“Hey Genie, You Got Me Thinking about My Menu Choices!” Impact of Proactive Feedback on User Perception and Reflection in Decision-making Tasks

MATEUSZ DUBIEL, LUIS A. LEIVA, KERSTIN BONGARD-BLANCHY, and ANASTASIA SERGEEVA, University of Luxembourg, Esch-sur-Alzette, Luxembourg

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 users’ reflection on choices in voice-only decision-making tasks. We conducted a within-subjects experiment to evaluate the effect of CA’s feedback delivery strategy at three levels (no feedback, unsolicited and solicited feedback) and the impact on users’ likelihood of changing their choices in an interactive food ordering scenario. We discovered that in both feedback conditions the CA was perceived to be significantly more persuasive than in the baseline condition, while being perceived as significantly less confident. Interestingly, while unsolicited feedback was perceived as less appropriate than the baseline, both types of proactive feedback led participants to relisten and reconsider menu options significantly more often. Our results provide insights regarding the impact of proactive feedback on CA perception and user’s reflection in decision-making tasks, thereby paving a new way for designing proactive CAs.

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

Additional Key Words and Phrases: Conversational Agents; Synthetic Speech; Design Ethics; Personalisation; Trust

ACM Reference format:
Mateusz Dubiel, Luis A. Leiva, Kerstin Bongard-Blanchy, and Anastasia Sergeeva. 2024. “Hey Genie, You Got Me Thinking about My Menu Choices!” Impact of Proactive Feedback on User Perception and Reflection in Decision-making Tasks. *ACM Trans. Comput.-Hum. Interact.* 31, 5, Article 56 (November 2024), 30 pages.
https://doi.org/10.1145/3685274

1 Introduction

Conversational Agents (CAs) such as Amazon Alexa, Apple Siri, or Google Home are becoming increasingly ubiquitous. Such CAs are currently available through various devices such as computers,

This research was 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 Luxembourg’s National Research Fund (FNR) through Grants IPBG2020/IS/14839977/C21 and IS/14717072.

Authors’ Contact Information: Mateusz Dubiel (corresponding author), University of Luxembourg, Esch-sur-Alzette, Luxembourg; e-mail: mateusz.dubiel@uni.lu; Luis A. Leiva, University of Luxembourg, Esch-sur-Alzette, Luxembourg; e-mail: luis.leiva@uni.lu; Kerstin Bongard-Blanchy, University of Luxembourg, Esch-sur-Alzette, Luxembourg; e-mail: kerstin.bongard-blanchy@lmddc.lu; Anastasia Sergeeva, University of Luxembourg, Esch-sur-Alzette, Luxembourg; e-mail: anastasia.sergeeva@uni.lu.

ACM 1557-7325/2024/11-ART56
https://doi.org/10.1145/3685274

smartphones and smart speakers [107]. According to The Smart Audio Report, in April 2022 over 82 million people owned a smart speaker in the United States alone [93]. While CAs are still predominantly used for simple tasks such as checking the weather, playing music or setting alarms [4], a growing number of users are expecting to use them routinely for purchasing products and services online [93, 125]. An analogous trend can be also observed for personalised conversational recommender systems [60, 65].

While personalised technologies have potential to support users in achieving behavioural change goals such as increasing physical activity [73] or improving productivity at work [71], they can also lead to complacent behaviour and lack of reflection [58, 128]. Lately, the prevalence of personalised technologies has become even more evident, as the COVID-19 pandemic resulted in a rapid increase in demand for touch-free conversational interactions [93, 126]. Yet, the impact of voice-only CA on user behaviour still remains under-researched.

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, for example, have led to a rapid improvement of the quality of synthetic voices in terms of intelligibility and naturalness, making them almost indistinguishable from human speech [49]. Research shows that CAs that sound like humans are generally perceived as significantly more trustworthy [40, 109], more pleasant to listen to [34] and likeable [76] 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 [29].

Proactivity in CAs can be considered as an autonomous initiation of a voice-based action, such as providing reminders, recommendations, or nudges, taken by the agent to support the user [104]. Since CAs are increasingly being used for transactional tasks that can affect users' agency [31], it is timely to explore how different proactive feedback interventions (i.e., feedback provided by the agent without user's request) could trigger reflection and, in turn, impact upon choices that users make in such tasks. Research shows that users tend to report increased trust in systems' capabilities as their familiarity with the system grows. In online shopping contexts, for example, a recent study found that an increased facility of purchasing products via 'one-click' buttons led to an increase in impulsive buying behaviours [58]. It has been also demonstrated that, if not treated, overtime, impulsive buying behaviour can lead to shopping addiction [103, 135]. Yet, proactive interventions that offer feedback to participants have potential to support participants to make better informed choices [7]. As highlighted by Penha et al. [105], providing more information in a voice purchasing scenario can help users to make better decisions. Research also shows that children are more likely to be influenced by a CA than a human [2].

Following Cox et al. [23], who advocated use of design frictions to support 'mindful' interactions, in this work we explore whether proactive CA's feedback interventions could foster participants' reflection in a food ordering scenario, consequently slowing down their decision-making process and potentially making it less impulsive. Our motivation for choosing a food ordering scenario is to provide a decision-making task that is familiar to participants and reflective of the emerging capabilities of CAs, while also offering an opportunity to apply proactive interventions to make participants reconsider their choices. Moreover, while many food decisions are made without much cognitive effort [25], developing a healthy diet is fundamental to long-term well-being [38].

Building on previous research that elicited user's expectations regarding proactive CA support (e.g., [21, 30, 83, 84, 114, 132, 137]), our work focuses on an interactive, voice-only decision-making scenario to explore the impact of proactive interventions of a CA. Specifically, we investigate (1) how proactively providing feedback regarding menu options affects perceived trustworthiness, persuasiveness and appropriateness of the CA and (2) if such feedback can foster reflection regarding selected menu choices. We use a simple food ordering scenario that provides a high degree of control over the interaction and allows us to test the impact of agent's interventions on participants' decisions.

In sum, the contribution of this article is two-fold:

(1) We provide empirical insights on the impact of proactive CA interventions on users' perceptions and behaviour in a standardised, voice-only decision-making scenario which approximates capabilities of present-day CAs.
(2) We offer design suggestions for CAs that can foster user reflection and slow down their decision-making process.

Our investigation indicates that proactive voice-only feedback has potential to trigger users' reflection in a food ordering task. We find that users consider proactive interventions of the CA as less appropriate than the baseline (no feedback). Nonetheless, we observed that both types of proactive feedback resulted in a more frequent relistening behaviour and more frequent changes in selected menu items. In addition, unsolicited feedback results in longer deliberation than solicited feedback, which indicates its potential to make users reflect on their choices. To the best of our knowledge, this is the first time that different forms of feedback provided by voice-only CAs have been evaluated in a decision-making context *while* considering both users' perception and corresponding behavioural outcomes. Our investigation builds on previous work on CAs which highlighted both openness and concerns of users regarding proactive feedback (discussed in Section 2.3.1), by applying a mixed-method approach and testing these expectations in a food ordering task.

As CAs are becoming more proactive, understanding the impact of different feedback interventions on the agent perception and user behaviour can shed light on their potential to foster reflection and provide assistance during simple decision-making tasks. It is our hope that our study will inform and help to advance future research developments in this area.

2 Background and Related Work

Speech has been shown to be an effective tool for promoting reflection in the educational context [89], improving focus on task [56] and increasing participants' involvement in an exploratory data analysis task [113]. While **Human–Computer Interaction (HCI)** researchers highlight that the role of **Artificial Intelligence (AI)** should be to empower people and amplify their skills rather than fully automate every task [121], *how* to provide the most effective and appropriate proactive support for the users during interactions with intelligent systems remains an open research problem [113]. To illustrate the complexity of this problem, here we present key concepts related to perception of proactive CAs (Section 2.1), followed by an overview of methods for promoting reflection and their potential benefits for users (Section 2.2), and conclude with discussion of some relevant research studies (Section 2.3).

2.1 Relevant Concepts and Constructs

In decision-making contexts, trust, benevolence, social influence and agency have been identified as the key factors that affect user's perception of CA's recommendations and their likelihood to rely on them [45, 48, 57, 66, 77, 79, 133]. In the following we summarise the major findings in this regard.

2.1.1 Trust. Trust is a complex and multi-layered construct with various definitions. One part of the academic community argues that trust should be considered as a 'belief' or an 'attitude' [79, 81], while the other part advocates a behavioural approach [87, 123]. In our work, we consider trust as a combination of both—an attitude that could be formed by beliefs that can in turn inspire certain behavioural outcomes (e.g., selection of menu choices). As noted by Porcheron et al. [108], lack of distinct behavioural scales makes measuring trust a challenging task. What is more, there may be a disconnect between perceived trust and user's trusting behaviour due to personal perceptions of a CA such as scepticism towards technology [5]. Some users can also be more inclined to share information with a more persuasive system, despite not trusting it [48]. Thus, combining self-reported measures (subjective metrics presented in Section 3.3) and behavioural measures (objective metrics presented in Section 3.4) offers a promising approach for CA evaluation, by providing a broader and more comprehensive picture. In this study, we combine perceived social impression of the agent (measured via questionnaire, explained in Section 3.3.1) with participants' willingness to reconsider their menu choices (determined by hesitations, relistening behaviour, and changes in selected menu items).

2.1.2 Benevolence. Often described as a part of the trust construct, benevolence can be summarised as a belief that one of the interaction's agents (a human [124], a computer [57] or a company [44]) is concerned about the well-being of the other, and motivated to create a mutually beneficial interaction. In HCI, it could be described as the user's belief that the system is acting in the user's best interest and is willing to help them to achieve their goals [57]. Since we examine the extent to which CAs could be considered by users as appropriate sources of feedback, we decided to analyse this dimension of trust separately. Our expectation is that the CA that provides advice to the user could be perceived as more benevolent compared to one that does not.

2.1.3 Social Influence. Related to trust and benevolence is the concept of 'social influence' which corresponds to individual's opinions, emotions and behaviours being affected by other people [77]. Social influence stems from Social Cognitive Theory, which posits that people are more likely to follow guidance from someone whom they trust and feel connected to [8]. In the context of HCI, social acceptability and appropriateness determine how the guidance should be provided. For example, in previous research, proactive robot behaviour was assessed in context of approaching a human in an appropriate manner [47, 66, 133]. In our study, we use a proxy term 'social impression' to refer to key perceived CA qualities that influence participants' behaviour in a decision-making task.

2.1.4 Reactance. The concept of Psychological Reactance arose from the fact that people tend to act against the interventions they perceive as a loss of their agency [97]. It could be perceived as a situational reaction factor to social or therapeutic intervention, but it can also be understood as a personality construct [64]. In that sense, people could have a generally higher or lower level of psychological reactance. In our study, the interaction with the CA involves different types of interventions, so we decided to use the psychological reactance as a co-variate variable, which could affect the participant's perception of the CA.

2.1.5 Parameters of Social Impression of the Agents. Social impression refers to the process of forming an impression of someone's personality, based on their appearance and behaviour [6]. This concept originated in social psychology and primarily applies to human-to-human interactions. However, studies have shown that people also extend it, among others, to robots [62], chatbots [11], voice interfaces [115] and virtual agents [16]. Factors contributing to a positive social impression are often considered as proxies for the user's trust in the agent [116]; therefore, they should be given high priority in its design.

