Doctor Who?: The Influence of AI on Affective Responses to Vaccine Calls

Cathy Sheng
Electrical Engineering and Computer Science
University of California, Berkeley
Berkeley, United States
shengcathy@gmail.com

Deanna Vuong
Mission Hills High School
San Marcos, United States
vng.deanna@gmail.com

Jason Zhang
Julia R. Masterman School
Philadelphia, United States
jasonzhang2021@gmail.com

Noha Yousif
James M. Bennett High School
Salisbury, United States
nohagyousif@gmail.com

Zhiyuan Ma
Mills High School
Millbrae, United States
zhizhizhima1@gmail.com

Danielle Amir-Lobel
Stanford Management Science & Engineering
Stanford University
Stanford, United States
daniamlo@stanford.edu

This work was supported by the Stanford Engineering Department.

***Abstract*—AI voice generators are becoming increasingly prevalent in personal and professional settings. Existing research postulates a negative human bias towards AI-generated text; however, there is a gap in the literature surrounding perceptions of artificial audio. Using a novel survey, this paper examines how the gender (male or female) and perceived identity (AI or human) of a caller affects receptiveness towards the message: in this context, a promotion for a fictitious vaccine. Contrasting previous literature, the results do not support a bias against AI-generated audio. These findings, in tandem with further research, could inform more effective implementation of artificial audios.**

***Index Terms*—vaccine attitudes, text-to-speech, artificial intelligence (AI), doctor gender**

I. Introduction

Whether through precision medicine, diagnosis, or virtual voice assistance, the potential impact of artificial intelligence (AI) in the medical field is undeniable [1] . But whether positive or negative, human perceptions of AI will impact its efficacy, regulation, and utilization [2]. The present study seeks to examine these perceptions and determine: *How do people's beliefs about AI's role in producing an audio recording providing medical advice affect their perception of that advice and behavioral intention?*

II. Conceptual Development

Previous studies have shown that when people are presented with identical text stimuli, they rate texts lower in terms of affective qualities (such as honesty, trustworthiness, and sincerity) if they know that ChatGPT was used to generate it [3]. Similar biases have been observed regarding general satisfaction ratings, which tend to be lower for content labeled as AI-generated. Zhang and Gosline attribute this phenomenon to human favoritism rather than AI aversion [4]. Thus, it was predicted that the disclosed AI voice will be rated lower on both affective attributes and general satisfaction than the undisclosed AI voice, regardless of voice gender (**H1**).

Furthermore, among studies featuring a one-way communication design, AI was significantly less effective in promoting behavioral intention than humans [5]. Therefore, it was anticipated that participants' self-reported intention to get, recommend, or seek more information about the norovirus vaccine would be lower in disclosed AI conditions compared to undisclosed ones, following a similar pattern as responses to textual stimuli (**H2**).

Gender also plays a pivotal role in this study, as past research assumes that male doctors are often perceived as more competent, professional and trustworthy, and female doctors more gentle, helpful, sentimental, and sociable [6]. However, it is important to note that past studies were primarily conducted on pools of real-life doctors and patients; thus, their findings were influenced by the divergent ways in which male and female doctors practically communicate. For example, female doctors often hold longer conversations with their patients and are also found to be more accepting of criticism. [7]. By contrast, the current study presents all respondents with the same dialogue, solely altering the gender of the voice. It was nonetheless predicted that this pattern of gender biases would result in higher ratings on affective traits for the female voice and higher ratings on non-affective traits for the male voice (**H3**).

Finally, potentially due to gaps in education, political beliefs, and lack of access to technology, individuals from lower-income backgrounds have been found to be more hesitant towards receiving vaccines [8]. People from lower-income backgrounds also tend to show a stronger aversion to adopting new technologies [9]. Thus, participants from low-income backgrounds were expected to rate the AI-disclosed voices less favorably and show more hesitancy towards receiving vaccines across conditions (**H4**).

III. Methodology

Using the AI voice generator ElevenLabs, four different recordings were created of a caller informing patients about a fictional norovirus vaccine: NoroPro. Each recording corresponded with a different condition: two recordings varied by the gender of the voice (male or female), and two sets of instructions and text content differed based on whether the speaker was identified as a doctor or an AI medical assistant. These conditions will be referred to in the paper as female human, female AI, male human, and male AI.

To account for possible differences in intonation and tone, both of the voices were selected from a subgroup of ElevenLabs voices classified as American, middle-aged, and confident. Prior to releasing the experimental survey, a pretest was conducted (n = 50) on CloudResearch's survey platform Connect, asking participants to rate recordings of these two voices from 1 ('Definitely human') to 7 ('Definitely AI'). When prompted with the choice, participants leaned towards believing that the recordings were machine-generated, with an average rating of 5.50 (σ = 1.56) and 5.79 (σ = 1.59) for the male and female voices respectively. However, since the difference between conditions was not significant (p = 0.515), these voices were used for the main experimental survey.

