Are you sure you want to order that?

On Appropriateness of Voice-only Proactive Feedback Strategies

Mateusz Dubiel
University of Luxembourg
Luxembourg
mateusz.dubiel@uni.lu

Kerstin Bongard-Blanchy
University of Luxembourg
Luxembourg
kerstin.bongard-blanchy@uni.lu

Luis A. Leiva
University of Luxembourg
Luxembourg
luis.leiva@uni.lu

Anastasia Sergeeva
University of Luxembourg
Luxembourg
anastasia.sergeeva@uni.lu

ABSTRACT

Conversational agents (CAs) that deliver proactive interventions can benefit users by reducing their cognitive workload and improving performance. However, little is known regarding how such interventions would impact perception of CA's appropriateness in voice-only, decision-making tasks. We conducted a within-subjects experiment (N=30) to evaluate the effect of CA's feedback delivery strategy at three levels (no feedback, unsolicited, and solicited feedback) in an interactive food ordering scenario. We discovered that unsolicited feedback was perceived to be more appropriate than solicited feedback. Our results provide preliminary insights regarding the impact of proactive feedback on CA perception in decision-making tasks.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing design and evaluation methods**; *Auditory feedback*.

KEYWORDS

Conversational Agents; Synthetic Speech; Design Ethics; Trust

1 INTRODUCTION

Conversational Agents (CAs) such as Amazon Alexa, Apple Siri, or Google Home are becoming increasingly ubiquitous. According to The Smart Audio Report, in April 2022 over 82 million people owned a smart speaker in the United States alone [24]. While CAs are still predominantly used for simple tasks such as checking the weather, playing music, or setting alarms [1], a growing number of users are expecting to use them routinely for purchasing products and services online [24]. An analogous trend can be also observed for personalised conversational recommender systems [15].

The growing popularity and increased usage of voice-based CAs could be partly attributed to their ever-improving natural language processing capabilities. Recent developments in Deep Learning have led to a rapid improvement in the quality of synthetic voices in terms of intelligibility and naturalness, making them almost indistinguishable from human speech [11]. Research shows that CAs that sound like humans are generally perceived as significantly more trustworthy [26] and likeable [18] than CAs with more 'robot like' voices. However, a recent study found that virtual agents with highly realistic, neural synthetic speech are perceived as more eerie and less trustworthy compared to agents with less natural, concatenative synthetic speech [7].

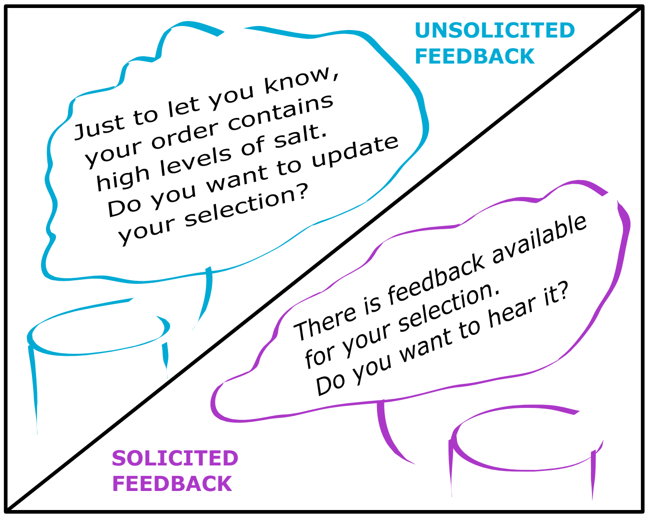
UNSOLICITED FEEDBACK
Just to let you know, your order contains high levels of salt. Do you want to update your selection?
There is feedback available for your selection. Do you want to hear it?
SOLICITED FEEDBACK

Figure 1: Comparison of solicited and unsolicited proactive CA feedback strategies. Solicited feedback is only provided once the user has acknowledged that they want to hear it.

Building on previous research that elicited users' expectations regarding proactive CA support (e.g., [4, 8, 19, 20, 29, 34, 35]), our work focuses on an interactive, voice-only decision-making scenario to explore the impact of proactive interventions of a CA. Specifically, we show that proactively providing feedback regarding menu options affects perceived appropriateness of the CA. Our investigation provides empirical insights on the impact of proactive CA interventions in a standardised, voice-only decision-making scenario which approximates capabilities of present day CAs.

