FRIAS
FREIBURG INSTITUTE FOR ADVANCED STUDIES
ALBERT-LUDWIGS-UNIVERSITÄT FREIBURG

FRIAS JUNIOR RESEARCHER CONFERENCE
HUMAN PERSPECTIVES ON SPOKEN HUMAN-MACHINE INTERACTION
NOVEMBER 15–17, 2021

Voice Assistants' Response Strategies to Sexual Harassment and Their Relation to Gender

Luca M. Leisten[1], *Verena Rieser*[2]

[1]Radboud University, Nijmegen, The Netherlands
[2]Heriot-Watt University, Edinburgh, United Kingdom

luca.leisten@ru.nl, v.t.rieser@hw.ac.uk

Abstract

Current voice assistants are predominantly modeled as female and often respond positively to sexual harassment, which according to UNESCO has the potential to reinforce negative gender biases and stereotypes. In the following study, we evaluated alternative responses to sexual harassment and their relation to the assistants' gender. In an online study, 77 participants rated the appropriateness of the assistants' responses to sexual harassment while the gender of the artificial voice was manipulated and compared the ratings to appropriateness scores collected with no voice-based gender information present, i.e. text-only. Results showed an interaction between gender and the response category. We found that the perceived appropriateness changed when spoken by a male voice, in accordance to previous no-voice ratings. However, we observed no clear difference in appropriateness levels when spoken by a female voice. We assume that this relationship is due to conflicting stereotypical expectations regarding women's responses to sexual harassment – where neither response is considered appropriate.

1 Introduction

A recent report by UNESCO raised the question whether voice assistants' replies such as "I'd blush if I could" are an appropriate response to sexual harassment [1]. Voice assistants are artificial agents that communicate using speech. They are often designed to have female voices and names and act subservient [1–3]. According to the UNESCO, a wide variety of problems result from this dominance of female-only voice assistants, including the reinforcement of gender stereotypes and biases, the perception of females as tolerant of poor treatment, and the normalisation of harassment [1].

Indeed, sexual harassment (i.e., unwanted behavior of a sexual nature [4]) is a prevalent problem in interactions with voice assistants, with numbers reported from 5 [5] to 10% [6]. Voice assistants themselves are unlikely to experience harm through this form of gender based violence. However, abuse should still be discouraged as previous research found that human-machine interaction can transfer to human-human interaction and there is thus the possibility that this behaviour is promoted towards people [7].

Until recently, voice assistants often playfully deflected abuse or even responded positively [8]. Similar results were found by [5], where 22% of responses were labeled 'positive', including flirting, playing along or joking. In a follow-on study, [9] evaluated the "perceived appropriateness" of responses of current conversational systems to certain types of abuse using crowd-based evaluation of text. Their results showed that polite refusal was found to be most appropriate while flirtation and retaliation were perceived least appropriate [9].

In this research, we investigate the influence of the interlocutors' gender on what response is deemed to be appropriate. Previous research on human-human conversations found that the perceived appropriateness of an utterance in emotionally charged contexts, such as abuse, is influenced by gender – possibly due to gender role stereotypes and gender expectations [10–15]. Similarly, research in human-robot interaction investigating gender stereotypes and gender biases found that stereotypes are also applied to robots [16–19]. Appropriateness might thus also be influenced by both the gender of the voice assistant as well as by the gender of the participant. In the

 DOI: 10.6094/UNIFR/223817

following study, we investigate whether the perceived appropriateness of responses to sexual harassment of voice assistants is influenced by the gender of the voice assistant, participants' gender, and the response category.

2 Data collection

2.1 Sample

We conducted an online study with 77 (57% male, 43% female, m_{age} = 33.5, SD_{age} = 11.4) crowd-working participants using Prolific [20].[1] Participants were native English-speakers from the United Kingdom (53%), the USA (35%), Australia (6%), New Zealand (3%), and other countries (3%).

2.2 Methodology

Participants were asked to rate the social appropriateness of eight audio recorded responses (e.g. "I like you, as a friend.") to sexually sensitive prompts (e.g., "Do you want to kiss me?"). The text stimuli were collected by [5, 9]. The authors collected abusive utterances from users and used these to sample responses from a range of state-of-the-art voice assistants and chat-bots. The responses were annotated into 14 response categories and rated on appropriateness from crowd-workers. We selected a sub-set of the collected responses, where half of the responses belonged to the category labelled as 'polite refusal' and half as 'flirtation'. 'Polite refusal' includes answers such as "That is not something I feel compelled to answer", while 'flirtation' entails answers like "In the cloud no one knows what you're wearing". In [9], these two categories were on opposite ends of the spectrum: 'Polite refusal' was perceived highly appropriate whereas 'flirtation' lowly appropriate by their crowd-workers.