Beyond the general quality of agents, connected with their functionality and ability to perform conversational tasks, previous work highlighted the importance of the agent being able to recognise and adequately react to the user's emotional state [14]. Moreover, it also indicated that keeping the interaction in style consistent with human's perception of personality features improves the interaction with the agent [110]. Following the findings in the field of human-to-human interaction, Kim et al. proposed competence and warmth as two key dimensions of the social impression of an agent [68]. Furthermore, drawing from the human-to-human educational process, previous work also suggest that an agent with not just a warm but also an enthusiastic personality can effectively facilitate interactions with users when acting as a coach or guide [131]. Speaking about competence in the agents, it is often operationalised through making the agent to appear 'intelligent' and 'expert' [13]. At the same time, previous studies in human-to-human presentation showed the connection between how confident the speaker presents some information and their perceived competence [119], which can be relevant in interactions with voice-only CAs.

2.1.6 Agency. User agency refers to their ability to actively participate in automated decision-making process [5]. In the context of CA interactions, user agency is crucial for retaining control over the decision-making process and avoiding being manipulated by the system [45]. In our work, the CA's feedback is considered as a proactive intervention that is designed to foster reflection and make users think about their choices.

2.1.7 Hesitations. Hesitations, also referred to as 'response latency,' are an implicit behavioural measure that is commonly used in social psychology research to quantify underlying mental representations, cognitive processes and motivation [102]. The context and nature of the task, as well as experimental design, are factors that determine the meaning of this measure and make it both a useful and an universal tool [42]. Nonetheless, it has to be noted that it is not possible to determine with full certainty what cognitive processes are indicated by hesitations. Here, we use hesitations as a proxy of participants' reflection, which indicates that participants take their time to consider the CA's feedback.

2.2 Promoting Reflection

Reflection can be defined as 'the act of thinking and re-evaluating prior actions, choice constructs, or available alternatives and search rules, triggered by conscious or unconscious experiences, resulting in the uptake of new perspectives' [9, p.78]. Among many applications, in HCI, reflection can be used to promote learning [28, 41], improve performance at work [71], support behavioural change [73], or enhance cognitive engagement [46]. While supporting reflection via interaction with computer-based systems is a long-standing field of research [10], still the potential of voice-based systems to actively assist users in their decision-making processes remains mostly unexplored. We seek to address this gap by investigating how proactive, voice-only CAs can support reflection.

2.2.1 Methods for Promoting Reflection. Bentvelzen et al. [12] summarise four design techniques for designing applications to trigger reflection, namely: (1) *conversation* (including conversation with technology), (2) *comparison* (helping users compare them with other users or themselves over time), (3) *temporal perspective* and (4) *discovery* (evoking users to discover something or use something in a new light) [12].

Even if the entire interaction with the system is not specifically designed to foster user reflection, several steps in the interaction flow may require reflective behaviour. This design for slow-thinking, reflective steps is widely discussed in HCI through the conception of design frictions [23], which are interface design solutions that slow down users and aim to prompt them to think about a specific step that they have taken. For example, an additional confirmation dialog box which appears when deleting an email can be considered a friction that aims to encourage the user to reflect on the action before it is executed [94].

Another important concept for incorporating reflection into digital interactions is nudging. A nudge is a part of the choice architecture that influences people's behaviour in the designer's preferred way, without constraining any options available to the user [127]. Digital mental health applications rely on both design frictions and nudging to provoke a thoughtful approach to therapeutic interventions [138]. Similarly, both approaches can also be used in well-being applications (e.g., for limiting the smartphone use) [3].

2.2.2 Reflective Potential of CAs. The statement that interactive systems should trigger users' reflection upon provided information is also widely discussed in the AI-assisted decision-making community, mostly in the contexts of excessive trust and overreliance on the systems, especially in high-stakes scenarios [15, 22, 51]. For example, when discussing AI-powered advice-giving systems, Miller advocated use of the advice-giving mode [95], which should not give users a straightforward solution but rather trigger their expertise in the domain to critically evaluate the AI's proposal. In a similar vein, Gajos and Mamykina [46] demonstrated that providing users with explanations rather than recommendations by an AI system can lead to higher cognitive engagement and consequently result in incidental learning.

At the same time, critical reflection on information can be beneficial beyond high-stakes scenarios. For example, in the domain of misinformation debunking support, Danry et al. [26] discussed a similar approach to the interventions proposed by Miller for high- and medium-stake scenarios and showed that triggering users' reflection via Socratic questioning (i.e., discovering answers by asking questions) can help users better recognise misinformation in online settings.

Moreover, critical reflection on information can not only be used to evaluate a piece of information itself critically but also as a proxy for a more reflective approach to users' personal behaviour across various domains. This potential for fostering reflection via conversations is highly recognised in the area of CAs for well-being and health. For example, Callejas and Griol [17] discussed the role of CAs in a mental health setting, including the delivery of reflective therapy. The goal of such therapy is to actively provoke users to self-reflect, and empower them to improve their mental health. Similarly, Kocielnik et al. [72] discussed text-based conversational interfaces (chatbots) as a medium to force users to reflect upon their health data.

A number of studies have also explored the potential of CAs in triggering reflection upon users' learning experiences and daily activities [67, 100], promoting energy saving behaviour [59, 129]. CAs promoting reflection are also broadly discussed in the context of helping people to avoid certain unwanted behaviours, such as compulsive smartphone use [82] or procrastination at work [69]. Nonetheless, despite their potential to support users, a positive behavioural change can be hindered if CAs provide responses that are considered inappropriate given the sensitivity of the topic (e.g., mental health) [96], or fail to adapt their own affective response to appropriately match the emotional state of the user.

The majority of the studies presented in the current section focused either on text-based interventions [15, 46, 67, 69, 82] or theoretical benefits of proactive support [17, 22, 51, 100], with only two studies exploring the impact of proactive interventions delivered via voice [59, 129]. Since human speech has potential to increase social presence [78] and facilitate more effective learning [90], it can be argued that proactive interventions delivered by voice-only CA can effectively support reflection.

2.3 Proactive CAs

Based on the theory vs. practice distinction, we divide studies on proactive CAs into two categories: (1) eliciting users' expectations regarding proactive support (e.g., [84, 132, 137]) and (2) empirical evaluations of CA prototypes (e.g., [33, 43, 134]).

2.3.1 Eliciting Users' Expectations. Luria et al. [84] 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, participants liked the idea of proactive agent but none of them was happy with an agent trying to limit their agency, for example by preventing them from ordering unhealthy food. In a similar vein, Zargham et al. [137] investigated when CA interventions are appropriate, they found that emergency support and health-related interventions are welcomed provided that the CA asks for user for permission to intervene. In another elicitation study, Völkel et al. [132] 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 and recommendations to solve complex problems.

The above studies, that highlight the openness of users to proactive CA's support, were our motivation to empirically investigate its social impression and impact on users' behaviour in a decision-making task.

2.3.2 Proactive CA Evaluation. Based on the interaction modality, empirical evaluations of CAs fall into two main categories; (1) voice-based and (2) text-based. In the following we present some relevant recent examples of both of these categories.

Voice-Based. In a Wizard of Oz [24] experiment, Dubiel et al. [33] investigated the impact of persuasive synthetic speech on CA perception and user behaviour. They found that while speech generated from a debating-style corpus was perceived as more persuasive than speech generated from an audio-book corpus, there was no significant effect on user behaviour—with participants following CA's recommendation same number of times in each condition. In a similar study, Fetwell et al. [43] developed a 'pushy' CA for broadening political views of users by exposing them to different stories across the range of political spectrum. The study showed that the CA was effective in attracting participants' attention, promoted reflection on viewpoints and fostered discussions with other members of the households where it was placed. Wei et al. [134] conducted a longitudinal study where a Google Home-based speaker with a proximity sensor proactively prompted participants to engage them in conversation using three different starters. The starters were: *baseline* (no opening starter), *earcon* (notification ringtone) and *utterance*. It was found that the earcon starter was the preferred way for the participants to start the conversation with the proactive speaker. The results of the study also indicated that mood, current activity, proximity to the speaker and co-presence of other people where the best indicators of participant's perceived availability and responsiveness, highlighting the importance of personal and contextual settings in implementing proactive interventions.

Text-Based. Most relevant to our investigation are studies by Musto et al. [99] and Tassiello et al. [126]. Musto et al. [99] proposed a natural language recommender for recipes that emphasises nutritional content, potential risks and health benefits of different dishes. Tassiello et al. [126] explored the role of a CA in the context of food choices, finding that users are more willing to purchase low-involvement products (i.e., cheap items such as a bottle of juice) rather than high-involvement products (e.g., a bottle of champagne), and that emotional involvement of the agent leads to bolder buying behaviour.

We study participants' expectations regarding voice-based CA's proactivity as expressed in previous work [84, 132, 137], and test different types of feedback interventions in an interactive decision-making task, in order to evaluate the impact on CA's proactive interventions on participants' perceptions of the agent, reflections, and, ultimately, undertaken choices. The main contribution of our work is that it provides a *full-simulation of in-person, voice-only interactions of* *users with a proactive CA that goes beyond online evaluations*, which are frequently limited to evaluation of isolated prompts lacking a broader context, and do not provide a real-time conversational experience.

3 Methods

We conducted a 3 × 2 within-subjects experiment to evaluate the effects of Food Genie, a CA that provides three feedback strategies: No Feedback (baseline), Unsolicited Feedback (without user's permission) and Solicited Feedback (with user's permission) on two sets of factors: (1) perceived trustworthiness, persuasiveness, and appropriateness of the agent, and (2) the number of times that menu item has been relistened and the number of times that a menu item has been changed. Each participant was exposed to each CA's feedback strategy once. The order of feedback strategies was randomised with the baseline strategy always provided as the starting condition.

3.1 Feedback Strategies

Research indicates that proactive dialogue strategies (i.e., taking initiative to actively provide feedback) can lead to more positive perception of a conversational partner, including more trust and higher compliance, as compared to passive interaction (no feedback) [50, 74, 75]. Furthermore, compared to unsolicited feedback, solicited feedback was found to be more satisfying [18], less face-threatening [53] (less harmful to one's self-image), and more likely to be utilised [27, 122]. These findings are inline with the Advice Response Theory [85, 86] which postulates that treats 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 the communication literature [130] 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 both user's perception of the CA (perception-based hypotheses) and lead to the user following the recommendations of the CA more frequently (behaviour-based hypotheses).

3.2 Research Hypotheses

Our literature review on human-human [18, 27, 53, 130] and human-computer dialogue [50, 74, 75, 117] indicates that proactive feedback is associated with initiative, which can lead to more trust and be perceived as more appropriate than no feedback, and in turn translate to higher compliance. Moreover, the results of previous research indicate that people consider CA's feedback valuable if it concerns urgent matters that affect their well-being (such as hazards and emergencies), or provides a thoughtful suggestion [84, 132, 137], and that automated feedback can motivate reflection [15, 112]. Therefore, based on the above insights, we formulate the following research hypotheses:

- *Perception-based hypotheses*: *A CA that provides feedback (both solicited and unsolicited) regarding food selections will be perceived as: (H1) more trustworthy and (H2) more persuasive than an agent that does not. While in terms of appropriateness, we hypothesised that: (i) a CA which provides solicited feedback will be perceived as more appropriate compared a CA with no feedback (H3a), and that (ii) a CA using unsolicited feedback will be perceived as less appropriate than a non-feedback CA (H3b).*
- *Behaviour-based hypotheses: A CA that asks the user for permission before providing feedback (solicited feedback) will lead to more reflection, resulting in (H4) more frequent relistening of menu options than a CA providing unsolicited feedback and a non-feedback CA. Moreover, a CA that provides solicited feedback will lead to more frequent reconsiderations of selected options than a CA providing unsolicited feedback and a non-feedback CA (H5).*

Table 1. Description of Experimental Variables

Variable	Survey Statement	Likert Scale	Range
Social impression			
Trust	Did Food Genie appear to be trustworthy?	Not at all Trustworthy to Very Trustworthy	[1–7]
Confidence	Did Food Genie sound confident?	Not Confident at all to Very Confident	[1–7]
Enthusiasm	Did Food Genie sound enthusiastic?	Not Enthusiastic at all to Very Enthusiastic	[1–7]
Persuasiveness	Was Food Genie persuasive?	Not Persuasive at all to Very Persuasive	[1–7]
Benevolence			
Interest	I believe that Food Genie will act in my best interest.	Strongly Disagree to Strongly Agree	[1–5]
Help	I believe that Food Genie will do its best to help me if I need help.	Strongly Disagree to Strongly Agree	[1-5]
Preferences	I believe that Food Genie is interested in understanding my needs and preferences.	Strongly Disagree to Strongly Agree	[1-5]
Cognitive load			
Confidence	I felt confident using the system	Strongly Disagree to Strongly Agree	[1–7]
Tense	I felt tense using the system	Strongly Disagree to Strongly Agree	[1–7]
Calm	I felt calm using the system	Strongly Disagree to Strongly Agree	[1–7]
Concentration	A high level of concentration is required when using the system	Strongly Disagree to Strongly Agree	[1–7]
Usage	The system is easy to use	Strongly Disagree to Strongly Agree	[1–7]
Appropriateness of feedback			
Behaviour	The behaviour of Food Genie during the conversation was:	Very Inappropriate to Very Appropriate	[1–11]
Satisfaction with selected menu options			
Satisfaction	Overall, I am satisfied with my menu choices	Strongly Disagree to Strongly Agree	[1–5]

These hypotheses are tested by using a combination of experimental variables (Table 1) that, together, provide key insights into how a CA like Food Genie is perceived by the user and, in turn, how likely the user is to rely on its feedback. In the following we describe them in detail.