2 BACKGROUND AND RELATED WORK

Speech has been shown to be an effective tool for promoting reflection in the educational context [23], improving focus on task [14] and increasing participants' involvement in an exploratory data analysis task [28]. While researchers highlighted that the role of Artificial Intelligence (AI) should be to empower people and amplify their skills rather than fully automate every task [32], as pointed out by Reicherts et al., *how* to provide the most effective and appropriate proactive support for the users during interactions with intelligent systems remains an open research problem [28]. Here we present key concepts related to perception of feedback appropriateness and discuss some studies that elicited users' requirements regarding proactive CAs.

2.1 Appropriateness of feedback

Research indicates that proactive dialogue strategies (taking initiative to actively provide feedback) can lead to more positive perception of a conversational partner (inlc. more trust and higher compliance) as compared to passive interaction (no feedback) [12, 16, 17]. Furthermore, compared to unsolicited feedback, solicited feedback was found to be: more satisfying [2], less face-threatening [13] (less harmful to one's self-image), and more likely to be utilised [6]. These findings are in line with the Advice Response Theory [21] which postulates that threats to esteem, face, and/or identity are the key factors that affect individuals' responses to persuasive and supportive communication. The common strategy to handle unsolicited advice recommended in communication literature [33] is to ask if the recipient is willing to receive the feedback before actually providing it. Accordingly, here, we follow this recommendation by hypothesising that solicited feedback will positively affect users' perception of a CA's appropriatness.

2.2 Prior studies on Proactive CAs

Luria et al. [20] proposed three degrees of proactivity: *reactive* (responds only when being directly asked), *proactive* (intervenes by providing additional information but without providing recommendations) and *proactive recommender* (intervenes and provides recommendations). In their study, most participants liked the idea of a proactive agent but none of the participants were happy with an agent trying to limit their agency, for instance by preventing them from ordering unhealthy food. In a similar vein, Zargham et al. [35] investigated when conversational agent interventions are appropriate. They found that emergency support and health-related interventions are welcomed provided that the CA asks for the user for permission to intervene. In another elicitation study, Völkel et al. [34] explored how users envision a conversation with a perfect voice assistant – the majority of their participants envisioned a CA that is smarter and more proactive than currently available agents and provides 'well thought-through' suggestions to solve problems.

In this paper, based on previous work that elicited users' expectations regarding voice-based CA's proactivity [20, 34, 35], we investigate how different types of feedback interventions in an interactive decision making task impact perceived appropriateness of the CA. The main contribution of this work is that it provides a full simulation of an in-person, voice-only interaction of users with a proactive CA that goes beyond online evaluations which are frequently limited to isolated prompts that lack a broader context, and do not provide a real-time conversational experience.

3 USER STUDY

We conducted a 3x2 within-subjects experiment to evaluate how the three feedback strategies of the CA *Food Genie* - No Feedback (baseline), Solicited Feedback (with user's permission), and Unsolicited Feedback (without user's permission) - affect the perceived appropriateness of the agent.. Each participant was exposed to each CA feedback strategy once. The order of feedback strategies was randomised with the baseline always used as the starting condition, to reflect a default, reactive interaction manner that is characteristic of the current state-of-the-art CAs [35].

3.1 Research Hypotheses

Following our literature review on human-human [2, 6, 13, 33] and human-computer dialogue [12, 16, 17, 30], and based on the results of previous research which indicates that, in certain conditions, users are open to receiving a proactive assistance from CAs [20, 34, 35], we formulate the following research hypotheses:

H1: A CA which provides **solicited feedback** will be perceived as **more appropriate** compared a CA with no feedback.

H2: A CA which provides **unsolicited feedback** will be perceived as **less appropriate** than a CA with no feedback.

To measure how appropriate the agent's conversational behaviour was during each intervention, we used a one-question appropriateness scale, described in Section 3.4.

3.2 Materials

The voice used in the experiment was developed with the TorToiSe text-to-speech (TTS) software.[1] Specifically, we have chosen an English voice called 'William' from the TorToiSe repository. TorToiSe was inspired by the 'zero-shot text-to-image generation approach' [27] —recently popularised by OpenAI's DALL-E[2] and Google's Imagen[3], among others— which uses an autoregressive decoder and a diffusion-based decoder. TorToiSe allows for high accuracy in capturing vocal qualities of the speaker, leading to a highly expressive and natural synthetic voices. However, due to slow synthesis time (two minutes for 7-10 word sentences on average) it is prohibitive to use it in real-time applications. Therefore, in our experiment we decided to adhere to highly structured scenarios.

