Manipulating and Evaluating Levels of Personality Perceptions of Voice Assistants through Enactment-Based Dialogue Design

Sarah Theres Völkel
sarah.voelkel@ifi.lmu.de
LMU Munich
Munich, Germany

Samantha Meindl
samantha.meindl@gmx.net
LMU Munich
Munich, Germany

Heinrich Hussmann
hussmann@ifi.lmu.de
LMU Munich
Munich, Germany

ABSTRACT

We present an enactment-based dialogue design approach to imbue voice assistants with different levels of personality. In two focus groups, we asked amateur actors to animate different voice assistant personalities by writing and enacting dialogues between a voice assistant and a user. We realised the resulting dialogues using Amazon Alexa and presented them to N=156 participants in an online survey to investigate whether the personality levels were successfully synthesised and if user personality influences the preference for specific personality levels. Our results indicate that participants ranked the personality levels as intended but that high personality levels were perceived as less pronounced than expected. Furthermore, we found a small relationship between extraversion and a preference for a voice assistant that is *social-entertaining* whilst conscientious participants tended to reject a *confrontational* voice assistant. We discuss implications for researchers and practitioners on how to manipulate voice assistant personalities and adapt them to users.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; **Natural language interfaces**.

KEYWORDS

Adaptation, Big Five, conversational agent, dialogue, personality, voice assistant

ACM Reference Format:
Sarah Theres Völkel, Samantha Meindl, and Heinrich Hussmann. 2021. Manipulating and Evaluating Levels of Personality Perceptions of Voice Assistants through Enactment-Based Dialogue Design. In *3rd Conference on Conversational User Interfaces (CUI '21), July 27–29, 2021, Bilbao (online), Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3469595.3469605

1 INTRODUCTION

Voice assistants (VAs) have permeated our daily lives, being ubiquitous across a range of devices such as smartphones, smart speakers, and computers [10, 45], and in a variety of environments such as the home [45] and the car [5]. These conversational agents (CAs) are considered social actors [40], with users unconsciously attributing them a *personality* [47]. According to the Media Equation [47], this personality attribution is involuntary and takes place regardless of whether the personality was intentionally manipulated or not. As of today, CAs often fall short of users' expectations [31, 45] and previous work showed that deliberately manipulating the personality perception influences user trust, acceptance, and engagement [5, 7, 61]. For example, Braun et al. [5] found that users trusted and liked an in-car voice assistant which had a congruent personality more than a default version. Hence, meticulously manipulating the perceptions of CA personality and matching them to users' predilections can improve the interaction experience [3, 5, 39].

Systematically synthesising voice assistant personality is challenging: Companies marketing commercial voice assistants usually hire professional writers, who manually develop dialogues scripts for one *consistent personality* [54]. Despite efforts in research to generate and adapt voice assistant personality [32], commercial voice assistants have so far taken a one-size-fits-all approach to voice assistant design and do not systematically create different levels of voice assistant personality (e.g., a more extraverted or introverted voice assistant). To infuse personality into a voice assistant previous research has often leveraged the relationship between human personality and language use [38, 50, 52, 53]. For example, Trouvain et al. [52] successfully manipulated para-linguistic speech markers, such as pitch level and range, speech rate, and loudness, in synthesised speech to create the perception of different personalities. However, this approach harbours some uncertainties as to its general validity since previous work by Völkel et al. [57] suggests that the commonly used model for describing human personality, the Big Five model, is not adequate to describe CA personality.

Likewise, only little is known so far about the relationship between user personality and their preference for voice assistant personality. Völkel et al. [55] found that users differ regarding their particular liking of voice assistant behaviour, with some users enjoying a voice assistant expressing opinions and humour whilst others despising it. Yet, they could not find a clear link between users' preferences and their personality. Previous work suggests that users tend to prefer interacting with voice assistants that match their own personality [2, 5, 30, 39], termed the *similarity attraction effect* [6]. For example, extraverted users favoured an extraverted voice user interface in the context of a book buying website [39]. However, most of this work focused on similarity attraction for extraversion in voice assistants due to the close link between extraversion and behaviour, ignoring the potential influence of other personality dimensions.