3.3 Subjective Variables

3.3.1 Social Impression. To understand the parameters of the Social Impression of the three Food Genie variants, we adapted the Social Impression Scale from the MOS-X questionnaire [106]. The instrument shows adequate reliability and is widely used in voice system evaluations. In our study, the same voice was used in all of the experimental conditions, therefore we slightly modified the questions to focus on addressing the system as a whole, rather than focusing on the voice parameters alone. Further, given that the questions of the scale contained items that are close to the social impression factors found in the line of previous studies of persuasion in virtual agents [29, 98, 136], we decided to analyse each question separately and create a social impression profile of each agent's interaction in terms of trustworthiness, enthusiasm, persuasiveness and confidence.

3.3.2 Benevolence. As an additional instrument to assess the benevolence dimension of trust to the CAs, we used the Benevolence subscale from the Human-Computer Trust Scale, proposed by Gulati et al. [57]. The Human-Computer trust scale has high validity, reliability and predictive power [57]. For our study, we added the name and functions of our CA to the subscale's template.

3.3.3 Cognitive Demand. We used the **Subjective Assessment of Speech System Interfaces (SASSI)** Cognitive Demands scale [63] to determine if any interventions of Food Genie increased the cognitive load of the participants. The SASSI questionnaire is systematically used in the assessment of interactive speech interfaces [80].

3.3.4 Appropriateness Scale. To measure how appropriate the agent's behaviour was during each intervention, we used a one-question appropriateness scale, where we asked participants to rate the appropriateness of agent's behaviour during the conversation (from 1—Very inappropriate to 11—Very appropriate).

3.4 Objective Variables

3.4.1 Hesitations. We measured the time interval between the end of the audio prompt 'Would you like to reconsider your choice?' and the subject's response, and used it as a proxy of participants' reflection on the presented menu items.

3.4.2 Relistening Behaviour. As another proxy of reflection, we considered the number of times that participants have relistened to the menu options provided by Food Genie before committing to their final menu item choice.

3.4.3 Changing Choices. We considered the number of times that participants changed their menu item selection as a behavioural metric. Change of option was possible after the intervention of the CA and before the final confirmation request.

3.5 Materials

The voices used in the experiment were developed with the **TorToiSe text-to-speech (TTS)** software.[1] TorToiSe is inspired by the 'zero-shot text-to-image generation approach' [111]—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 very slow synthesis time (two minutes for 7–10 word sentences on average) at the time the study was conducted, owing to hardware limitations, it was 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 high fidelity synthetic speech that outperformed any alternative open source system in terms of naturalness and ability of reflect human emotions. 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 [39], 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 ($Z_1 = 2.86, p_1 = 0.004; Z_2 = 2.9, p_2 = 0.004; Z_3 = 2.01, p_3 = 0.036$).

TorToiSe TTS has been used to prepare experimental prompts. All prompts were then arranged following the order in which they were to be presented in the experiment (e.g., '1_*Introduction*', '2_*StarterA*', etc.) The interaction process for each experimental condition is presented in Figure 1.

[1] https://github.com/neonbjb/tortoise-tts

[2] https://github.com/openai/DALL-E

[3] https://imagen.research.google/

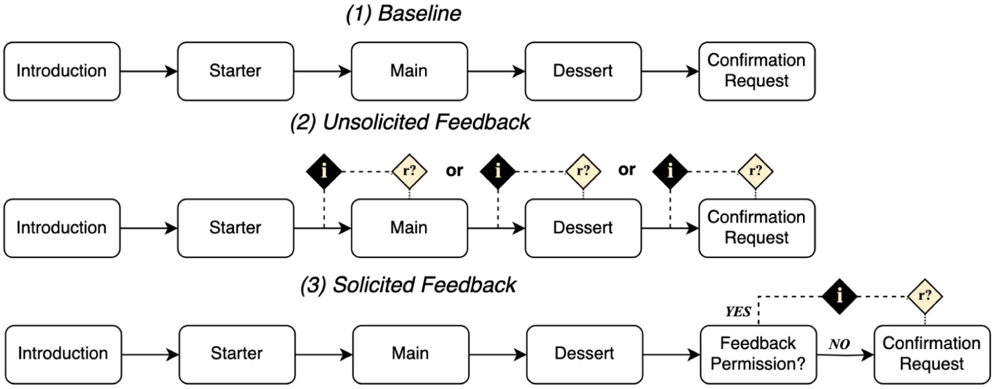


Fig. 1. Overview of Food Genie interaction process for the three experimental conditions: (1) Baseline, (2) Unsolicited Feedback and (3) Solicited Feedback. Note: Black diamonds indicate CA's feedback interventions and white diamonds indicate CA asking: 'Was this feedback relevant?'

Each menu item contained two options to facilitate choice and avoid overloading the participants. For CA's feedback interventions, we 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 a balanced diet and preventing cardiovascular diseases [19, 88], and therefore were likely to be considered as relevant to the participants. In the 'Unsolicited feedback' condition was provided after one of the menu items has been selected (randomised order), while in the 'Solicited feedback' condition the users were asked once all of their menu choices have been selected. Our rationale for providing solicited feedback at the end of the interaction was to make it less obtrusive and abrupt. Moreover, we wanted to make it easier for participants to distinguish between these two conditions given that they had to rely only on audio cues to distinguish between both feedback strategies. For any out-of-scope query, a 'Sorry, this functionality is not supported at the moment.' prompt was used.

3.6 Procedure

The experiment was conducted as a Wizard of Oz scenario [24], 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 (featured in Figure 2 on the right). At the beginning of interaction, once the wake-word was used, the colour of the bulb has changed to blue to indicate that the CA was active.

The experiment consisted of three stages: (1) a pre-interaction questionnaire; (2) a series of three interactive food ordering tasks, each followed by questionnaire on CA's social impression, benevolence, trust and appropriateness and (3) a semi-structured interview. The experiment took place in the HCI lab of the University of Luxembourg. One week prior to the experiment, participants were asked to fill in a pre-evaluation questionnaire that contained questions on demographics and reactance. To control the level of psychological reactance we used an 11-item version of the Hong Psychological Reactance Scale [64] as it provides better results than the 14-item one in factor structure and shows a good level of validity and reliability [120].

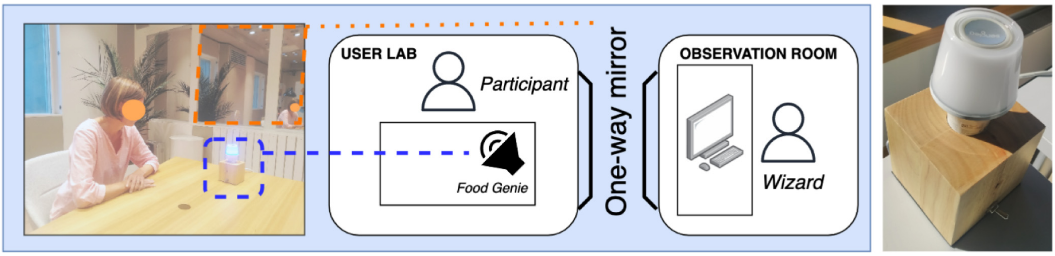


Fig. 2. Illustration of experimental setup, with participant facing Food Genie (left) and a close-up of the CA (right).

On the day of the experiment, upon arrival to our research facility, 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 search scenarios, where participants interacted with the CA to book a three course meal. After each scenario, participants filled in a series of questionnaires (see Table 1 for details). Finally, having completed all tasks, participants were invited to a semi-structured interview, where we asked them questions about their experience with Food Genie, including the interaction, CA interventions and the voice itself. 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. To avoid ordering effects, we randomised the tasks for each participant. There were two food options available for each menu item. We advised the participants that they should not speak when the CA is speaking. This step was taken to ensure consistency between the interaction and to replicate the functionality of current commercial systems that do not support 'barge-in'. An overview of the experimental setup is illustrated 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. Examples of participants' dialogues under each experimental condition are presented in Table 2 (see also Tables 1–3 in the supplementary materials for additional examples). 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 to help ensure that interaction consistency has been preserved across all participants.

3.7 Participants

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. We used a 5-point Likert-type scale to determine participants' level of familiarity with audio-based CAs: 1—'Not at all,' 2—'Rarely (around once or twice a month),' 3—'Sometimes (a couple times a week),' 4—'Once or twice a day,' 5—'Many times a day.' The data suggested that for most of the users the CA is not part of everyday experience (Mdn = 2). More detailed information about each participant is provided in Table 5 of the supplementary materials.

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

Table 2. Food Genie's Interaction Strategies Followed by Food Genie (FGE) When Interacting With Participants (PAR)

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:** No. **FGE:** Thank you, your selection has been registered. (continues...)

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.'

3.8 Design

An alpha level of 0.05 was used for all statistical tests. Considering that for most of our dependent variables there were departures from normality (verified by Shapiro-Wilk tests) and homoscedasticity (verified by Levene's tests) and since all dependent variables are discrete and have a natural order, we used a **Linear Mixed-Effects (LMEs)** model to analyse the data.

LME models are quite robust to violations of several distributional assumptions [118] and previous studies have demonstrated its suitability for analysing Likert-type data [70].

We fit the LME models and computed the estimated marginal means for specified factors. Each dependent variable is explained by each condition. The participants are considered a random effect. We also considered the effect of order and reactance, to control the potential influence of these factors. We then ran pairwise comparisons (also known as *contrasts* in LME parlance) with Bonferroni-Holm correction to guard against over-testing the data. When referring to CA's feedback strategies we use following abbreviations: Base (Baseline), Sol. (Solicited feedback) and Unsol. (Unsolicited feedback).

4 Results

4.1 Descriptive Statistics

On average it took participants 7 min and 58 s (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 feedback (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 the main variables under study is presented in Figure 3. Based on the results, we can see that for the social impression and appropriateness questions, participants tend to provide positive and extremely positive answers regardless of the condition.