All in all, our motivation for choosing TorToiSe was that: (1) the software is open source, and (2) it is capable of creating a high fidelity synthetic speech that outperformed any alternative open source system in terms of naturalness. In order to validate our selection, we conducted listening tests ($N = 14$), where participants were asked to rate three corresponding speech samples generated with TorToiSe and Tacotron 2 [10], alternative state-of-the art TTS system. The results of Wilcoxon Signed-Rank tests (Bonferroni-Holm corrected) indicated that in all three comparisons, TorToiSe was perceived as significantly more natural than Tacotron 2 ($Z_1 = 2.86, p_1 = .004; Z_2 = 2.9, p_2 = .004; Z_3 = 2.01, p_3 = .036$).

Experimental prompts were generated with TorToiSe TTS. Each menu item contained two options to facilitate choice and avoid overloading the participants. For CA's feedback interventions, we

[1] https://github.com/neonbjb/tortoise-tts
[2] https://github.com/openai/DALL-E
[3] https://imagen.research.google/

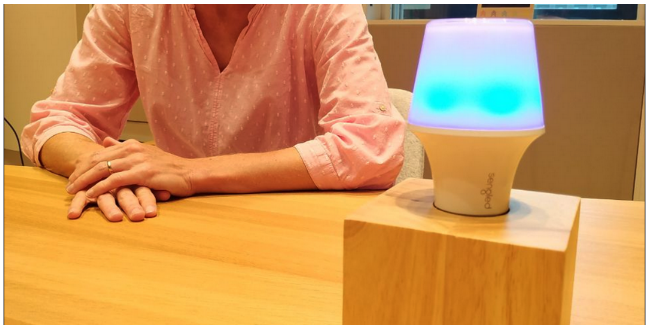

Figure 2: Food Genie (CA prototype) in 'active' mode.

decided to use prompts providing nutritional remarks regarding selected menu items. Specifically, we used the following two prompts: 'Just to let you know, your selection is high in cholesterol. Would you like to reconsider your choice?' (Unsolicited Feedback) and 'OK, here is your feedback. Your [MENU ITEM] is high in salt. Would you like to reconsider your choice?' (Solicited Feedback). We focused on cholesterol and salt as these nutrients are crucial for balanced diet and preventing cardiovascular diseases [3, 22], and therefore were likely to be considered as relevant to the participants. In the 'Unsolicited' feedback condition, feedback was provided after one of the menu items has been selected (randomised order), while in the 'Solicited' condition the users were asked if they wanted to hear feedback once all menu choices have been selected (see Figure 1 for an abstract illustration). For any out-of-scope queries, 'Sorry, this functionality is not supported at the moment.' prompt was used.

3.3 Procedure

The experiment was conducted as a Wizard of Oz (WoZ) scenario [5], where the CA was simulated by a member of the research team who selected synthesised prompts that were played through a wireless speaker with the *Sengled Solo* light-bulb (shown in Figure 2).

Participants were instructed to say 'Hey Genie' at the start of each task in order to initiate the conversation with the CA. We used this wake word in order to reflect the interaction conventions of modern smart speakers. Once the wake-word has been used, the colour of the bulb has changed to blue to indicate that the CA was active. We conducted three internal pilot studies to test and refine both the design and implementation of our CA in all experimental conditions. We also conducted additional training sessions for the Wizard to develop their competency and experience in running the studies and help to ensure that interaction consistency has been preserved across all participants.

The experiment took place in the HCI lab of the University of Luxembourg. Upon arrival to the lab, participants were briefed about the study and told that they will be interacting with a prototype of an interactive food ordering CA called Food Genie. The next stage was a food ordering task which consisted of three interactive food ordering scenarios, where participants interacted with the CA to book a three course meal. After each scenario, participants answered the question regarding their perception about how appropriate was Food Genie in delivering feedback. Finally, having completed all tasks, participants were invited to a semi-structured interview (due to space constraints we do not provide a full qualitative analysis in this paper), where we asked them questions about their experience with Food Genie. At the end, we informed them that the CA was operated by a human.