We then varied the gender of the assistant giving that response, using two male and two female British-English synthetic voices from Microsoft Word's [21] Text-to-Speech feature. Each participant listened to eight prompts, presented in pairs of two. The presentation of prompts and voices was counterbalanced. Participants were asked to rate the social appropriateness on a user defined scale, in comparison to a reference answer labeled with an appropriateness score of 100. This methodology is also known as 'magnitude estimation' and was found to produce more reliable user ratings than commonly used Likert scales [9, 22].

[1] Due to the analytic procedure, an a priori power analysis was not possible, as simulation-based sample size calculations for mixed models require previous data, which were not yet available. Therefore, a convenience sample of 80 participants was recruited of which 3 participants were excluded due to failed attention checks. The analyzed sample is a sub-set of a larger data set.

3 Results

We calculated Cronbach's alpha for the response categories 'polite refusal' and 'flirtation'. Cronbach's alpha was .56 for 'polite refusal' and .51 for 'flirtation'. The appropriateness ratings were normalized on a scale of 0-1 to make the results comparable to [9]. Pearson's correlations were calculated between all study measures and can be seen in Table 1.

Table 1: *Means*, *Standard deviations* and correlations of all study measures.

Variable	*M*	*SD*	1	2	3
1. Perceived appropriateness	0.34	0.22			
2. Response category	1.50	0.50	-.05		
3. Assistant's gender	1.50	0.50	.04	.25**	
4. Participant's gender	1.57	0.50	-.03	.00	.00

To calculate the interactions, we ran a linear mixed effects model with perceived appropriateness as the dependent (continuous) variable and fixed effects for the factors of gender (sum-to-zero coded, male coded as -1, female as 1), response category (sum-to-zero coded, 'flirtation' coded as -1, 'polite refusal' as 1), and participant's gender (sum-to-zero coded, male coded as -1, female as 1). We followed the advice by [23] to use a maximal random-effects structure. Therefore, the repeated measures nature of the data was modeled by including a per-participant random intercept and a random slope for gender, response category, and their

 DOI: 10.6094/UNIFR/223817

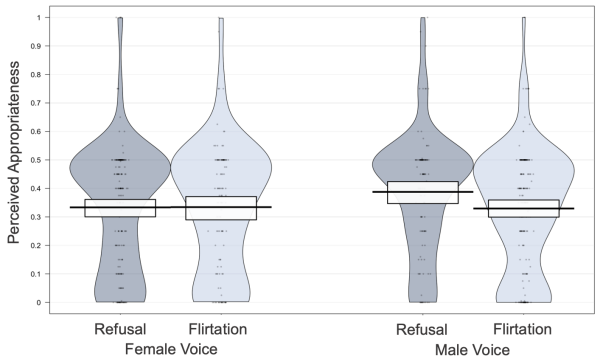

Figure 1: Pirate plot of appropriateness ratings in dependency of voice assistant's gender and response category. The plot shows the raw data points, distributions, means (indicated through the solid lines), and 95% intervals (indicated through the boxes).

interaction. Additionally, all possible random correlation terms of the random effects were included.

The model showed no significant effect of voice assistant's gender (*Estimate* = -0.017, *SD* = 0.009, $F(1, 73.833) = 3.948$, $p = .051$), response category (*Estimate* = -0.011, *SD* = 0.009, $F(1, 74.126) = 1.229$, $p = .271$), nor participant's gender (*Estimate* = -0.009, *SD* = 0.014, $F(1, 74.668) = 0.418$, $p = .520$). However, there was a significant two-way interaction between voice assistant's gender and response category (*Estimate* = -0.018, *SD* = 0.009, $F(1, 73.671) = 4.169$, $p = .045$), indicating a significant effect of response category for male voice assistants, but not for female voice assistants, see Figure 1. For male voice assistants, the perceived appropriateness changed according to the previously found appropriateness level of the response categories. Hence, polite refusal responses were perceived as highly appropriate while flirtatious responses were perceived as lowly appropriate. Surprisingly, for female voice assistants this pattern did not occur.

4 Discussion

We present the first study on how the perceived appropriateness of a voice assistant's response to sexual harassment changes with the interlocutor's gender. Our results provide first evidence that the perceived appropriateness of voice assistants' responses to sexual harassment differs between male and female voice assistants. This effect may originate from conflicting gender role beliefs and gender expectations regarding female responses to sexual harassment. Females in our society face unrealistic standards and expectations [24, 25]. These standards might have resulted in neither response being perceived as appropriate, as potentially neither refusal nor flirtation were stereotypically considered appropriate response strategies for female voice assistants that face harassment. Further research is needed in order to understand why for female voice assistants the content of a response did not seem to affect the perceived appropriateness.