A summary of the multiple comparison component of the LMEs is given in Table 3. The results show statistically significant differences between the baseline and both feedback conditions (unsolicited and solicited), with both feedback conditions being more persuasive ($p \leq 0.03$) but less confident ($p \leq 0.011$) than the baseline. We also found a statistically significant difference

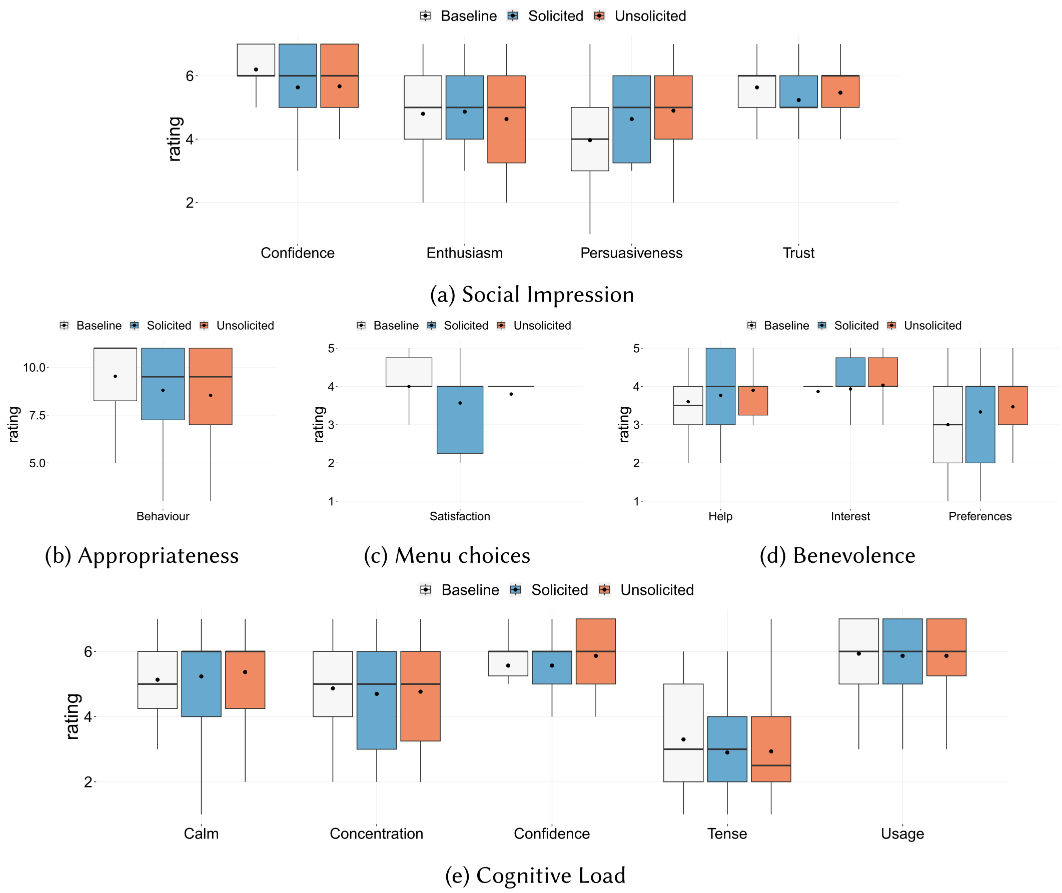
Table 3. Summary of Contrasts Tests

Contrast	**Estimate**	**SE**	**df**	***t*.ratio**	***p*.value**	**Effect Size**
Social Impression: Trustworthiness						
Unsol. - Sol.	0.233	0.176	56	1.327	.379	.174
Unsol. - Base	−0.167	0.176	56	−0.948	.379	.125
Sol. - Base	−0.400	0.176	56	−2.276	.080	.290
Social Impression: Confidence						
Unsol. - Sol.	0.033	0.186	56	0.179	.858	.023
Unsol. - Base	−0.533	0.186	56	−2.865	**.011***	.357
Sol. - Base	−0.566	0.186	56	−3.044	**.010***	.376
Social Impression: Enthusiasm						
Unsol. - Sol.	−0.233	0.206	56	−1.131	.789	.149
Unsol. - Base	−0.166	0.206	56	−0.808	.845	.107
Sol. - Base	0.066	0.206	56	0.323	.845	.043
Social Impression: Persuasiveness						
Unsol. - Sol.	0.267	0.271	56	0.985	.328	.130
Unsol. - Base	0.933	0.271	56	3.449	**.003***	.418
Sol. - Base	0.667	0.271	56	2.464	**.033***	.312
Appropriateness of feedback						
Unsol. - Sol.	−0.267	0.342	56	−0.779	.439	.103
Unsol. - Base	−1.000	0.342	56	−2.920	**.015***	.363
Sol. - Base	−0.733	0.342	56	−2.142	.073	.275
Satisfaction with selected menu options						
Unsol. - Sol.	0.233	0.193	56	1.206	.465	.159
Unsol. - Base	−0.200	0.193	56	−1.034	.465	.136
Sol. - Base	−0.433	0.193	56	−2.240	.087	.286
Benevolence						
Unsol. - Sol.	0.367	0.414	56	0.885	.380	.117
Unsol. - Base	0.933	0.414	56	2.252	.084	.288
Sol. - Base	0.567	0.414	56	1.367	.354	.179
Cognitive load						
Unsol. - Sol.	0.533	0.442	56	1.207	.697	.159
Unsol. - Base	0.000	0.442	56	0.000	1.00	.000
Sol. - Base	−0.533	0.442	56	−1.207	.697	.159

Note: p-values adjusted by the Bonferroni-Holm method for 3 tests.

between the baseline condition and the unsolicited feedback condition in terms of appropriateness, where the baseline was perceived as significantly more appropriate ($p = 0.013$). All other comparisons were not found to be statistically significant, however effect sizes suggest a moderate practical importance of the results when comparing solicited feedback vs. baseline in terms of trust ($r = 0.294$), persuasiveness ($r = 0.317$) and appropriateness ($r = 0.267$). Therefore, our results suggest an actual difference between solicited feedback and no feedback in practice, for most of the dependent variables analysed.

4.1.1 Cognitive Demand Check. The data showed no statistically significant differences between conditions in terms of cognitive demand (Unsol. - Sol.: $t.ratio = 1.207, p = .697$; Unsol. - Base: $t.ratio = 0.000, p = 1.00$; Sol. - Base: $t.ratio = -1.207, p = 0.697$). This supports our modelling assumption that our interventions do not significantly differ in terms of the cognitive load imposed on participants, which otherwise could have distorted the experimental results. If the conditions



(a) Social Impression

(b) Appropriateness

(c) Menu choices

(d) Benevolence

(e) Cognitive Load

Fig. 3. Boxplots comparing 'Food Genie' perception ratings. Dots denote mean values.

were significantly different in terms of cognitive load, we could attribute the relistening behaviour not to a reflection on the proposed choices, but merely to an attempt to understand the content of the message.

4.1.2 Model Covariates. As hinted previously, we included order and reactance into the LME models as fixed effects. No statistically significant differences were found against the LME models that did not consider these (neither as covariates nor random effects). Therefore, we conclude that neither order nor reactance have had an effect on the results. We report our findings using LMEs that consider order and reactance as model covariates, and users as random effects.

4.2 Main Findings

We did not find statistically significant differences between either feedback condition and the baseline in the metrics related to trustworthiness (Unsol. - Base: $t.ratio = -0.948, p = 0.379$; Sol.-Base: $t.ratio = -2.276.p = 0.080$) and benevolence (Unsol. - Base: $t.ratio = 2.252, p = 0.084$; Sol. - Base: $t.ratio = 1.367, p = 0.354$). Therefore we reject hypothesis *H1*. Nonetheless, the effect sizes suggest a moderate practical importance of the results when comparing solicited feedback vs. baseline in terms of trustworthiness ($r = 0.290$).

The data show statistically significant differences between baseline and both intervention conditions in terms of persuasiveness (Unsol. - Base: $t.ratio = 3.449, p = 0.003$; Sol. - Base: $t.ratio = 2.464, p = 0.033$). That is, both interventions are perceived as more persuasive than the baseline, therefore we validate *H2*.

We did not find statistically significant differences between the baseline and solicited feedback on the appropriateness scale ($t.ratio = -2.142, p = 0.073$), and thus we reject *H3a*. Nonetheless, against our assumption, a moderate effect size ($r = 0.275$) indicates that solicited feedback might have been perceived as less appropriate than the baseline.

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 ($t.ratio = -2.920, p = 0.015$). Therefore we found the evidence in support of *H3b*.

We run χ^2 tests to analyse the effect of feedback type (no feedback, solicited and unsolicited) on participants' behaviour in selecting menu choices, both in terms of relistening of menu options (*H4*) and reconsideration of recommendations (*H5*). For relistening behaviour, pairwise *post hoc* tests of proportions (Bonferroni-Holm corrected) revealed statistically significant differences between baseline condition and both solicited ($p < 0.01$) and unsolicited ($p < 0.001$) feedback. In both cases, proactive feedback led to significantly more relistening compared with the baseline. However, we did not find a statistically significant difference between solicited and unsolicited feedback ($p > 0.05$). Therefore, we partially confirm H4. For reconsideration behaviour, in contrast to our hypothesis, we observed the same results as in relistening behaviour. The baseline condition was significantly different than solicited ($p < 0.01$) and unsolicited ($p < 0.001$) feedback. While proactive feedback provided significantly more reconsiderations, no differences were found between solicited and unsolicited feedback ($p > 0.05$). Therefore, we partially confirm H5.

In addition, the results of the feedback's relevance question showed that in both unsolicited and solicited feedback conditions, most of the participants found the feedback relevant (76% and 86% of participants, respectively). However, we did not find a significant difference between the perception of relevance in both experimental conditions ($\chi^2(1, N = 60) = 1.002, p = 0.317$).

4.3 Exploratory Findings

4.3.1 Enthusiasm and confidence. We decided to analyse the additional questions, connected to the social impression of audio-based CAs: enthusiasm and confidence. We found no statistically significant differences in the perception of CA enthusiasm (Unsol. - Base: $t.ratio = -0.808, p = 0.845$; Sol. - Base: $t.ratio = 0.323, p = 0.845$), which was expected because we used the same voice to generate the prompts. But we did find statistically significant differences in perceived confidence, as in both solicited and unsolicited conditions the CA was perceived as significantly less confident (Unsol. - Base: $t.ratio = -2.865, p = 0.011$; Sol. - Base: $t.ratio = -3.044, p = 0.010$).

4.3.2 Hesitation Lengths. We used hesitations as a proxy metric to assess the inception of the reflection process (i.e., a participant reconsidering their menu item selections). Specifically, for all experimental conditions, we examined the time gap between the CA's question 'Would you like to reconsider your choice' and the onset of user's response to this question. Hesitation lengths for all experimental conditions are presented in Figure 4. For the other experimental conditions (solicited and unsolicited feedback), the hesitation duration was measured right after feedback intervention (see Figure 1 for intervention patterns for each experimental condition), in milliseconds. The results indicated that, in the unsolicited feedback condition the time taken to before the onset of participant's response was significantly longer compared to the baseline condition (paired-sample $t(29) = -2.771, p = 0.010, 95\%CIs[-1{,}174.036; -176.96]$). We did not

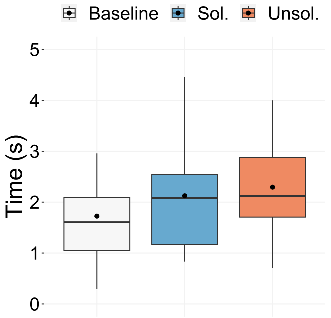


Fig. 4. Comparison of hesitation duration for all experimental conditions.