During each task, participants were instructed to interact with the CA to book a three-course meal (starter, main, and dessert). There were three tasks in total, each with a distinct type of menu. There were two food options available for each menu item.

Examples of participants' dialogues under each experimental condition are presented in Table 1.

3.4 Appropriateness Scale

After each food ordering scenario, participants were asked to rate the appropriateness of Food Genie's behaviour during the conversation on a scale from 1 (Very inappropriate) to 11 (Very appropriate). We did not provide the participants with any additional instructions beyond asking them to treat Food Genie as a proxy system for ordering their meal.

3.5 Participants and recruitment

Thirty participants took part in the experiment (16 F and 14 M). The average age of participants was 28 years (SD = 3.8). They were recruited through the internal network of our institution, targeting students and staff. Participants included both native English speakers and non-native speakers who were fluent in English. Due to the nature of our study (i.e., encouraging reflection on menu choices), to avoid unintended psychological risks, the inclusion criterion was that participants had not been diagnosed with a food disorder and were comfortable discussing food-related topics. Participants performed the assigned tasks in a dedicated laboratory room and were rewarded with a 30 EUR gift voucher upon completion of the experiment. The study was approved by the Ethics Review Panel of the University of Luxembourg with the ID: ERP 22 − 051 C21LL.

4 RESULTS

On average it took participants 7 minutes and 58 seconds (SD = 1 min and 40 s) to complete the three tasks: Baseline (M = 2 min and 16 s, SD = 35 s), Solicited feedback (M = 2 min and 55 s, SD = 35 s), and Unsolicited (M = 2 min and 46 s, SD = 49 s). The average time of a semi-structured interview was 17 min and 12 s (SD = 6 min and 6 s).

The data distribution of our independent variable is presented in Figure 3. An alpha level of .05 was used for all statistical tests. We should note that we cannot use ANOVA because our dependent variable is non-continuous and residuals are not normally distributed (verified by the Shapiro-Wilk test). Therefore, we used a linear mixed-effects (LME) model to analyse the data.

To control for the potential influence of confounding factors, presentation order is considered an interaction effect (model covariate) and users are considered a random effect. The model is coded in R as:

```
appropriateness ~ condition * presentation + (1|user)
```

where `condition` is the feedback strategy. We then ran pairwise comparisons (also known as *contrasts*) with Bonferroni-Holm correction to guard against over-testing the data. When referring to

Table 1: Examples of conversational strategies followed by Food Genie (FGE) when interacting with participants (PAR). Notes: In the Unsolicited condition, the feedback was provided after participants selected either their starter, main course, or dessert. In the Solicited condition, the feedback was only provided at the end of the session and only if a participant agreed to hear it.

Baseline	Unsolicited	Solicited
FGE: (Presents menu options) Which option would you like to choose? **PAR:** Option one. **FGE:** Thank you, all selections have been now registered. Would you like to update any of your choices? **[No Feedback]** **PAR:** No. **FGE:** Your order is now complete, thank you for using this service. (conversation ends.)	**FGE:** (Presents menu options) Which option would you like to choose? **PAR:** Option two. **FGE:** Thank you, your selection has been registered. Just to let you know, your selection is high in cholesterol. Would you like to reconsider your choice? **[Unsolicited Feedback]** **PAR:** No. **FGE:** Thank you, your selection has been registered. By the way, did you find this information relevant? **PAR:** No. **FGE:** Thank you, your selection has been registered. (continues...)	**FGE:** (Presents menu options) Which option would you like to choose? **PAR:** Option two. **FGE:** Thank you, your selection has been registered. Just to let you know, feedback is available for one of your selections. Would you like to hear it? **[Solicited Feedback]** **PAR:** Yes. **FGE**: Ok, here is your feedback. Your main course selection contains high level of salt. Would you like to reconsider your choice? **PAR:** Yes. **FGE:** Thank you, your selection has been registered. (continues...)

CA feedback strategies we use following abbreviations: Base (Baseline), Sol. (Solicited feedback), and Unsol. (Unsolicited feedback). Based on the results, it can be observed that participants tend to provide positive and extremely positive answers regardless of the condition.

We found a statistically significant difference between the baseline condition and the unsolicited feedback condition in terms of feedback appropriateness, where the baseline was perceived as significantly more appropriate ($p = .013$).