In the main survey, participants were randomly assigned to one of the four recordings to listen to before answering two sets of questions via Connect. The first set of questions asked participants to evaluate the audio recording's effectiveness by rating their likelihood to receive the vaccine on a seven-point Likert scale. Then, respondents were asked to rate the recording on eight criteria on a scale from 1 (extremely poor) to 7 (extremely strong). These criteria included affective attributes (trustworthiness, genuineness, friendliness, and expressiveness) as well as non-affective attributes (clarity, specificity, professionalism, and efficiency). This rating concluded with a question asking for participant's overall satisfaction with the call. Next, participants rated their likelihood to advise their friends and family to get the vaccine, seek further information about the vaccine, or hang up before the recording ended (on

a scale of 1 to 7, where 1 is extremely unlikely and 7 is extremely likely). The second set of questions inquired about respondents' backgrounds and former experiences. Participants disclosed their gender, race, household income, highest level of education, age, and political ideology (with the option to decline to answer). Then, participants were asked about their former experiences with both vaccines and norovirus to control for any heterogeneity in past knowledge or experience. Finally, participants self-reported the frequency at which they interacted with general AI and voice AI.

While 448 participants were initially recruited via CloudResearch Connect, after participants who failed the attention check were excluded from analyses, 433 valid responses remained, with 111 participants in the male AI condition, 109 in the female AI condition, 107 in the male human condition, and 106 in the female human condition. Of these participants, 218 identified as female, 212 identified as male, and 3 chose not to disclose their gender. The average age was 39.5 0 (σ = 11.65), and most participants were white (289), followed distantly by Asian (62).

IV. Results

Disregarding conditions, participants generally leaned slightly toward being likely to seek more information ($\overline{x}$ = 4.82, σ = 1.85), get the vaccine ($\overline{x}$ = 4.26, σ = 1.90), and advise others to get the vaccine ($\overline{x}$ = 4.12, σ = 1.89) after listening to the recordings. However, they also reported being somewhat likely to hang up early if they had received the call in a real life setting ($\overline{x}$ = 4.30, σ = 2.03).

Affective and non-affective summary variables were created by averaging the respective attribute ratings. These summary variables indicated that participants ranked the recordings higher on non-affective traits than affective ones under all four conditions (See Table I). Performing a one-tailed t-test on the summary variables showed that this difference is statistically significant in all conditions ($p < 0.001$ for all).

Linear regression was used to analyze the effect of the four conditions on the participants' ratings. For the dependent variables, Likert scales were treated as continuous. Under all conditions, male participants tended to rate recordings lower on both affective and non-affective traits than female participants. This is most pronounced in the female AI condition for both affective traits (male participants: $\overline{x}$ = 4.77, σ = 1.39, female participants: $\overline{x}$ = 5.35, σ = 1.08, p = 0.008) and non-affective traits (male participants: $\overline{x}$ = 5.50, σ = 1.35, female participants: $\overline{x}$ = 6.07, σ = 0.92, p = 0.006). While the same trend exists for the other three conditions, these differences were not statistically significant.

Most attributes did not differ significantly in their ratings between conditions. The exception was that participants in the female AI condition assigned a higher friendliness rating to their recording (see Friendliness Model 1 in Table II, Model 2 described in later paragraph). The difference in ratings is likely due to the gender of the voices rather than their labels (AI or human) as participants in the female human condition also gave their recording a much higher rating (though the p-value is higher than 5%). Thus, contrary to **H1** and **H2,** the disclosed AI identity did not have a significant main effect on dependent variables.

TABLE I
Average Affective Traits Rating Versus Non-Affective Traits Rating by Condition

Condition	**Affective Rating**		**Non-Affective Ratings**	
	$\overline{x}$	σ	$\overline{x}$	σ
Female AI	5.04	1.28	5.78	1.18
Male AI	5.00	1.36	5.98	0.97
Female Human	5.00	1.20	5.79	1.04
Female AI	4.06	1.34	5.87	1.06

TABLE II
Regression Results for Friendliness Models Without and With Control Variables

	Friendliness Model 1	**Friendliness Model 2**
Constant	4.7944*** (0.139)	4.4963*** (0.298)
Male AI	+0.1786 (0.195)	+0.1431 (0.197
Female Human	+0.3660* (0.197)	+0.2931 (0.199)
Female AI	+0.4074** (0.196)	+0.3838* (0.197)
Controls		✓

*p<0.1; **p<0.05; ***p<0.01

Interestingly, adding an interaction term between gender and condition to the regression models shows that, for some affective and non-affective traits, participants gave a higher ranking to recordings with the same gender as them. For instance, for the expressiveness rating, the interaction term "condition female AI:participant gender female" is significant (p = 0.041). Examining the fitted values of the model (See Table III) reveals that male participants tended to rate male voices to be more expressive than female voices. However, the opposite is true for female participants; they tended to rate female voices higher than male. These results demonstrate that **H3** may need to be modified for future experiments, as participants with different genders may have divergent perceptions toward the voices.