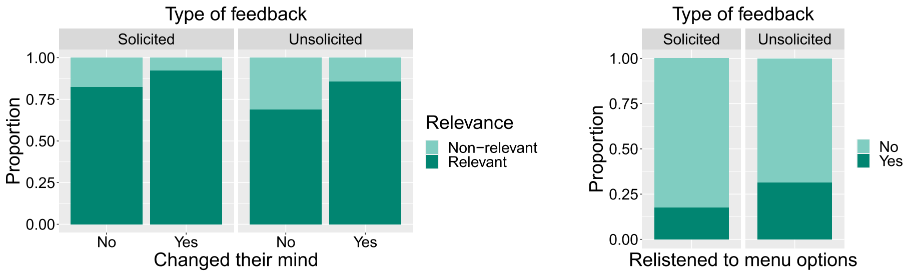


Fig. 5. Comparison of 'Food Genie' behavioural metrics.

observe significant differences in hesitation length between the baseline and solicited feedback conditions (paired-sample $t(29) = -1.616, p = 0.117, 95\%CIs[-1{,}347.856;\ 158.19]$). Statistical tests are Bonferroni-Holm corrected for two t-tests.

4.3.3 Relistening Behaviour. As for menu options selection, there was no difference in how frequently participants changed their mind in either of the feedback conditions ($\chi^2(1, N = 60) = 2.038, p = 0.153$). This result indicates that, while in the unsolicited condition, the CA made the participants curious to reconsider their menu choices, it did not translate to them changing their selections. Nonetheless, participants may have felt reassured with their selection after having relistened to the options. Figure 5 illustrates perceived relevance of feedback for both behavioural outcomes (change vs. no change) for both feedback conditions (left) and relistening to menu options (right).

4.4 Qualitative Results—Participants' Perceptions of Food Genie

We conducted semi-structured interviews after the experiment, where participants reflected on their interactions with Food Genie. We analysed the interviews via inductive coding, following the content analysis approach [91]. Given the straightforward questions and answers, we followed the practice recommended by McDonald [92] and had a single author thematically analyse the data. This resulted in responses categorised into the topics: (1) attitude towards the CA, including the user experience, and the perception of the voice and interaction and (2) attitude towards the CA's feedback, including the appropriateness and relevance of the feedback. As the study explored the social impression parameters of the interaction, we specifically asked participants to compare their experiences with Food Genie and a human waiter. This comparison aimed to assess how the fact that Food Genie is not human affected the overall perception of the interaction, particularly regarding the appropriateness of CA evaluations of human behaviour in a human-to-human scenario [35].

4.4.1 General Perceptions of the CA. All participants perceived Food Genie positively, explicitly praising the pleasant voice (26/30 participants) and the smooth interaction (23 participants). For instance, P7 said: 'The assistant sounds very natural. It is just like talking to another human being.' Similarly, P23 noted that they felt that Food Genie was realistic and empathetic: 'It felt like talking to a real human being. It gave me a sense of empathy as if it understood my needs.'

Several participants commented about the *sense of presence* created by the voice. Referring to their interaction experience with Food Genie, P23 said: 'It felt very comfortable. The voice was very pleasant. I felt a bit like being in the restaurant, like there is a man standing next to me. It was a bit like storytelling...' Similarly, P25 commented that talking to Food Genie felt like a highly realistic experience: 'You really feel like it is a person. Maybe in some conditions, it would be beneficial to make it sound more robotic so that you won't confuse it with the person. It is a little funny that I felt guilty because it sounded like a real person.' [Note: P25 felt guilty because they decided not to follow Food Genie's recommendation.]

Despite the positive reception, five participants nevertheless said that they prefer to order from humans. The main critiques regarding the interaction with the CA pointed to a *higher cognitive load* caused by the audio-only interaction (13), with eight participants suggesting that additional visuals would improve the experience. Other critiques expressed by the participants included their wish for more (11) and ideally personalised (11) options to select from, greater conversational capacities (12), and, foremost, a broader set of communicated information, including an order summary, nutritional details, pricing and delivery (24 participants).

4.4.2 Attitude towards the CA's Feedback. The extent to which participants found Food Genie's interventions appropriate varied based on their perceived reliability and relevance to the users. Many participants found the information trustworthy (20 participants), and they *valued it as beneficial for their wellbeing* (24). As a consequence, twenty expressed reconsidering their food choices after the feedback.

P8 praised the feedback that they received from Food Genie, comparing it to his experience with sports apps: 'I find it interesting because I rely on these kinds of things, so I know that they have a relatable source.' Similarly, P28 perceived the CA's intervention as benevolent: 'I really liked the extra information. Restaurants tend to use a lot of additives to the detriment of health, but the restaurants would not inform you about it. Food Genie was being open and honest about food so that I can make even better choices. It gave me a very transparent impression. It was very open and honest, which increased my trust.' The proactive feedback was also appreciated by P26, who said: 'It is a very new, creative and knowledgeable way to provide insights about my health. I find it important to know what kind of food is good for me and healthy for my body,' and P29, who compared Food Genie to his current CA, which does not offer any proactive support: 'I felt that feedback was really good. When I ask Siri it only answers the questions. Here the feedback is proactive because it is something that I may not have thought about.' Some participants (P1, P3 and P12) pointed out that personalisation of the feedback (i.e., CA accounting for user allergies and health issues) would make it even more beneficial (11). Eleven participants perceived the feedback as *patronising and unwelcome*, and some felt judged for their choices (8). P2 commented that 'it is slightly strange to receive suggestions from a robot. It was really surprising. When I heard the recommendation, I felt like: Who are you to be telling me that?' while P6 noted, 'It felt suspicious because I did not really know how the system works under the hood.' P13 said that proactive interventions go beyond what is traditionally expected of CAs: 'I don't like when machines are trying to pass for humans. I find this kind of technology creepy.' Similarly, P22 commented about the inappropriateness of the automated feedback: 'I did not find suggestions very helpful. I don't want to hear something that is insulting. This information would not make me change my selection. It would just make me think: Why are you saying bad things about food I am going to eat?'

4.4.3 Receiving Feedback from Human vs CA. Regarding the source of feedback, most participants preferred to receive it from a CA rather than a human (12). Only four would have preferred it from a human, and three explicitly expressed indifference. The participants who argued in favour of human feedback noted that it is faster: 'It is pretty cool to receive feedback from AI, but people can do it faster' (P3), more familiar: 'I am more accustomed to being judged by people' (P6), more reliable 'I rely more on people.' (P11) and less annoying 'It would have annoyed me less if the feedback came from a person' (P13). The preference for CA-based feedback was motivated by its unbiased character and the assumption that it comes from a *trustworthy and rich data source.* P18 pointed out that suggestions from CA can provide more informed insights: 'AI suggestions feel a bit awkward. It is not something that I am used to. However, I did not feel that it was misinformation or wrong advice. The advice was there for a reason, probably. I am assuming that CA recommendation is based on automated, informed choices such as tables with sugar levels.'

Several participants praised the *impersonal and non-judgemental character* of Food Genie, which makes it a preferable source of feedback. For instance, P4 said, 'It would be more offensive to hear this type of suggestion from a waiter. When it comes from a CA, it is just like advice.' P7 noted: 'If it were human, it would have been more judgemental, so it is better to receive suggestions from a robot.' Another participant (P28) commented that when the feedback comes from a person, it could be considered more biased and potentially spiteful: 'I don't think you would have experienced it in a restaurant. It would be more negative when a person said it. When CA says it, it comes across as intentional and better catered. For me, it is more positive to have it from a CA rather than a person. It would feel like the waiter had a bad day and wanted to take it on clients.'

4.4.4 Preferred Feedback Strategy. As we asked questions about the overall perception of Food Genie after three cycles of interaction, it was not possible to measure the separate effect of feedback type on the general impression parameters. However, when invited to consider their preference regarding how feedback was provided by the CA, of those who expressed their opinion, nine did not perceive a difference between the two strategies, and eleven favoured the solicited feedback. No participant preferred the unsolicited feedback, which six participants described as intrusive and impolite.

Participants in favour of solicited feedback noted that, 'It gives you more choice, it does not come across as so judgemental' (P6), 'Unsolicited feedback was highly inappropriate, it felt pretty invasive. 'The second option is high in cholesterol' (does the CA's voice). Like, who the f*** are you to judge me? (laughter) When the feedback was provided at the end, it was more acceptable.' (P12). P28 noted that, on balance, solicited feedback provides more agency: 'I like both ways, but generally speaking, it is better to be asked first. It is always nice to receive a choice. I know that some people are more sceptical about AI, but the more choice you have, the more at ease it will be for you. I personally would prefer to have this kind of feedback.'

Across the experiment, we saw that the *timing for the feedback intervention was critical.* Participants recognised when the timing was not ideal, e.g., 'In the third condition (solicited feedback), it was more inappropriate because I was not asked at the right moment. It sounded a bit like an afterthought. It is like you have already made your choice, and someone is trying to convince you to change it.' (P18). Similarly, P11 and P19 highlighted the importance of prompt feedback to enable a smooth interaction: 'For me, it is best to receive the feedback immediately.' (P11), 'It felt more relevant to get the feedback directly right after I submitted my option. I did not like the feedback but I feel that it was more appropriate coming right after my selection.' (P19)

Use Cases Beyond Food Ordering. Twenty participants reflected on use cases that could equally benefit from CA's feedback interventions. As such they proposed CAs as sports and health coaches (3 participants), shopping assistants (2), cooking and diet assistants (2), personal activity assistants for learning (2), and travel guides (3).

5 Discussion

Our study provides empirical evidence that a proactive, voice-only CA which provides users with unsolicited feedback on food selections is perceived as less appropriate and less confident than a no-feedback CA. However, both types of proactive feedback (i.e., solicited and unsolicited) can promote reflection by making users relisten and reconsider the options more frequently compared to the baseline (no feedback). This result indicates that, while the type of CA's feedback may not directly translate to participants changing their minds, being exposed to CA's feedback gives them a nudge to reconsider their selection. The results of our investigation provide a stepping stone towards design of proactive personalised CAs for food ordering scenarios that can slow down decision-making processes and consequently promote more considerate choices.

5.1 Perceptions of Food Genie

As noted in Section 3.2, the results of previous studies indicated openness of users to receiving proactive feedback from their CAs [84, 132] and a positive perception of CAs that provide solicited health-related suggestions [114, 137]. These results led us to hypothesise that a proactive CA which provides nutrition-related feedback would be perceived more trustworthy (*H1*) and persuasive (*H2*) as compared to a passive CA that does not provide any feedback. We further hypothesised that solicited feedback would be considered as more appropriate than unsolicited feedback (*H3a*), while the unsolicited feedback would be considered as less appropriate than no feedback (*H3b*).

5.1.1 Trustworthiness—H1. While there were no statistical differences in terms of trustworthiness, moderate effect sizes indicate that the baseline condition is more trustworthy than the solicited feedback condition (see Section 4.2). This result is quite surprising, as it goes against our assumption that a CA that asks for user's permission to provide feedback should have been perceived as more trustworthy. One plausible explanation for this result is that in the solicited condition, the CA's feedback was provided once all of the selections have been made, which potentially could have created an impression that the CA is hiding something from the participants by not providing the feedback immediately. As indicated by Edwards et al. [36], CA's spoken interruptions should be delivered sooner if the task is considered urgent. It could be argued that, in our study, the unsolicited feedback was perceived as more trustworthy and less of a hindrance towards the task completion, since the CA provided the required information 'just at the right time'.

5.1.2 Persuasiveness—H2. Both feedback conditions were perceived as significantly more persuasive than the baseline. This result was expected since in the feedback conditions the CA asks participants if they would like to reconsider their choice by providing arguments, while in the baseline condition there is no argument in favour of reconsideration (see Table 2). Overall, we registered high trust ratings for both feedback conditions. In some cases, participants were surprised by CA's recommendations but still followed its feedback. For example, P8 remarked that Food Genie sounded like a person which made her feel guilty not to follow its recommendations.