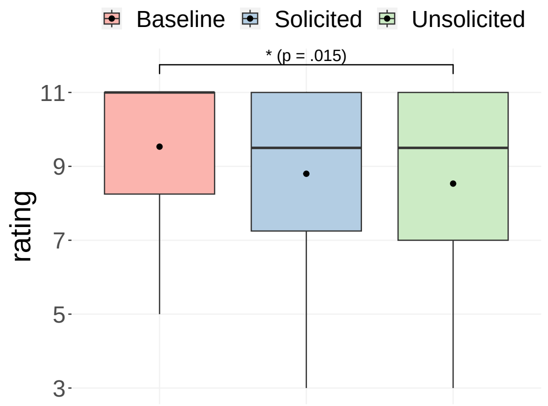

Figure 3: Boxplot comparing perception ratings regarding feedback appropriateness. Dots denote mean values.

All participants replied ‘yes’ to a solicited feedback offer. We did not find statistically significant differences between the baseline and solicited feedback on the appropriateness scale ($t.ratio = -2.051, p = .09$), and thus we reject **H1**.

We found statistically significant differences between the baseline condition and the unsolicited feedback on the appropriateness scale, where the baseline was perceived as significantly more appropriate than the unsolicited feedback condition ($t.ratio = -2, 927, p = .015$). Therefore we found the evidence in support of **H2**.

5 DISCUSSION

As noted in Section 3.1, the results of previous studies indicated openness of users to receiving proactive feedback from their CAs [20, 34] and positive perception of CAs that provide solicited health-related suggestions [29, 35]. This led us to hypothesise that solicited feedback will be considered as more appropriate than no feedback (**H1**), while the unsolicited feedback will be considered as less appropriate than no feedback (**H2**).

We have not found support for our H1, with no statistically significant difference between solicited feedback and the baseline condition. This result can be attributed to the so called ‘ceiling effect’, since participants’ appropriateness ratings for the baseline condition approached the highest possible score. While this result goes against our assumption, it also indicates that solicited feedback may be considered equally appropriate as the baseline, status quo, condition. On the other hand, as hypothesised in H2, the baseline strategy has been found significantly more appropriate than unsolicited feedback (cf. Figure 3). This result could be linked to scepticism regarding CA making dietary suggestions reported by Luria et al. [20] and concerns regarding participants’ agency presented by Reicherts et al. [29] and Zargham et al. [35], or the belief that a CA should not have, or express its own views [8]. While participants of our study were more in favour of receiving feedback from a CA rather than a human (9 out of 13 who expressed opinion on the subject during the semi-structured interview), some found CA feedback inappropriate (e.g., P6 ‘Why are you saying bad things about the food that I am going to eat?’). Other participants have also questioned the authority of the agent to provide them with this kind of feedback (e.g., P2: ‘Who are you to be telling me that?’). It could be argued that in both feedback conditions, Food Genie violated participants social expectations regarding conversational agents, which consequently yielded lower appropriateness scores.

The fact that solicited feedback was not considered as more appropriate than the baseline may be explained by the timing of the intervention. In solicited condition, the CA's feedback was provided once all of the selections have been made (cf. Table 1), which potentially could have created an impression that the CA is hiding something from the participants by not disclosing the feedback immediately. As indicated by Edwards et al. [9], CA's spoken interruptions should be delivered sooner if the task is considered urgent.

5.1 Limitations

We are mindful that our study is subject to some limitations. First, TorToiSe does not support real-time speech synthesis, which led us to design a WoZ scenario with a limited number of menu options to choose from. However, this design decision provided us with more control over the experiment and helped to ensure high consistency between trials. Second, we only used a male voice in our experiment and left the exploration of the impact of female-synthesised voices with the same software for future work, as the perceived appropriateness of an agent may vary based on its gender [25, 31]. Nonetheless, it should be noted that the majority of participants (N = 19) remarked that the voice of Food Genie was very natural, friendly and pleasant to listen to.

6 CONCLUSION AND FUTURE WORK

We have investigated how a proactive feedback behaviour of a voice-only CA affects its perceived appropriateness through an interactive ordering scenario. We found that unsolicited feedback strategy was perceived as less appropriate than the baseline condition (no feedback). While our findings are preliminary, the investigation of the impact of providing feedback during decision-making tasks is pertinent, as CAs are starting to exhibit more proactive capabilities and thus have potential to influence and even modify the user's behaviour.