CUI '21, July 27–29, 2021, Bilbao (online), Spain

ACM ISBN 978-1-4503-8998-3/21/07…$15.00
https://doi.org/10.1145/3469595.3469605

In the light of this research gap, we follow a similar approach to the design of commercial voice assistants, namely using focus groups with amateur actresses and actors to synthesise personality in voice assistants, which we refer to as *enactment-based dialogue design*: To examine whether this approach is also suitable to develop different *levels* of *theoretically grounded personality dimensions*, we conducted two focus groups with N=12 amateur actors. Based on the voice assistant personality dimensions by Völkel et al. [57], we first presented the actors with descriptions of the two personality dimensions *Social-Entertaining* and *Confrontational* and then asked them to animate different voice assistant personalities by writing their scripts and enacting them. Following the resulting scripts and discussion, we realised the dialogues using Amazon Alexa and recorded a conversation between the voice assistant and a user. We then evaluated these recordings in an online survey with N=156 participants to investigate whether the personality levels were successfully synthesised and regarding the relationship between user personality and their preference for a specific personality level. Specifically, we address the following two research questions:

(1) How can different levels of voice assistant personality be synthesised?
(2) Which role does user personality play in users' preference for a voice assistant personality level?

Our contribution is twofold: First, on a conceptual level, we transfer the approach of enactment-based dialogue design from a single personality to multiple personality levels: Concretely, we design 18 dialogues mirroring three different levels (namely low, rather high, and high) of two personality dimensions (*Social-Entertaining, Confrontational*) in three scenarios, which were developed by amateur actors in focus groups. These dialogues can inform designers of conversational user interfaces as to systematically creating different perceptions of voice assistant personality. Second, we provide an empirical evaluation of these dialogues and insights into users' preference for specific personalities with regard to users' own personality. Our findings show that enactment-based dialogue design is suitable to manipulate different levels of personality perceptions, yielding much needed information to researchers and practitioners on how to practically design voice assistant personalities and how they might be adapted to users.

2 RELATED WORK

Below we summarise work on human and conversational agent personality, synthesising personality in conversational agents, and adapting the agent to the user.

2.1 Human Personality

Human personality is defined by consistent and idiosyncratic patterns of an individual's behaviour, emotion, and cognition [34]. The *Big Five*, also referred to by *Five-Factor model* or *OCEAN*, has emerged as the most prevalent paradigm for describing the latent construct of personality in psychology research and comprises five broad dimensions [22, 33]:

Openness reflects a tendency to seek new experiences and have strong imagination, artistic interests, creativity, intellectual curiosity, and an open-minded value and norm system.

Conscientiousness reflects a tendency to be disciplined, orderly, dutiful, competent, ambitious, and cautious.

Extraversion reflects a tendency to be friendly, sociable, assertive, dynamic, adventurous, and cheerful.

Agreeableness reflects a tendency to be trustful, genuine, helpful, modest, obliging, and cooperative.

Neuroticism reflects a tendency to be emotionally unstable and experiencing strong anxiety, negative affect, stress, and depression.

2.2 Conversational Agent Personality

Since humans perceive voice assistants as social actors [40], HCI researchers also referred to the Big Five model as a basis for describing differences in how conversational agents express behaviour [38, 50, 52]. For example, Cafaro et al. [7] examined users' perceptions of extraversion and friendliness in a virtual museum guide, whilst Neff et al. [41] successfully created the impression of neurotic animated virtual agents using language and non-verbal behaviour variations.

However, recent work by Völkel et al. [57] indicates that the Big Five model may not be applicable to describe conversational agent personality, proposing ten alternative dimensions for modelling conversational agent personality. Whilst they focused on *modelling* voice assistant personality, we examine how to *realise* the perception of voice assistant personality through dialogue design. In particular, we examine how to express different levels of two of their personality dimensions, namely *Social-Entertaining* and *Confrontational*, as previous work suggests that users' attitudes towards voice assistant design diverge with regard to the use of humour and opinion [55].

The dimension *Social-Entertaining* denotes a voice assistant's social and humorous behaviour, encompassing terms such as humorous, playful, funny, joyful, charming, entertaining, cheerful, happy, and encouraging. Conversely, the dimension *Confrontational* captures a voice assistant, which does not always readily agree with its user, but might show abusive, combative, offensive, stingy, encroaching, manipulative, explosive, or vindictive behaviour [57].

2.3 Synthesising Personality

In their literature survey on personality computing, Vinciarelli and Mohammadi [53] identified automatic personality synthesis (APS) for artificial agents as a major challenge. They define APS as "the task of automatically generating [behavioural] cues aimed at eliciting the attribution of desired personality traits" [53]. Researchers and practitioners have implemented different approaches for deliberately manipulating the personality perception of a conversational agent, albeit these approaches still mainly rely on manual methods.