5.1.3 Appropriateness—H3a and H3b. We have not found support for our H3a, 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 condition, which is status quo. On the other hand, as hypothesised in H3b, the baseline strategy has been found significantly more appropriate than unsolicited feedback (see Figure 3). This result could be linked to scepticism regarding CAs making dietary suggestions reported by Luria et al. [84] and concerns regarding participants' agency presented by Reicherts et al. [114] and Zargham et al. [137], or the belief that a CA should not have or express its own views [30]. While participants of our study were more in favour of receiving feedback from a CA rather than a human (9 out of 13), some found CA's 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 CAs, which consequently yielded lower appropriateness scores.

5.2 Behavioural Analysis

Based on the results of previous studies on eliciting participants' expectations regarding proactive CAs [132, 137] and research on decision-making systems [15, 112], we hypothesised that a proactive CA (which asks for user's permission before providing feedback) will foster more reflection by: making them relisten their choices more frequently (*H4*) and consequently lead to more menu selection changes (*H5*) than a CA that provides feedback without permission and CA that does not provide any feedback.

5.2.1 Relistening Behaviour—H4. Contrary to our expectations, both unsolicited and solicited feedback conditions have led to significantly more frequent relistening behaviour compared to the baseline. It seems that the abruptness of the intervention was equally effective in making participants question their selection and consequently encouraged them to listen to the available menu options again. This result is related to findings reported by Graesser et al. [54] who noted that, despite their negative perception, system interruptions that are unexpected and confusing can have a positive contribution to users' learning experience by increasing engagement in the task. The longer hesitations before making the decision about the user's choice in unsolicited feedback conditions can also be interpreted as a sign of higher engagement in the task and deeper reflection about their choice.

5.2.2 Changing Choices—H5. We have not registered any statistically significant differences between both proactive feedback strategies when it comes to changing selected menu options. A plausible explanation is that in both conditions the CA provided transparent feedback about nutritional values of menu items without trying to manipulate the participants. This finding is similar to results of the study by Dubiel et al. [33] where persuasive perception of the agent did not translate to changes in participants' selections.

5.3 Takeaways and Reflections for Proactive Decision-making Support

Our study indicates that, while proactive feedback interventions may be considered as less trustworthy and less appropriate than the standard, no feedback interaction (offered by the majority of the current voice-only CAs) provides a way to impact the users' decision-making process. Additionally, our data showed that up to one third of participants did not perceive the difference between the types of feedback, while for the rest the preferences of feedback were distributed practically equally. This trend may indicate that all three types of feedback (i.e., baseline, solicited and unsolicited) can be perceived as useful by different groups of users. It is also possible that the preferred type of feedback can be connected with communication norms that apply in different cultures, and differences in main personality traits such as conscientiousness, agreeableness, neuroticism, openness to experience and extraversion (i.e., the Big Five Personality Traits [52]). However, the specific impact of these factors should be explored in future work.

As indicated by our experimental data (see Figure 3), participants had a general propensity to trust the automated feedback. Interestingly, even unsolicited feedback interventions did not seem to affect the level of trust and the belief in the benevolent and impartial nature of the CA as a source of credible feedback (possibly explaining the skewed distributions of plots in Figure 3(a)). It should be noted that this 'credit of trust' towards proactive CA intervention offers an opportunity to benefit the users by making their decisions more considerate, but it can also be used to manipulate users' decisions to their disadvantage.

Moreover, as mentioned by our participants during the semi-structured interviews, the feedback should provide a right balance between the timeliness of intervention and appropriateness. Solicited feedback that is provided too late could potentially be perceived as inappropriate and manipulative. On the one hand, unsolicited feedback may be preferable for the situations when we want to slow down the user's decision-making process and foster reflection (e.g., combating impulsive purchases). On the other hand, while theoretically more appropriate, solicited feedback can be more deceptive as, by giving the agent the opportunity to present the feedback, we potentially become more likely to get influenced without conscious control [1] by unwittingly giving consent to be exposed to persuasive attempts [20]. In our study, all participants decided to take the feedback when it was available, however, some participants felt 'guilty' and that is why they followed the recommendation as not to 'upset the agent'. One way of addressing this problem, as suggested by Dubiel et al. [31] would be to provide users with control over the desired level of CAs' proactivity and personalised support.

Going beyond the 'food ordering' domain, we believe that there are many more areas where proactive voice-based CAs can proliferate. One such area is health and well-being, where the agent can proactively challenge the user to become more active by making them reflect on their lifestyle. For example, this could be achieved by adding alternative exercise suggestions to self-tracking CAs (such as the ones discussed by Kocielnik et al. [72]). Proactive CAs can also be used in AI-supported decision-making. Specifically, building on research by Gajos and Mamykina [46], a CA can suggest alternative problem solutions to promote caution, reduce overreliance, and make users reflect on their own reasoning. Yet another potential application area is combating procrastination and addictive behaviours such as excessive use of smartphones. Recently, Li et al. [82] proposed a text-based chatbot, called 'StayFocused' to promote reflection on their smartphone usage. One possible extension of this system would be to incorporate proactive voice interventions during extensive periods of social media usage, to alert users regarding that they have spent on the particular platform and suggest switching to other, more productive activities.

While our study provides a use case for a proactive reflection CA, it should be noted that development and implementation of such systems will require overcoming several challenges. Firstly, the design of proactive CAs will require balancing between the goals of users and service providers which may be contradictory (e.g., healthy diet vs. more profit). Consequently, future research should consider tradeoffs between different stakeholders' goals in the design process and their implementation. Secondly, effective personalisation of proactive CAs in the food recommendations will be required. Previous research on recommender systems indicates that personalisation can lead to filter bubbles, preventing discovery of new food options and effectively lead to a less healthy and less varied diet, as compared to a generic system [128]. Another important consideration when implementing proactive voice assistance is the setting in which interaction takes place. Previous work shows that location (e.g., work vs. home) and social context (e.g., using CA when alone vs. in shared spaces) are crucial factors that determine social acceptability of CA interactions [32, 37].

5.4 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 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 persuasiveness agent may vary based on its gender [136]. Third, we did not gather participants' preferences to create personalised menu options. However, this decision was taken to limit privacy implications associated with creating personalised user profiles. Fourth, it should be noted that our experiment was limited to one-off interaction, therefore different results could have been observed over long-term CA usage, given that trust is a concept that is known to develop over time [61]. Fifth, it is possible that the difference in intervention points in proactive feedback conditions could have impacted participants perception of the CA in terms of trust and appropriateness, since the timing of feedback was different. However, as explained in Section 3.5, this design choice was taken to facilitate distinction of both strategies and make solicited feedback less obtrusive. Related to the fifth point above, we gathered the qualitative feedback from the users after all three ordering tasks have been completed, which could have impacted their ability to recall specific characteristics of Food Genie's behaviour in each individual condition. Finally, we would like to point out that, due to transient nature of speech, a graphical user interface would have provided participants with an ability to better scrutinise and reflect on provided menu options. However, since voice-only CAs currently allow users to order products and services, we decided to explore this modality in the current study.

On the note of generalisability, we acknowledge that since our study was set in the lab, some participants may have been affected by social desirability bias (i.e., the tendency to act in a way that is considered more socially desirable rather one that reflects their true intentions) [55]. Related to this point, participants' response times could have been affected by the 'pressure to perform'. Therefore, it is possible that the selected menu items could have been different if participants were ordering food at home. Finally, we would like to note that our findings should be considered in the context of a food ordering scenario and may not generalise beyond this domain.

5.5 Future Work

Future work should experiment with different domains in which users' autonomy can be compromised. In our protocol, the CA made interventions aligned with the users' values that might fall under the category of 'nudges'. As Noggle [101] explains, nudges that highlight information that is necessary for the user to make a decision are not manipulative, like in our case. However, it is necessary to test when proactive messages can foster reflection in scenarios where the intervention is not aligned with the users' values and with hidden manipulative techniques. In our study, CA's persuasive interventions provided information to users in a transparent way and without any manipulative intent.

Furthermore, to ensure ecological validity, future studies should be conducted in real interaction contexts and consider behavioural metrics over longer interaction periods. For example, when analysing longitudinal interactions with CAs, researchers can focus on the duration of interactions, user's response latency to proactive interventions and frequency of following vs. rejecting CA's recommendations. Subject to users' consent, such data could be analysed as interaction logs, or by following the ethnomethodological approach proposed by Porcheron et al. [107] where an external device is used to record user interactions that are then analysed using conversational analysis tools.

6 Conclusion

We have investigated how the proactive feedback interventions of a CA affects its perceived trustworthiness, persuasiveness and appropriateness and explored its impact on reflection and choices in a decision-making task. We found that unsolicited feedback strategy was perceived as less appropriate than the baseline condition (no feedback). Interestingly, while unsolicited feedback was generally perceived as less appropriate than solicited feedback, it led to equally more frequent relistening behaviour and effectively slowed down the decision-making process. Our investigation of the impact of providing feedback in support 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. Our Supplementary Materials, codebook, analysis scripts and interview questions are available at https://github.com/asyasergeeva/foodgenie.

Acknowledgement

We thank Eliott Bonte for his help with transcribing participants' dialogues.

References

[1] Martin Adam, Michael Wessel, and Alexander Benlian. 2021. AI-based chatbots in customer service and their effects on user compliance. *Electronic Markets* 31, 2 (2021), 427–445.

[2] Hugues A. Mehenni, Sofiya Kobylyanskaya, Ioana Vasilescu, and Laurence Devillers. 2021. Nudges with conversational agents and social robots: A first experiment with children at a primary school. In *Conversational Dialogue Systems for the Next Decade*. Luis Fernando D'Haro, Zoraida Callejas and Satoshi Nakamura (Eds.), Springer Singapore, 257–270.

[3] Sultan Almoallim and Corina Sas. 2022. Functionalities review of digital wellbeing apps: Towards research-informed design implications for interventions limiting smartphone use. *JMIR Form Res* 6 (2022), e31730.

[4] Tawfiq Ammari, Jofish Kaye, Janice Y Tsai, and Frank Bentley. 2019. Music, search, and IoT: How people (really) use voice assistants. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 3 (2019), 17–1.

[5] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H. De Vreese. 2020. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & Society* 35, 3 (2020), 611–623.

[6] Solomon E Asch and Henri Zukier. 1984. Thinking about persons. *Journal of Personality and Social Psychology* 46, 6 (1984), 1230.

[7] Simone Balloccu, Ehud Reiter, Matteo G. Collu, Federico Sanna, Manuela Sanguinetti, and Maurizio Atzori. 2021. Unaddressed challenges in persuasive dieting chatbots. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, 392–395.

[8] Albert Bandura. 1986. Social foundations of thought and action. *Englewood Cliffs, NJ* 1986, 23–28 (1986), 2.

[9] Sanchayan Banerjee and Peter John. 2024. Nudge plus: Incorporating reflection into behavioral public policy. *Behavioural Public Policy* 8, 1 (2024), 69–84.

[10] Eric P. S. Baumer, Vera Khovanskaya, Mark Matthews, Lindsay Reynolds, Victoria Schwanda Sosik, and Geri Gay. 2014. Reviewing reflection: on the use of reflection in interactive system design. In *Proceedings of the 2014 Conference on Designing Interactive Systems*, 93–102.

[11] Austin Beattie, Autumn P. Edwards, and Chad Edwards. 2020. A bot and a smile: Interpersonal impressions of chatbots and humans using emoji in computer-mediated communication. *Communication Studies* 71, 3 (2020), 409–427.

[12] Marit Bentvelzen, Paweł W. Woźniak, Pia S. F. Herbes, Evropi Stefanidi, and Jasmin Niess. 2022. Revisiting reflection in HCI: Four design resources for technologies that support reflection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–27.