In future work we will explore the relationship between the perceived appropriateness of a CA and its trustworthiness and persuasiveness, and examine how these qualities map to participants' behaviour in decision making scenarios beyond the one we have explored in this paper. We will also conduct a qualitative coding of the data collected during the semi-structured interviews to obtain a better understanding of the participants' perceptions of Food Genie, as well as their concerns and expectations regarding proactive conversational support.

ACKNOWLEDGMENTS

We thank Sophie Doublet for helping with the semi-structured interviews. We also thank Lorena Sanchez Chamorro, Johannes Schöning, and Benjamin Cowan for their helpful feedback on earlier versions of this paper. This research is supported by the Horizon 2020 FET program of the European Union (grant CHIST-ERA-20-BCI-001), the European Innovation Council Pathfinder program (SYMBIOTIK project, grant 101071147), and the Luxembourg National Research Fund (FNR) IPBG2020/IS/14839977/C21 and IS/14717072.

REFERENCES

[1] Tawfiq Ammari, Jofish Kaye, Janice Y Tsai, and Frank Bentley. 2019. Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Trans. Comput. Hum. Interact.* 26, 3 (2019), 17–1.

[2] Yulia E Chentsova Dutton. 2012. Butting in vs. being a friend: Cultural differences and similarities in the evaluation of imposed social support. *J. Soc. Psychol.* 152, 4 (2012), 493–509.

[3] Michael Chourdakis, Thrasivoulos Tzellos, Chryssa Pourzitaki, Konstantinos A Toulis, George Papazisis, and Dimitrios Kouvelas. 2011. Evaluation of dietary habits and assessment of cardiovascular disease risk factors among Greek university students. *Appetite* 57, 2 (2011), 377–383.

[4] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, et al. 2019. What makes a good conversation? Challenges in designing truly conversational agents. In *Proc. CHI.* 1–12.

[5] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies—why and how. *Knowl Based Syst.* 6, 4 (1993), 258–266.

[6] Janna T Deelstra, Maria CW Peeters, Wilmar B Schaufeli, Wolfgang Stroebe, Fred RH Zijlstra, and Lorenz P Van Doornen. 2003. Receiving instrumental support at work: when help is not welcome. *J. Appl. Psychol.* 88, 2 (2003), 324.

[7] Tiffany D Do, Ryan P McMahan, and Pamela J Wisniewski. 2022. A New Uncanny Valley? The Effects of Speech Fidelity and Human Listener Gender on Social Perceptions of a Virtual-Human Speaker. In *Proc. CHI.* 1–11.

[8] Philip R Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R Cowan. 2019. Mapping perceptions of humanness in intelligent personal assistant interaction. In *Proc. MobileHCI.* 1–12.

[9] Justin Edwards, Christian Janssen, Sandy Gould, and Benjamin R Cowan. 2021. Eliciting spoken interruptions to inform proactive speech agent design. In *Proc. CUI.* 1–12.

[10] Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Jia Ye, R. J. Skerry-Ryan, and Yonghui Wu. 2021. Parallel Tacotron 2: A Non-Autoregressive Neural TTS Model with Differentiable Duration Modeling. *CoRR* abs/2103.14574 (2021).

[11] Andrew Gibiansky, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. 2017. Deep voice 2: Multi-speaker neural text-to-speech. *Adv Neural Inf Process Syst.* 30 (2017).

[12] Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Acad. Manag. Ann.* 14, 2 (2020), 627–660.

[13] Daena J Goldsmith. 2000. Soliciting advice: The role of sequential placement in mitigating face threat. *Commun. Monogr.* 67, 1 (2000), 1–19.

[14] Ted Grover, Kael Rowan, Jina Suh, Daniel McDuff, and Mary Czerwinski. 2020. Design and evaluation of intelligent agent prototypes for assistance with focus and productivity at work. In *Proc. IUI.* 390–400.

[15] Dietmar Jannach. 2022. Evaluating conversational recommender systems. *Artif. Intell. Rev.* (2022).

[16] Matthias Kraus, Nicolas Wagner, and Wolfgang Minker. 2020. Effects of proactive dialogue strategies on human-computer trust. In *Proc. UMAP.* 107–116.

[17] Matthias Kraus, Nicolas Wagner, Nico Untereiner, and Wolfgang Minker. 2022. Including Social Expectations for Trustworthy Proactive Human-Robot Dialogue. In *Proc. UMAP.* 23–33.