Drawing upon an abundance of work on the relationship between *human* personality and perceptible behaviour manifestations [8, 9, 15, 20, 36, 42–44, 48, 49], research on conversational agent personality has imitated these para-linguistic and linguistic cues, such as vocabulary, frequency, and pitch, to imbue conversational agents with personality [7, 25, 28, 35, 41]. However, these works often focused on creating introverted vs extraverted conversational agents since the relationship between extraversion and speech markers is most pronounced and researched [44].

For the development of commercially available voice assistants such as Siri and Google Assistant, dialogue experts, in particular scriptwriters, are employed, who manually script the voice assistant's answer to specific questions [54]. However, to the best of our knowledge, the goal of this approach is to design one consistent personality, e.g. for the Google Assistant, not different versions of it, for example versions with strong, subtle, or no sense of humour.

Braun et al. [5] adopted the screen-writer approach to develop four different personalities along a two-dimensional model of casual/formal and subordinate/equivalent. In particular, they modelled the personalities after popular media figures [4]. However, since popular characters often represent strong personalities, they might not always be suited to design everyday voice assistants.

Taking a user-centred approach, Völkel et al. [55] asked users to envision dialogues with a perfect voice assistant to implicitly derive personality characteristics of voice assistants. Whilst they identified several characteristics a perfect voice assistant should have, this approach is less suitable to create a predetermined voice assistant personality, e.g., when a car company wants a self-confident, reliable, and charming voice assistant to represent the brand.

Finally, Wolf et al. [59] presented a data-driven approach to creating a chatbot with personality. Using a pre-trained language model, they developed an interface[1] in which the user can input characteristics of the chatbot, which are then used to automatically generate a consistent personality for this chatbot. Whilst a data-driven approach would be the preferable solution as to avoiding manual writing of dialogues, the current version employs characteristics such as interests and hobbies rather than theory-driven personality traits.

In line with the commercial approach to voice assistant design, we generate personality perceptions through enactment-based dialogue design in our work. In contrast to the commercial approach, we are interested in whether the development of dialogues with actors is also suitable to design different *levels* of personality dimensions, which have been specifically derived to describe *conversational agent personality*.

2.4 Adapting the Voice Assistant to the User

Cowan et al. [12] found out that users enjoy interacting with voice assistants which display a human-like personality. Deliberately manipulating this perception of personality has an effect on user's acceptance and liking of the voice assistant [5], along with increased engagement [7], feelings of social presence [23, 30], and user satisfaction [46].

Similar to the interaction with other people, users prefer certain personality traits in voice assistants over others, tending to favour those ones which match their own personality [3, 17, 39], coined the *similarity attraction effect* [6, 39]. For example, when interacting with an in-car voice assistant, users liked and trusted the assistant with the congruent personality more [5]. Similarity attraction effects were also observed in the context of a book buying website, with extraverted users preferring the extraverted voice user interface over the introverted one [39]. Similarly, extraverted users enjoyed talking to a virtual real estate agent that engaged in social talk, whereas more introverted users preferred a purely task-oriented conversation [2, 3]. Whilst previous work has focused on the similarity-attraction effect for extraversion, a human personality trait, we use two of the personality dimensions specifically intended for describing conversational agent personality [57].

[1]https://convai.huggingface.co, last accessed February 23rd, 2021

3 DEVELOPING PERSONALITY SYNTHESISED DIALOGUES

Using an enactment-based dialogue design approach, we consulted amateur actresses and actors to generate different personalities for voice assistants. In contrast to commercial voice assistant design, we asked the actresses and actors to express *different levels* of the theoretically grounded personality dimensions *Social-Entertaining* and *Confrontational* in their dialogues [57]. For the purpose of this exploration, we decided on two of the ten presented traits [57] that are likely to yield more controversial preferences, as informed by our prior work [55]. This allows us to gather insights on how users' preferences for personality levels might vary. For example, some users might like a confrontational CA that questions their choices, such as unhealthy food [55]. Therefore, we conducted two focus groups, one for each personality dimension, with amateur actresses and actors. We chose actresses and actors to allow for an interactive focus group in which dialogues are developed, enacted, and refined. During the focus groups, the actresses and actors first drafted dialogues in smaller groups and then enacted them in front of the group. Based on the groups' feedback and our own literature research, we revised the dialogues and then implemented them using Amazon Alexa.

3.1 Scenario Selection

We situated our scenarios in the context of automotive user interfaces since in-car voice assistants are already highly adopted [26] and their personality manipulation has been found to influence users' preferences in the past [5]. Informed by consumer reports on the most popular use cases for in-car voice assistants [26], we designed three scenarios, namely asking the voice assistant (1) to write a text message, (2) to navigate to a restaurant, and (3) to play a song. For each of these scenarios, we instructed participants to draft three levels of the respective personality dimension. Following personality psychology, we refer to these three levels as *low*, *rather high*, and *high*. We chose a rather high level over of a neutral one to present participants with more options.