Additional regression models (without interaction terms) were created, but with covariables added for income, age, vaccine attitude, political ideology, AI experience, voice AI experience, vaccine attitude, and time of last vaccine received voluntarily (see Table IV for examples of regression set-up). The male human condition remains the baseline. Ten responses were dropped as the participants preferred not to report one of these covariables, leaving 423 responses to train the models. Interestingly, the friendliness rating ceases to be significant (see Friendliness Model 2 in Table II), suggesting that the impact of the conditions is reduced after taking account of participants' backgrounds. However, participants' likelihood to get the vaccine does become significant for the female human condition. (see Table IV).

As for **H4**, when the model only includes the conditions and

TABLE III
FITTED VALUES FOR EXPRESSIVENESS MODEL (WITH INTERACTION) BY CONDITION AND PARTICIPANTS' GENDER

	Male Participants	**Female Participants**
Male Human	4.79	4.83
Male AI	4.61	5.08
Female Human	4.72	4.98
Female AI	4.46	5.39

TABLE IV
REGRESSION RESULTS FOR MODELS WITH COVARIABLES FOR SELECTED DEPENDENT VARIABLES

	Likelihood to Get Vaccine	**Likelihood to Advise Others to Get Vaccine**	**Likelihood to Hang Up**	**Efficiency Rating**
Constant	5.7795*** (0.298)	5.2731*** (0.315)	4.3486*** (0.415)	5.9172*** (0.264)
Male AI	+0.0541 (0.197)	+0.1843 (0.208)	+0.0951 (0.274)	+0.0905 (0.174)
Female Human	-0.4091** (0.199)	-0.1554 (0.210)	+0.0346 (0.277)	-0.2199 (0.176)
Female AI	-0.0455 (0.197)	+0.1629 (0.208)	-0.0429 (0.274)	-0.0336 (0.174)
Income	+0.0811 (0.055)	+0.0578 (0.058)	+0.0319 (0.076)	-0.0149 (0.048)
Vaccine Attitude	-0.5557*** (0.057)	-0.5086*** (0.060)	+0.2130* (0.080)	-0.2021*** (0.051)
Last Vaccine	-0.3449*** (0.071)	-0.3331*** (0.075)	+0.0910 (0.099)	-0.0002 (0.063)
Political Ideology	+0.0236 (0.072)	+0.0823 (0.076)	-0.1109 (0.100)	+0.0147 (0.064)
AI Experience	+0.0400 (0.056)	+0.0378 (0.059)	-0.0322 (0.078)	+0.1247** (0.050)
Voice AI Experience	+0.1511*** (0.057)	+0.1457** (0.060)	-0.2109*** (0.079)	-0.0393 (0.050)

*p<0.1; **p<0.05; ***p<0.01

income as the regressors, income significantly affect likelihood to get the vaccine (β = +0.2407, p = 0.001) and likelihood to advise others to get the vaccine (β = +0.2116, p = 0.002), with higher likelihoods corresponding with higher annual income.

Finally, the more often participants used or interacted with voice AI tools, the more likely participants were to say they would seek more information about the vaccine, get the vaccine, or advise others to get the vaccine. They are also more satisfied with the call and less likely to hang up before the call is over. (See Table IV for models with 'Likelihood to Get Vaccine,' 'Likelihood to Advise Others to Get Vaccine,' and 'Likelihood to Hang Up'.). This trend does not occur with the AI experience covariable. Interestingly, the opposite is true with affective and non-affective traits, where increased AI Experience positively correlates with most traits (except clarity, professionalism, and friendliness), but voice AI experience does not (see Table IV for efficiency rating as an example).

As may be expected, a more negative self reported vaccine attitude correlates with a lower rating for most ratings.

V. DISCUSSION

Participants rated the recordings higher on non-affective attributes than they did affective attributes, reinforcing previous studies that indicate AI voice technologies are perceived as more efficient and functional than emotional and genuine [10]. Furthermore, the survey results show a correlation between greater self-reported AI experience and higher ratings for both affective and non-affective traits, suggesting that individuals more familiar with AI tend to have more positive perceptions of its attributes.