[18] Katharina Kühne, Martin H Fischer, and Yuefang Zhou. 2020. The human takes it all: Humanlike synthesized voices are perceived as less eerie and more likable. evidence from a subjective ratings study. *Frontiers in neurorobotics* (2020), 105.

[19] Ewa Luger and Abigail Sellen. 2016. " Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. In *Proc. CHI.* 5286–5297.

[20] Michal Luria, Rebecca Zheng, Bennett Huffman, Shuangni Huang, John Zimmerman, and Jodi Forlizzi. 2020. Social boundaries for personal agents in the interpersonal space of the home. In *Proc. CHI.* 1–12.

[21] Erina L MacGeorge, Bo Feng, Ginger L Butler, and Sara K Budarz. 2004. Understanding advice in supportive interactions: Beyond the facework and message evaluation paradigm. *Hum. Commun. Res.* 30, 1 (2004), 42–70.

[22] Janne A Martikainen, Erkki JO Soini, David E Laaksonen, and Leo Niskanen. 2011. Health economic consequences of reducing salt intake and replacing saturated fat with polyunsaturated fat in the adult Finnish population: estimates based on the FINRISK and FINDIET studies. *Eur J Clin Nutr.* 65, 10 (2011), 1148–1155.

[23] Manolis Mavrikis, Beate Grawemeyer, Alice Hansen, and Sergio Gutierrez-Santos. 2014. Exploring the potential of speech recognition to support problem solving and reflection. In *Int. J. Technol. Enhanc. Learn.* Springer, 263–276.

[24] National Public Media. 2022. The Smart Audio Report | National Public Media — nationalpublicmedia.com. https://www.nationalpublicmedia.com/insights/reports/smart-audio-report/. [Accessed 04-Apr-2023].

[25] John W Mullennix, Steven E Stern, Stephen J Wilson, and Corrie-lynn Dyson. 2003. Social perception of male and female computer synthesized speech. *Comput. Hum. Behav.* 19, 4 (2003), 407–424.

[26] Lingyun Qiu and Izak Benbasat. 2009. Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems. *J Manag Inf Syst* . 25, 4 (2009), 145–182.

[27] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proc. ICML.* PMLR, 8821–8831.

[28] Leon Reicherts, Yvonne Rogers, Licia Capra, Ethan Wood, Tu Dinh Duong, and Neil Sebire. 2022. It's Good to Talk: A Comparison of Using Voice Versus Screen-Based Interactions for Agent-Assisted Tasks. *ACM Trans. Comput. Hum. Interact.* 29, 3 (2022), 1–41.

[29] Leon Reicherts, Nima Zargham, Michael Bonfert, Yvonne Rogers, and Rainer Malaka. 2021. May I Interrupt? Diverging Opinions on Proactive Smart Speakers. In *Proc. CUI*. 1–10.

[30] Julian B Rotter. 1980. Interpersonal trust, trustworthiness, and gullibility. *American psychologist* 35, 1 (1980), 1.

[31] Katie Seaborn, Norihisa P Miyake, Peter Pennefather, and Mihoko Otake-Matsuura. 2021. Voice in human–agent interaction: a survey. *ACM Comput. Surv.* 54, 4 (2021), 1–43.

[32] Ben Shneiderman. 2020. Human-centered artificial intelligence: Three fresh ideas. *AIS Trans. Hum.-Comput. Interact.* 12, 3 (2020), 109–124.

[33] Kathleen S Verderber, Rudolph F Verderber, and Cynthia Berryman-Fink. 2004. *Inter-act: Interpersonal communication concepts, skills, and contexts.* Oxford University Press New York.

[34] Sarah Theres Völkel, Daniel Buschek, Malin Eiband, Benjamin R Cowan, and Heinrich Hussmann. 2021. Eliciting and Analysing Users' Envisioned Dialogues with Perfect Voice Assistants. In *Proc. CHI*. 1–15.

[35] Nima Zargham, Leon Reicherts, Michael Bonfert, Sarah Theres Völkel, Johannes Schöning, Rainer Malaka, and Yvonne Rogers. 2022. Understanding Circumstances for Desirable Proactive Behaviour of Voice Assistants: The Proactivity Dilemma. In *Proc. CUI*.