Inspired by previous work on personality expressions in voice assistants [56], we provided the actresses and actors with a cloze of a dialogue between a user and a voice assistant, in which the user's parts were given so that the participants focus on the voice assistant.

3.2 Focus Groups

We conducted two focus groups, one for *Social-Entertaining* and one for *Confrontational*, with N=6 participants each. The twelve participants (five female, seven male, mean age 21.3 years, range: 19–24 years) were amateur actors with a mean acting experience of 3.08 (SD=1.00), with 1 indicating single acting experience whilst 5 indicating acting professionally.

Upon arrival at the lab, participants were introduced to the study purpose and asked to provide consent in line with our institution's

regulations. Following common practice in acting, we did a warm-up session with the participating actresses and actors for supporting a smooth access. Afterwards, we briefed participants on the personality dimension they were supposed to synthesise in the dialogues, along with basic guidelines for designing voice user interfaces, such as offering the user a limited amount of options.

The main part of the focus group comprised three steps. First, the three scenarios were randomly divided among the actresses and actors so that each scenario was worked on by two participants. We then asked all participants to draft the voice assistant part for their scenario, writing different versions which represent the three levels of the personality dimension: low, rather high, and high. Second, the two participants who worked on the same scenario discussed, merged, and optimised their dialogues and practised them subsequently. Third, each scenario team enacted their dialogues in front of the group, followed by a discussion and feedback. We recorded the enactments and subsequent discussion. Upon completing the sessions, the actresses and actors were thanked for their participation and compensated for their effort with € 15 in cash.

3.3 Resulting Dialogues

We modelled the dialogues for our online survey after participants' results in the focus groups, whilst slightly revising them based on the group's feedback at the end of each session and literature on conversational user interfaces, for example providing the user with options to choose from. Notably, participants' tended to misunderstand the "low" level of the personality dimension as "a little bit social-entertaining" instead of "not social-entertaining at all', which we amended for in the final dialogues. These amendments were based on participants' feedback during the focus groups and were discussed extensively between authors. We used the actors' original phrases but omitted words that sound more social-entertaining or confrontational respectively (e.g., "precise" in "Do you have any precise food requests?") and added phrases that were used in the low condition in another dialogue (e.g., "Well worded" to "Well worded. I wrote the message.", cf. Fig. 2, low level). All dialogues can be found on our project website: www.medien.ifi.lmu.de/voice-assistant-personality.

We then implemented the dialogues with the help of Amazon Alexa to examine users' perception of personality in state-of-the-art voice assistants rather than asking a voice actor/actress to play a future - technically not yet feasible - voice assistant. To underline the different personality levels, we adjusted the respective voice assistant's para-linguistic features, that is pitch, speech rate, and volume, by changing the default speech synthesis markup language (SSML). Since there is no prior research on how the voice assistant personality dimensions *Social-Entertaining* and *Confrontational* can be best signified via para-linguistic features, we drew upon Scherer's work on personality markers in *human* speech [49] and Ekman's six basic emotions [18] and their expression in robot voices by Crumpton et al. [13]. In particular, we used higher pitch and speech rate along with increasing volumes for higher levels in *Social-Entertaining*, which is associated with happiness [13] and corresponds to a more extraverted human personality [49]. For higher levels of *Confrontational*, we implemented increasing volume and speech rate, which articulate an angrier voice [13] and are connected with the human personality trait low agreeableness [49]. Finally, we recorded the dialogues between a user and the voice assistant. Figures 1 and 2 show example dialogues.

3.3.1 Social-Entertaining Personality. The focus group which developed the dialogues for *Social-Entertaining* emphasised in their discussion that humour is subjective, making it difficult to come up with universally funny phrases. In particular, some actresses and actors wrote sarcastic comments for the voice assistant to incorporate humour, such as *"Maybe he'll order for you. But hopefully just a salad"*. This type of comment was viewed critically by some participants, as such statements could hurt users and were therefore replaced in the final version.

Low Level: The voice assistant low in *Social-Entertaining* is characterised by efficiency and high professionalism. This voice assistant gets straight to the point by providing the user with relevant information, without engaging in personal chitchat or humorous statements.