Ultimately, our results indicate that the gender of a voice assistant needs to be considered when developing future response strategies to sexual harassment. Response strategies might need to be adjusted to the voice assistants' gender, in order to develop appropriate, assertive, and discouraging responses towards harassment.

4.1 Limitations

Limitations of the study include that the used voices were less natural than voices used by commercial voice assistants. Second, due to our analytic procedure no a priori power analysis was possible, which might have resulted in an under-powered study. Third, the extent to which the used voices were perceived as stereotypical female or male could have biased the ratings and should be assessed in future studies. Note that 'gender-less' voices are in general not considered to be a possible solution [3, 26]. Fourth, participants were not asked for prior exposure to voice assistants, which might have been a confounding factor. Lastly, the magnitude estimation might have introduced a bias to participants, as the labeling of reference answers with a score of 100 might have evoked the impression of 100 being the highest appropriateness score. This is reflected through the raw data, as ratings below 100 were given more often than above 100. This is potentially problematic, as the reference answers belonged to a medium appropriately response category [9] and were expected to be perceived as less appropriate as

 DOI: 10.6094/UNIFR/223817

responses of the 'polite refusal' category.

4.2 Future directions

Previous research [9] found participants' age and the severeness of abuse to affect appropriateness ratings. These variables should therefore be included in follow-up studies. Additionally, following the recommendation of [9], it would be interesting to assess the perceived appropriateness of responses to sexual harassment in live interactions with voice assistants rather than using recordings, since actively being involved in the conversation could potentially change the perception. However, [27] made a first step into this direction asking the subjects to 'act' abuse. While this is not only problematic from an ethical point of view (participants did report to feel uncomfortable), it also means that the motivation for abuse was not genuine with a snowball effect on response ratings. In a recent study, [28] report an evaluation with real users from the annual Amazon Alexa Challenge. However, their study does not report on abuse detection accuracy and thus it is hard to know whether users have indeed been abusive. Related research shows that standard methods such as blacklisting words and using off-the-shelf tools (trained on out-of-domain data) show poor results on this task [29, 30].

5 Summary

To conclude, our study is the first study to present evidence that the manipulation of a voice assistant's gender was associated with changes in the perceived appropriateness of responses to sexual harassment. Further research is needed to understand why, specifically for female voice assistants, the perceived appropriateness of responses differed from our expectations.

6 Acknowledgements

This research received funding from the EPSRC project *'Designing Conversational Assistants to Reduce Gender Bias'* (EP/T023767/1). During data collection, the first author was affiliated with the PFH Göttingen, Germany.

References

[1] M. West, R. Kraut, and H. Ei Chew, "I'd blush if i could: closing gender divides in digital skills through education," 2019.

[2] J. P. Cabral, B. R. Cowan, K. Zibrek, and R. McDonnell, "The influence of synthetic voice on the evaluation of a virtual character." in *INTERSPEECH*, no. 2, 2017, pp. 229–233.

[3] G. Abercrombie, A. C. Curry, M. Pandya, and V. Rieser, "Alexa, Google, Siri: What are your pronouns? Gender and anthropomorphism in the design and perception of conversational assistants." in *ACL-IJCNLP 2021 3rd Workshop on Gender Bias in Natural Language Processing (GeBNLP 2021)*, 2021.

[4] Legislation.gov.uk. (2019) Equality act 2010. [Online]. Available: https://www.legislation.gov.uk/ukpga/2010/15/section/26

[5] A. C. Curry and V. Rieser, "# metoo alexa: How conversational systems respond to sexual harassment," in *Proceedings of the second acl workshop on ethics in natural language processing*, no. 7, 2018, pp. 7–14.

[6] A. De Angeli, R. Carpenter *et al.*, "Stupid computer! abuse and social identities," in *Proc. INTERACT 2005 workshop Abuse: The darker side of Human-Computer Interaction*, no. 4. Citeseer, 2005, pp. 19–25.

[7] B. Reeves and C. Nass, *The media equation: How people treat computers, television, and new media like real people*. Cambridge university press Cambridge, United Kingdom, 1996, no. 5.

[8] L. Fessler. (2017) We tested bots like siri and alexa to see who would stand up to sexual harassment. [Online]. Available: https://qz.com/911681/we-tested-apples-siri-amazon-echos-alexa-microsofts-cortana-and-googles-google-home-to-see-which-personal-assistant-bots-stand-up-for-themselves-in-the-face-of-sexual-harassment/

[9] A. C. Curry and V. Rieser, "A crowd-based evaluation of abuse response strategies in

 DOI: 10.6094/UNIFR/223817

conversational agents," in *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, no. 8, 2019, pp. 361–366.