Additionally, since friendliness ratings for the female AI condition were significantly high in comparison to the other three conditions, further investigations into gendered AI voices could reveal similar correlations. Previously studied biases towards females as more friendly may have influenced participants in the female voice groups, although future research is needed to further examine this correlation in AI contexts [11]. Interestingly, as demonstrated by the expressiveness ratings, participants' gender may have an effect on their perception of the recordings, as they appear somewhat biased in favor of recordings with the same gender as them. This has implications for the design of AI voice assistants.

One limitation observed was that, although the male and female voices were designed to contain minimal differences in tone and accent, subtle variations in accent and tone could have skewed certain results. Additionally, although the recordings were deemed humanlike enough to continue with the main experiment, participants in the pretest survey were more prone to believe the recordings were AI-generated than human-generated. Therefore, the labels may not have had the intended effects on participants' perceptions, as a participant in the human condition may still have suspected the voice was AI-generated.

Further, this study assessed human biases towards a subject towards which people often hold strong preconceived beliefs, negative and positive [12]. Thus, attempting to answer similar questions surrounding AI aversion and gender biases in a less politically-charged context could aid the generalizability of the results.

Overall, this study illuminates potential biases in perceptions of AI voices, underscoring the need for further investigation. Continued exploration of human perceptions of AI can help reduce these biases and support the effective and fair integration of AI voice technology in both medical settings and society at large.

ACKNOWLEDGMENT

We thank the staff of the Stanford Compression Forum's SHTEM program, especially our mentor Danielle Amir-Lobel, for their guidance and unwavering support.

REFERENCES

[1] B. Meskó and M. Görög, "A Short Guide for Medical Professionals in the Era of Artificial Intelligence," *Npj Digital Medicine*, vol. 3, no. 1, Sep. 2020, doi: https://doi.org/10.1038/s41746-020-00333-z.

[2] R. Modhvadia, "How do people feel about AI?," *www.adalovelaceinstitute.org*, Jun. 06, 2023. Available: https://www.adalovelaceinstitute.org/report/public-attitudes-ai/

[3] I. Rae, "The Effects of Perceived AI Use On Content Perceptions," May 2024, doi: https://doi.org/10.1145/3613904.3642076.

[4] Y. Zhang and R. Gosline, "Human favoritism, not AI aversion: People's perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation," *Judgment and Decision Making*, vol. 18, p. e41, 2023. doi: 10.1017/jdm.2023.37.

[5] G. Huang and S. Wang, "Is artificial intelligence more persuasive than humans? A meta-analysis," *Journal of Communication*, Aug. 2023, doi: https://doi.org/10.1093/joc/jqad024.

[6] V. T. Haile, "Gender as a Factor in the Physician and Patient Interaction: From the Service Quality Perspective," *Journal of Marketing and Consumer Behaviour in Emerging Markets*, vol. 2/2018, no. 8, pp. 21–32, Dec. 2018, doi: https://doi.org/10.7172/2449-6634.jmcbem.2018.2.2.

[7] H. Löffler-Stastka, T. Seitz, S. Billeth, B. Pastner, I. Preusche, and C. Seidman, "Significance of Gender in the Attitude Towards doctor-patient Communication in Medical Students and Physicians," *Wiener Klinische Wochenschrift*, vol. 128, no. 17, pp. 663–668, 2016, doi: https://doi.org/10.1007/s00508-016-1054-1.

[8] V. Casigliani et al., "Vaccine hesitancy and cognitive biases: Evidence for tailored communication with parents," *Vaccine: X*, vol. 11, p. 100191, Aug. 2022, doi: https://doi.org/10.1016/j.jvacx.2022.100191.

[9] C. Srinuan and E. Bohlin, "Understanding the digital divide: A literature survey and ways forward," *www.econstor.eu*, 2011. Available: https://hdl.handle.net/10419/52191.

[10] M. Reis, F. Reis, and W. Kunde, "Beyond the stethoscope: Labelling medical advice as AI- compared to human-generated reduces perceived reliability and empathy," Feb. 2024, doi: https://doi.org/10.31234/osf.io/35hn8.

[11] Leyla Dogruel and S. Joeckel, "Gender Stereotypes and Voice assistants: Do Users' Gender and Conversation Topic matter?," *Behaviour & Information Technology*, pp. 1–11, Jul. 2023, doi: https://doi.org/10.1080/0144929x.2023.2235021.

[12] S. Pourrazavi, Z. Fathifar, M. Sharma, and H. Allahverdipour, "COVID-19 vaccine hesitancy: A Systematic review of cognitive determinants," *Health Promotion Perspectives*, vol. 13, no. 1, pp. 21–35, Apr. 2023, doi: https://doi.org/10.34172/hpp.2023.03.