[13] Kirsten Bergmann, Friederike Eyssel, and Stefan Kopp. 2012. A second chance to make a first impression? How appearance and nonverbal behavior affect perceived warmth and competence of virtual agents over time. In *Proceedings of the 12th International Conference on Intelligent Virtual Agents* (IVA '12). Springer, 126–138.

[14] Markus Blut, Cheng Wang, Nancy V. Wünderlich, and Christian Brock. 2021. Understanding anthropomorphism in service provision: A meta-analysis of physical robots, chatbots, and other AI. *Journal of the Academy of Marketing Science* 49 (2021), 632–658.

[15] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[16] Angelo Cafaro, Hannes Högni Vilhjálmsson, Timothy Bickmore, Dirk Heylen, Kamilla Rún Jóhannsdóttir, and Gunnar Steinn Valgarðsson. 2012. First impressions: Users' judgments of virtual agents' personality and interpersonal attitude in first encounters. In *Proceedings of the 12th International Conference on Intelligent Virtual Agents* (IVA '12). Springer, 67–80.

[17] Zoraida Callejas and David Griol. 2021. Conversational agents for mental health and wellbeing. In *Dialog Systems: A Perspective from Language, Logic and Computation.* T. Lopez-Soto (Ed.), Springer, 219–244.

[18] Yulia E Chentsova Dutton. 2012. Butting in vs. being a friend: Cultural differences and similarities in the evaluation of imposed social support. *The Journal of Social Psychology* 152, 4 (2012), 493–509.

[19] 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.

[20] Robert Cialdini. 2016. *Pre-Suasion: A Revolutionary Way to Influence and Persuade*. Simon and Schuster.

[21] 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 *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

[22] N. A. J. Cornelissen, R. J. M. van Eerdt, H. K. Schraffenberger, and Willem F. G. Haselager. 2022. Reflection machines: Increasing meaningful human control over decision support systems. *Ethics and Information Technology* 24, 2 (2022), 19.

[23] Anna L. Cox, Sandy J.J. Gould, Marta E. Cecchinato, Ioanna Iacovides, and Ian Renfree. 2016. Design frictions for mindful interactions: The case for microboundaries. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1389–1397.

[24] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies—Why and how. *Knowledge-Based Systems* 6, 4 (1993), 258–266.

[25] Kahneman Daniel. 2017. *Thinking, Fast and Slow*, Macmillan.

[26] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't just tell me, ask me: AI systems that intelligently frame explanations as questions improve human logical discernment accuracy over causal AI explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–13.

[27] Janna T. Deelstra, Maria C. W. Peeters, Wilmar B. Schaufeli, Wolfgang Stroebe, Fred R. H. Zijlstra, and Lorenz P. Van Doornen. 2003. Receiving instrumental support at work: When help is not welcome. *Journal of applied psychology* 88, 2 (2003), 324.

[28] Giada Di Stefano, Francesca Gino, Gary P. Pisano, Bradley Staats, and Giada Di-Stefano. 2014. *Learning by Thinking: How Reflection Aids Performance*. Harvard Business School, Boston, MA.

[29] 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 *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–11.

[30] Philip R. Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R. Cowan. 2019. Mapping perceptions of humanness in intelligent personal assistant interaction. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, 1–12.

[31] Mateusz Dubiel, Sylvain Daronnat, and Luis A. Leiva. 2022. Conversational agents trust calibration: A user-centred perspective to design. In *Proceedings of the 4th Conference on Conversational User Interfaces*, 1–6.

[32] Mateusz Dubiel, Martin Halvey, and Leif Azzopardi. 2018. A survey investigating usage of virtual personal assistants. arXiv:1807.04606. Retrieved from https://doi.org/10.48550/arXiv.1807.04606

[33] Mateusz Dubiel, Martin Halvey, Pilar O. Gallegos, and Simon King. 2020. Persuasive synthetic speech: Voice perception and user behaviour. In *Proceedings of the 2nd Conference on Conversational User Interfaces*, 1–9.

[34] Mateusz Dubiel, Anastasia Sergeeva, and Luis A. Leiva. 2024. Impact of voice fidelity on decision making: A potential dark pattern?. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, 181–194.

[35] Chad Edwards, Autumn Edwards, Fatima Albrehi, and Patric Spence. 2021. Interpersonal impressions of a social robot versus human in the context of performance evaluations. *Communication Education* 70, 2 (2021), 165–182.

[36] Justin Edwards, Christian Janssen, Sandy Gould, and Benjamin R. Cowan. 2021. Eliciting spoken interruptions to inform proactive speech agent design. In *Proceedings of the 3rd Conference on Conversational User Interfaces (CUI '21)*, 1–12.

[37] Christos Efthymiou and Martin Halvey. 2016. Evaluating the social acceptability of voice based smartwatch search. In *Information Retrieval Technology: 12th Asia Information Retrieval Societies Conference, AIRS 2016*, Beijing, China, November 30–December 2, 2016, Proceedings 12. Springer, 267–278.

[38] Ayoub E. Majjodi, Alain D. Starke, and Christoph Trattner. 2022. Nudging towards health? Examining the merits of nutrition labels and personalization in a recipe recommender system. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, 48–56.

[39] 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. arXiv.2103.14574. Retrieved from https://doi.org/10.48550/arXiv.2103.14574

[40] 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 (2013), 897–913.

[41] Michael Eraut*. 2004. Informal learning in the workplace. *Studies in Continuing Education* 26, 2 (2004), 247–273.

[42] Russell H. Fazio. 1990. A practical guide to the use of response latency in social psychological research. (1990). In *Research Methods in Personality and Social Psychology.* C. Hendrick & M. S. Clark (Eds.), Sage Publications, Inc., 74–97.

[43] Tom Feltwell, Gavin Wood, Phillip Brooker, Scarlett Rowland, Eric P. S. Baumer, Kiel Long, John Vines, Julie Barnett, and Shaun Lawson. 2020. Broadening exposure to socio-political opinions via a pushy smart home device. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

[44] Carlos Flavián, Miguel Guinalíu, and Raquel Gurrea. 2006. The role played by perceived usability, satisfaction and consumer trust on website loyalty. *Information & Management* 43, 1 (2006), 1–14.

[45] Brian J. Fogg. 2002. Persuasive technology: Using computers to change what we think and do. *Ubiquity* 2002 (December 2002), 2.

[46] Krzysztof Z, Gajos and Lena Mamykina. 2022. Do people engage cognitively with AI? Impact of AI assistance on incidental learning. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, 794–806.

[47] Anais Garrell, Michael Villamizar, Francesc Moreno-Noguer, and Alberto Sanfeliu. 2017. Teaching robot's proactive behavior using human assistance. *International Journal of Social Robotics* 9, 2 (2017), 231–249.

[48] Aimi Shazwani Ghazali, Jaap Ham, Emilia Barakova, and Panos Markopoulos. 2019. Assessing the effect of persuasive robots interactive social cues on users' psychological reactance, liking, trusting beliefs and compliance. *Advanced Robotics* 33, 7–8 (2019), 325–337.

[49] 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. In *Proceedings of the Advances in Neural Information Processing Systems,* Vol. 30, 1–9.

[50] Ella Glikson and Anita W. Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals* 14, 2 (2020), 627–660.

[51] Kate Goddard, Abdul Roudsari, and Jeremy C. Wyatt. 2011. Automation bias–A hidden issue for clinical decision support system use. *International Perspectives in Health Informatics* 164 (2011), 17–22.

[52] Lewis R. Goldberg. 2013. An alternative "description of personality": The big-five factor structure. In *Personality and Personality Disorders.* Routledge, 34–47.

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

[54] Arthur Graesser, Sidney D'Mello, Patrick Chipman, Brandon King, and Bethany McDaniel. 2007. Exploring relationships between affect and learning with AutoTutor. In *Proceedings of the International Conference on AIED*, 16–23.

[55] Pamela Grimm. 2010. Social desirability bias. *Wiley International Encyclopedia of Marketing*.

[56] 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 *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 390–400.

[57] Siddharth Gulati, Sonia Sousa, and David Lamas. 2019. Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology* 38, 10 (2019), 1004–1015.

[58] Yunha Han, Hwiyeon Kim, Hyeshin Chu, Joohee Kim, Hyunwook Lee, Seunghyeong Choe, Dooyoung Jung, Dongil Chung, Bum Chul Kwon, and Sungahn Ko. 2021. Wait, let's think about your purchase again: A study on interventions for supporting self-controlled online purchases. In *Proceedings of the Web Conference 2021*, 2476–2487.

[59] Tianzhi He, Farrokh Jazizadeh, and Laura Arpan. 2022. AI-powered virtual assistants nudging occupants for energy saving: proactive smart speakers for HVAC control. *Building Research & Information* 50, 4 (2022), 394–409.

[60] Iris Hendrickx, Federica Cena, Erkan Basar, Luigi Di Caro, Florian Kunneman, Elena Musi, Cataldo Musto, Amon Rapp, and Jelte van Waterschoot. 2021. Towards a new generation of personalized intelligent conversational agents. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, 373–374.

[61] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57, 3 (2015), 407–434.

[62] Guy Hoffman, Gurit E. Birnbaum, Keinan Vanunu, Omri Sass, and Harry T. Reis. 2014. Robot responsiveness to human disclosure affects social impression and appeal. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, 1–8.

[63] Kate S. Hone and Robert Graham. 2000. Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering* 6, 3–4 (2000), 287–303.

[64] Sung-Mook Hong and Salvatora Faedda. 1996. Refinement of the Hong psychological reactance scale. *Educational and Psychological Measurement* 56, 1 (1996), 173–182.

[65] Dietmar Jannach. 2022. Evaluating conversational recommender systems. *Artificial Intelligence Review* 56 (2022), 2365–2400.

[66] Yusuke Kato, Takayuki Kanda, and Hiroshi Ishiguro. 2015. May i help you?-design of human-like polite approaching behavior. In *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI '15).* IEEE, 35–42.

[67] Alice Kerly, Richard Ellis, and Susan Bull. 2007. CALMsystem: A conversational agent for learner modelling. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*. Springer, 89–102.

[68] Ahyeon Kim, Minha Cho, Jungyong Ahn, and Yongjun Sung. 2019. Effects of gender and relationship type on the response to artificial intelligence. *Cyberpsychology, Behavior, and Social Networking* 22, 4 (2019), 249–253.

[69] Everlyne Kimani, Kael Rowan, Daniel McDuff, Mary Czerwinski, and Gloria Mark. 2019. A conversational agent in support of productivity and wellbeing at work. In *Proceedings of the 8th international conference on affective computing and intelligent interaction (ACII '19).* IEEE, 1–7.

[70] Johannes Kizach. 2014. Analyzing Likert-scale data with mixed-effects linear models: A simulation study. *Poster Presented at Linguistic Evidence 2014, Tübingen, Germany*.

[71] Rafal Kocielnik, Daniel Avrahami, Jennifer Marlow, Di Lu, and Gary Hsieh. 2018. Designing for workplace reflection: a chat and voice-based conversational agent. In *Proceedings of the 2018 Designing Interactive Systems Conference*, 881–894.

[72] Rafal Kocielnik, Gary Hsieh, and Daniel Avrahami. 2018. Helping users reflect on their own health-related behaviors. In *Studies in Conversational UX Design.* R. Moore, M. Szymanski, R. Arar, and G. J. Ren (Eds.), 85–115.

[73] Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. 2018. Reflection companion: A conversational system for engaging users in reflection on physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–26.

[74] Matthias Kraus, Nicolas Wagner, and Wolfgang Minker. 2020. Effects of proactive dialogue strategies on human-computer trust. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, 107–116.

[75] Matthias Kraus, Nicolas Wagner, Nico Untereiner, and Wolfgang Minker. 2022. Including social expectations for trustworthy proactive human-robot dialogue. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, 23–33.

[76] 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* 14 (2020), 105.