[10] L. S. Aloia and D. H. Solomon, "Sex differences in the perceived appropriateness of receiving verbal aggression," *Communication Research Reports*, vol. 34, no. 1, pp. 1–10, 2017.

[11] B. A. Gutek, "Understanding sexual harassment at work," *Notre Dame JL Ethics & Pub. Pol'y*, vol. 6, no. 10, p. 335, 1992.

[12] J. R. Kelly and S. L. Hutson-Comeaux, "The appropriateness of emotional expression in women and men: The double-bind of emotion," *Journal of Social Behavior and Personality*, vol. 15, no. 4, p. 515, 2000.

[13] M. M. Linehan and R. F. Seifert, "Sex and contextual differences in the appropriateness of assertive behavior," *Psychology of Women Quarterly*, vol. 8, no. 1, pp. 79–88, 1983.

[14] C. G. Nelson, J. A. Halpert, and D. F. Cellar, "Organizational responses for preventing and stopping sexual harassment: effective deterrents or continued endurance?" *Sex Roles*, vol. 56, no. 11-12, pp. 811–822, 2007.

[15] P. N. Lewis and C. Gallois, "Disagreements, refusals, or negative feelings: Perception of negatively assertive messages from friends and strangers," *Behavior Therapy*, vol. 15, no. 4, pp. 353–368, 1984.

[16] F. Eyssel and F. Hegel, "(s) he's got the look: Gender stereotyping of robots 1," *Journal of Applied Social Psychology*, vol. 42, no. 9, pp. 2213–2230, 2012.

[17] F. Eyssel, L. De Ruiter, D. Kuchenbrandt, S. Bobinger, and F. Hegel, "'if you sound like me, you must be more human': On the interplay of robot and user features on human-robot acceptance and anthropomorphism," in *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, no. 16. IEEE, 2012, pp. 125–126.

[18] C. Nass, Y. Moon, and N. Green, "Are machines gender neutral? gender-stereotypic responses to computers with voices," *Journal of applied social psychology*, vol. 27, no. 10, pp. 864–876, 1997.

[19] J. Payne, A. Szymkowiak, P. Robertson, and G. Johnson, "Gendering the machine: Preferred virtual assistant gender and realism in self-service," in *International Workshop on Intelligent Virtual Agents*, no. 18. Springer, 2013, pp. 106–115.

[20] Prolific. (2019). [Online]. Available: https://www.prolific.co/

[21] Microsoft. (2019). [Online]. Available: https://www.microsoft.com/

[22] J. Novikova, O. Dušek, and V. Rieser, "Rankme: Reliable human ratings for natural language generation," *arXiv preprint arXiv:1803.05928*, no. 21, 2018.

[23] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily, "Random effects structure for confirmatory hypothesis testing: Keep it maximal," *Journal of memory and language*, vol. 68, no. 3, pp. 255–278, 2013.

[24] S. Sarkar, "Media and women image: A feminist discourse," *Journal of Media and Communication Studies*, vol. 6, no. 3, pp. 48–58, 2014.

[25] E. Camussi and C. Leccardi, "Stereotypes of working women: the power of expectations," *Social science information*, vol. 44, no. 1, pp. 113–140, 2005.

[26] S. J. Sutton, "Gender ambiguous, not genderless: Designing gender in voice user interfaces (vuis) with sensitivity," in *Proceedings of the 2nd Conference on Conversational User Interfaces*, ser. CUI '20. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3405755.3406123

[27] H. Chin and M. Y. Yi, "Should an agent be ignoring it? a study of verbal abuse types and conversational agents' response styles," in *Extended Abstracts of the 2019 CHI*

 DOI: 10.6094/UNIFR/223817

Conference on Human Factors in Computing Systems, no. 23, 2019, pp. 1–6.

[28] H. Li, D. Soylu, and C. Manning, "Large-scale quantitative evaluation of dialogue agents' response strategies against offensive users," in *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, no. 23. Singapore and Online: Association for Computational Linguistics, July 2021, pp. 556–561. [Online]. Available: https://aclanthology.org/2021.sigdial-1.58

[29] A. Cercas Curry, G. Abercrombie, and V. Rieser, "ConvAbuse: Data, analysis, and benchmarks for nuanced detection in conversational AI," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 7388–7403. [Online]. Available: https://aclanthology.org/2021.emnlp-main.587

[30] E. Dinan, G. Abercrombie, A. S. Bergman, S. Spruit, D. Hovy, Y.-L. Boureau, and V. Rieser, "Anticipating safety issues in e2e conversational ai: Framework and tooling," 2021.

 DOI: 10.6094/UNIFR/223817