interested in exploring how these perceptions might be studied in in-person experiments.

Ordinal logistic regression models will be used to evaluate the relationship between these multimodal features and the ordinal trust labels from baseline experiments. Machine learning models will be trained to predict user trust ratings based on speech and visual features and the most important features for prediction will be identified. This research will further test whether trust perception in multimodal conditions involves a simple additive combination of visual and auditory trust perception, or whether there are more complex interactions between multimodal cues that contribute to the overall perception of trustworthiness.

It is crucial to recognize the potential harms that can arise from manipulating trust-inducing features in voices and face images for personal gain. Ensuring that the trustworthiness of the application we design aligns with the actual transparency and intentions of the system is paramount. This alignment is necessary to maintain genuine user trust and prevent exploitation. Ethical considerations and regulatory oversight must guide the development of these technologies to protect users and foster a trustworthy digital environment.

5 CONCLUSION

Understanding the linguistic and visual factors that affect human trust perception of virtual agents can inform the design of more effective and trustworthy conversational user interfaces. This research can help develop guidelines and best practices for designing virtual agents that inspire trust and confidence in users. It is essential to ensure that the trustworthiness conveyed by the application aligns with the transparency and integrity of the system itself. Moreover, understanding these trust-inducing features can guide the HCI community, particularly CUI researchers, in creating interfaces that foster genuine user trust while safeguarding against potential misuse. The strong correlations observed between specific features, both visual and auditory, and perceived trustworthiness highlight the importance of incorporating a comprehensive range of nonverbal cues and auditory signals into interface design. Interfaces that effectively convey warmth, emotional stability, and sociability through visual cues such as facial expressions, as well as auditory cues such as speaking rate and intonation, are likely to foster higher levels of trust and engagement among users. Additionally, our findings suggest that lexical features, such as the presence of negation markers, distribution of parts of speech, and other linguistic characteristics including speech disfluencies and level of detail in verbal communication, may also contribute to trust perceptions within conversational interfaces. All features associated with negative traits, whether they are filled pauses or negation, may undermine user trust and satisfaction.

ACKNOWLEDGMENTS

This work is supported by the funds provided by the National Science Foundation and by DoD OUSD (R&E) under Cooperative Agreement PHY-2229929 (The NSF AI Institute for Artificial and Natural Intelligence).

REFERENCES

[1] Wilma A. Bainbridge, Phillip Isola, and Aude Oliva. 2013. The intrinsic memorability of face photographs. *Journal of experimental psychology. General* 142 4 (2013), 1323–34. https://api.semanticscholar.org/CorpusID:30042653

[2] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018. Social robots for education: A review. *Science Robotics* 3 (2018). https://api.semanticscholar.org/CorpusID:52033756

[3] Xi (Leslie) Chen, Sarah Ita Levitan, Michelle Levine, Marko Mandic, and Julia Hirschberg. 2020. Acoustic-Prosodic and Lexical Cues to Deception and Trust: Deciphering How People Detect Lies. *Transactions of the Association for Computational Linguistics* 8 (04 2020), 199–214. https://doi.org/10.1162/tacl_a_00311 arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00311/1923366/tacl_a_00311.pdf

[4] Aaron C. Elkins and Douglas C. Derrick. 2013. The Sound of Trust: Voice as a Measurement of Trust During Interactions with Embodied Conversational Agents. *Group Decision and Negotiation* 22, 5 (September 2013), 897–913. https://doi.org/10.1007/s10726-012-9339-x

[5] Kurt Kraiger, Alyssa K. McGonagle, and Diana R. Sanchez. 2020. What's in a Sample? Comparison of Effect Size Replication and Response Quality across Student, MTurk, and Qualtrics Samples 1. https://api.semanticscholar.org/CorpusID:252690288

[6] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of Human Factors and Ergonomics Society* 46 (2004), 50 – 80. https://api.semanticscholar.org/CorpusID:5210390

[7] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016). https://api.semanticscholar.org/CorpusID:1036498

[8] Aaron M. Ogletree and Benjamin Katz. 2021. How Do Older Adults Recruited Using MTurk Differ From Those in a National Probability Sample? *The International Journal of Aging and Human Development* 93, 2 (2021), 700–721. https://doi.org/10.1177/0091415020940197 arXiv:https://doi.org/10.1177/0091415020940197 PMID: 32683886.

[9] Nikolaas N. Oosterhof and Alexander Todorov. 2008. The functional basis of face evaluation. *Proceedings of the National Academy of Sciences* 105, 32 (2008), 11087–11092. https://doi.org/10.1073/pnas.0805664105 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.0805664105

[10] Joshua C. Peterson, Stefan Uddenberg, Thomas L. Griffiths, Alexander T. Todorov, and Jordan W. Suchow. 2022. Deep models of superficial face judgments. *Proceedings of the National Academy of Sciences of the United States of America* 119, 17 (26 April 2022). https://doi.org/10.1073/pnas.2115228119

[11] Minjin Rheu, Ji Youn Shin, Wei Peng, and Jina Huh-Yoo. 2021. Systematic Review: Trust-Building Factors and Implications for Conversational Agent Design. *International journal of human-computer interaction* 37, 1 (2021), 81–96. https://doi.org/10.1080/10447318.2020.1807710

[12] Shuo Tian, Wenbo Yang, Jehane Michael Le Grange, Peng Wang, Wei Huang, and Zhewei Ye. 2019. Smart healthcare: making medical care more intelligent. *Global Health Journal* 3 (10 2019). https://doi.org/10.1016/j.glohj.2019.07.001

[13] Alexander Todorov, Sean G. Baron, and Nikolaas N. Oosterhof. 2008. Evaluating face trustworthiness: A model based approach. *Social cognitive and affective neuroscience* 3, 2 (June 2008), 119–127. https://doi.org/10.1093/scan/nsn009

[14] Ilaria Torre, Emma Carrigan, Rachel McDonnell, Katarina Domijan, Killian McCabe, and Naomi Harte. 2019. The Effect of Multimodal Emotional Expression and Agent Appearance on Trust in Human-Agent Interaction. In *Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games* (Newcastle upon Tyne, United Kingdom) *(MIG '19)*. Association for Computing Machinery, New York, NY, USA, Article 14, 6 pages. https://doi.org/10.1145/3359566.3360065

[15] Tibert Verhagen, Jaap van Nes, Frans Feldberg, and Willemijn van Dolen. 2014. Virtual Customer Service Agents: Using Social Presence and Personalization to Shape Online Service Encounters. *Journal of Computer-Mediated Communication* 19, 3 (2014), 529–545. https://doi.org/10.1111/jcc4.12066 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/jcc4.12066