[77] Dominika Kwasnicka, Stephan U. Dombrowski, Martin White, and Falko Sniehotta. 2016. Theoretical explanations for maintenance of behaviour change: a systematic review of behaviour theories. *Health Psychology Review* 10, 3 (2016), 277–296.

[78] Eun-Ju Lee. 2010. What triggers social responses to flattering computers? Experimental tests of anthropomorphism and mindlessness explanations. *Communication Research* 37, 2 (2010), 191–214.

[79] John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 1 (2004), 50–80.

[80] James R. Lewis. 2016. Standardized questionnaires for voice interaction design. *Voice Interaction Design* 1, 1 (2016), 1–16.

[81] Michael Lewis, Katia Sycara, and Phillip Walker. 2018. The role of trust in human-robot interaction. In *Foundations of Trusted Autonomy*. Springer, Cham, 135–159.

[82] Zhuoyang Li, Minhui Liang, Hai Trung Le, Ray Lc, and Yuhan Luo. 2023. Exploring design opportunities for reflective conversational agents to reduce compulsive smartphone use. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, 1–6.

[83] Ewa Luger and Abigail Sellen. 2016. " Like having a really bad PA" the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5286–5297.

[84] 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 *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12.

[85] 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. *Human Communication Research* 30, 1 (2004), 42–70.

[86] Erina.L MacGeorge, Lisa M. Guntzviller, Lisa K. Hanasono, and Bo Feng. 2016. Testing advice response theory in interactions with friends. *Communication Research* 43, 2 (2016), 211–231.

[87] Bertram F. Malle and Daniel Ullman. 2021. A multidimensional conception and measure of human-robot trust. In *Trust in Human-Robot Interaction.* Chang S. Nam and Joseph B. Lyons (Eds.), Elsevier, 3–25.

[88] Janne A. Martikainen, Erkki J. O. 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. *European journal of clinical nutrition* 65, 10 (2011), 1148–1155.

[89] Manolis Mavrikis, Beate Grawemeyer, Alice Hansen, and Sergio Gutierrez-Santos. 2014. Exploring the potential of speech recognition to support problem solving and reflection. In *Proceedings of the European Conference on Technology Enhanced Learning*. Springer, 263–276.

[90] Richard E. Mayer. 2014. Principles based on social cues in multimedia learning: Personalization, voice, image, and embodiment principles. *The Cambridge Handbook of Multimedia Learning* 16 (2014), 345–370.

[91] Philipp Mayring. 2004. Qualitative content analysis. *A Companion to Qualitative Research* 1, 2 (2004), 159–176.

[92] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[93] National Public Media. 2022. The Smart Audio Report — National Public Media — nationalpublicmedia.com. Retrieved January 09, 2023 from https://www.nationalpublicmedia.com/insights/reports/smart-audio-report/

[94] Thomas Mejtoft, Sarah Hale, and Ulrik Söderström. 2019. Design friction. In *Proceedings of the 31st European Conference on Cognitive Ergonomics*, 41–44.

[95] Tim Miller. 2023. Explainable AI is dead, long live explainable AI! Hypothesis-driven decision support using evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 333–342.

[96] Adam S. Miner, Arnold Milstein, Stephen Schueller, Roshini Hegde, Christina Mangurian, and Eleni Linos. 2016. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA Internal Medicine* 176, 5 (2016), 619–625.

[97] Anca M. Miron and Jack W. Brehm. 2006. Reactance theory-40 years later. *Zeitschrift für Sozialpsychologie* 37, 1 (2006), 9–18.

[98] John W. Mullennix, Steven E. Stern, Stephen J. Wilson, and Corrie-lynn Dyson. 2003. Social perception of male and female computer synthesized speech. *Computers in Human Behavior* 19, 4 (2003), 407–424.

[99] Cataldo Musto, Alain D. Starke, Christoph Trattner, Amon Rapp, and Giovanni Semeraro. 2021. Exploring the effects of natural language justifications in food recommender systems. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, 147–157.

[100] Ha Nguyen. 2022. Examining teenagers' perceptions of conversational agents in learning settings. In *Proceedings of the Interaction Design and Children*, 374–381.

[101] Robert Noggle. 2018. Manipulation, salience, and nudges. *Bioethics* 32, 3 (Mar 2018), 164–170.

[102] Brian A. Nosek, Carlee B. Hawkins, and Rebecca S. Frazier. 2011. Implicit social cognition: From measures to mechanisms. *Trends in Cognitive Sciences* 15, 4 (2011), 152–159.

[103] Thomas C. O'Guinn and Ronald J. Faber. 1989. Compulsive buying: A phenomenological exploration. *Journal of Consumer Research* 16, 2 (1989), 147–157.

[104] Jeesun Oh, Wooseok Kim, Sungbae Kim, Hyeonjeong Im, and Sangsu Lee. 2024. Better to ask than assume: Proactive Voice assistants' communication strategies that respect user agency in a smart home environment. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–17.

[105] Gustavo Penha, Eyal Krikon, Vanessa Murdock, and Sandeep Avula. 2022. Helping Voice Shoppers Make Purchase Decisions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems Extended Abstracts.* ACM, 1–8.

[106] Melanie D. Polkosky and James R. Lewis. 2003. Expanding the MOS: Development and psychometric evaluation of the MOS-R and MOS-X. *International Journal of Speech Technology* 6, 2 (2003), 161–182.

[107] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice interfaces in everyday life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12.

[108] Martin Porcheron, Minha Lee, Birthe Nesset, Frode Guribye, Margot van der Goot, Roger K. Moore, Ricardo Usbeck, Ana Paiva, Catherine Pelachaud, Elayne Ruane, Björn Schuller, Guy Laban, Dimosthenis Kontogiorgos, Matthias Kraus, and Asbjørn Følstad. 2022. Definition, conceptualisation and measurement of trust. *Dagstuhl Reports* 11, 8 (2022), 101–105.

[109] Lingyun Qiu and Izak Benbasat. 2009. Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems. *Journal of Management Information Systems* 25, 4 (2009), 145–182.

[110] Nicole Radziwill and Morgan Benton. 2017. Evaluating quality of chatbots and intelligent conversational agents. *Software Quality Professional* 19, 3 (2017), 25–36.

[111] 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 *Proceedings of the International Conference on Machine Learning*. PMLR, 8821–8831.

[112] Leon Reicherts and Yvonne Rogers. 2020. Do make me think! How CUIs can support cognitive processes. In *Proceedings of the 2nd Conference on Conversational User Interfaces*, 1–4.

[113] 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 Transactions on Computer-Human Interaction* 29, 3 (2022), 1–41.

[114] Leon Reicherts, Nima Zargham, Michael Bonfert, Yvonne Rogers, and Rainer Malaka. 2021. May i interrupt? Diverging opinions on proactive smart speakers. In *Proceedings of the 3rd Conference on Conversational User Interfaces (CUI '21)*, 1–10.

[115] Chong E. Rhee and Junho Choi. 2020. Effects of personalization and social role in voice shopping: An experimental study on product recommendation by a conversational voice agent. *Computers in Human Behavior* 109 (2020), 106359.

[116] 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.

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

[118] Holger Schielzeth, Niels J. Dingemanse, Shinichi Nakagawa, David F. Westneat, Hassen Allegue, Céline Teplitsky, Denis Réale, Ned A. Dochtermann, László Zsolt Garamszegi, and Yimen G. Araya-Ajoy. 2020. Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution* 11, 9 (2020), 1141–1152.

[119] Barry R. Schlenker and Mark R. Leary. 1982. Audiences' reactions to self-enhancing, self-denigrating, and accurate self-presentations. *Journal of Experimental Social Psychology* 18, 1 (1982), 89–104.

[120] Lijiang Shen and James P. Dillard. 2005. Psychometric properties of the Hong psychological reactance scale. *Journal of Personality Assessment* 85, 1 (2005), 74–81.

[121] Ben Shneiderman. 2020. Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction* 12, 3 (2020), 109–124.

[122] Jacqui Smith and Jacqueline J. Goodnow. 1999. Unasked-for support and unsolicited advice: Age and the quality of social experience. *Psychology and Aging* 14, 1 (1999), 108.

[123] Matthias Söllner, Axel Hoffmann, and Jan M. Leimeister. 2016. Why different trust relationships matter for information systems users. *European Journal of Information Systems* 25, 3 (2016), 274–287.

[124] Binghai Sun, Xiajun Yu, Xuhui Yuan, Changkang Sun, and Weijian Li. 2021. The effect of social perspective-taking on interpersonal trust under the cooperative and competitive contexts: The mediating role of benevolence. *Psychology Research and Behavior Management* 14 (2021), 817.

[125] Madiha Tabassum, Tomasz Kosiński, Alisa Frik, Nathan Malkin, Primal Wijesekera, Serge Egelman, and Heather Richter Lipford. 2019. Investigating users' preferences and expectations for always-listening voice assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–23.

[126] Vito Tassiello, Jack S. Tillotson, and Alexandra S. Rome. 2021. "Alexa, order me a pizza!": The mediating role of psychological power in the consumer–voice assistant interaction. *Psychology & Marketing* 38, 7 (2021), 1069–1080.

[127] Richard H. Thaler and Cass R. Sunstein. 2009. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Penguin.

[128] Christoph Trattner and David Elsweiler. 2017. Investigating the healthiness of internet-sourced recipes: Implications for meal planning and recommender systems. In *Proceedings of the 26th International Conference on World Wide Web*, 489–498.

[129] Iis Tussyadiah and Graham Miller. 2019. Nudged by a robot: Responses to agency and feedback. *Annals of Tourism Research* 78 (2019), 102752.

[130] Kathleen S. Verderber, Rudolph F. Verderber, and Cynthia Berryman-Fink. 2004. *Inter-Act: Interpersonal Communication Concepts, Skills, and Contexts*. Oxford University Press, New York.

[131] Carla Viegas and Malihe Alikhani. 2021. Towards designing enthusiastic AI agents. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, 203–205.

[132] Sarah T. Völkel, Daniel Buschek, Malin Eiband, Benjamin R Cowan, and Heinrich Hussmann. 2021. Eliciting and analysing users' envisioned dialogues with perfect voice assistants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15.

[133] Nicolas Wagner, Matthias Kraus, Niklas Rach, and Wolfgang Minker. 2021. How to address humans: System barge-in in multi-user HRI. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*. Erik Marchi, Sabato Marco Siniscalchi, Sandro Cumani, Valerio Mario Salerno and Haizhou Li (Eds.). Springer, 147–152.

[134] Jing Wei, Tilman Dingler, and Vassilis Kostakos. 2021. Understanding user perceptions of proactive smart speakers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–28.

[135] Sunghwan Yi and Hans Baumgartner. 2011. Coping with guilt and shame in the impulse buying context. *Journal of Economic Psychology* 32, 3 (2011), 458–467.

[136] Catherine Zanbaka, Paula Goolkasian, and Larry Hodges. 2006. Can a virtual cat persuade you? The role of gender and realism in speaker persuasiveness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1153–1162.

[137] Nima Zargham, Leon Reicherts, Michael Bonfert, Sarah T. Völkel, Johannes Schöning, Rainer Malaka, and Yvonne Rogers. 2022. Understanding circumstances for desirable proactive behaviour of voice assistants: The proactivity dilemma. In *Proceedings of the 4th Conference on Conversational User Interfaces* (CUI '22), 1–14.

[138] James M. Zech, Morgan Johnson, Michael D. Pullmann, Thomas D. Hull, Tim Althoff, Sean A. Munson, Nicole Fridling, Boris Litvin, Jerilyn Wu, and Patricia A. Areán. 2023. An integrative engagement model of digital psychotherapy: Exploratory focus group findings. *JMIR Formative Research* 7 (2023), e41428.

Received 12 July 2023; revised 28 May 2024; accepted 26 June 2024