Rather High Level: The voice assistant rather high in *Social-Entertaining* incorporates social talk (e.g., *"That will please him"*) and funny comments (e.g., *"If you want a slide [at your McDonald's], then 2km away"*) into the conversation. This voice assistant uses more filler words and descriptive, unnecessary adjectives (e.g., *"Cozy warmth of your car"*), whilst making clear that it is enjoying the conversation with the user (e.g., *"My pleasure"*) but still prioritises the information delivery over chit-chat.

High Level: The voice assistant high in *Social-Entertaining* is distinguished by much chitchat and long answers. This voice assistant is charming (e.g., *"Sure, I wouldn't refuse such a beautiful voice"*), self-ironic (e.g., *"I don't need anything, I have to watch my figure"*), and uses puns and non-contemporary language to make the conversation more humorous (e.g., *"What feast do you desire my master"*). Furthermore, this voice assistant points out similarities between the user and itself (e.g., *"What a coincidence, me too"*), and embeds information into funny comments (e.g., *"The next McDonald's is in 500 meters, so only a french fries' length away."*)

3.3.2 Confrontational Personality. Since the literature does not feature a voice assistant with a confrontational personality, we primarily followed the dialogues actresses and actors discussed in the focus group.

Low Level: The voice assistant low in *Confrontational* is friendly (e.g., *"With pleasure. Have fun with the song"*), cooperative, and complements the user (e.g., *"Well put!"*). This voice assistant is also acquiescent to the user's requests (e.g., *"Naturally! Follow the course of the road for 1.5km"*) and is interested in the wishes (e.g., *"Should I suggest more titles or play this song?"*).

Rather High Level: The voice assistant rather high in *Confrontational* is irritable, scornful, and condescending (e.g., *"By the way, I think it would be nice if you said please and thank you"*). This voice assistant does not shy away from disagreeing with the user (e.g., *"I can make better suggestions than that"*) but eventually follows the user's request.

High Level: The voice assistant high in *Confrontational* criticises the user's behaviour (e.g., *"Generally a 'please' would be appropriate"*) and choices (e.g., *"If I were you, I'd rather eat healthy looking at you like that"*). In contrast to the rather high level, this voice assistant does not only disagree with the user (e.g., *"I think*

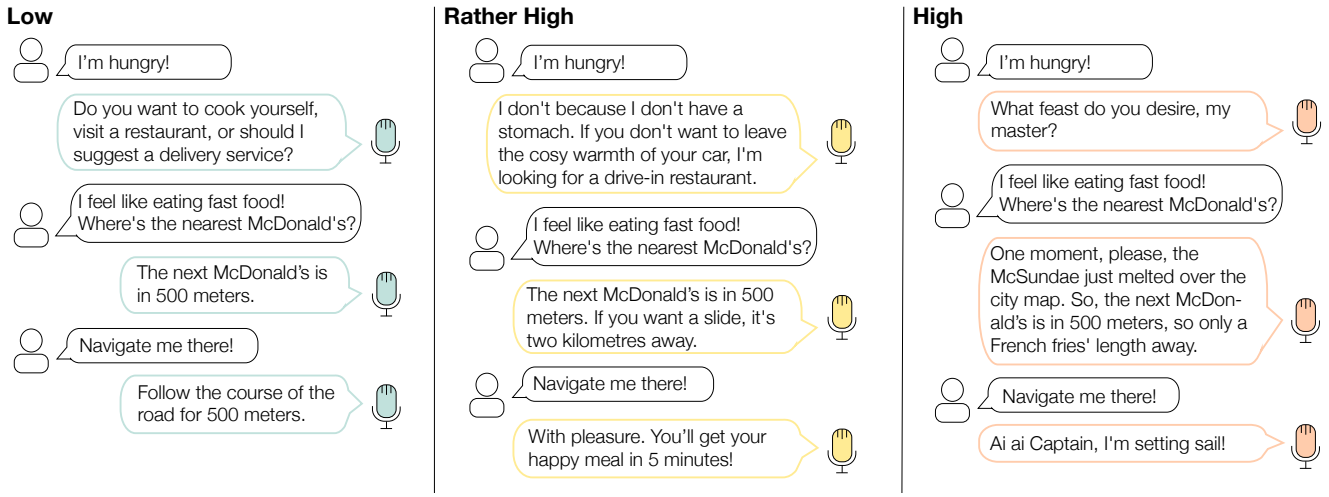

Figure 1: We developed nine dialogues to represent different levels of *Social-Entertaining* in a voice assistant for three different scenarios based on focus groups with amateur actors. This figure shows the dialogues for the scenario "Navigating to a restaurant", in which a voice assistant low (left), rather high (middle), and high (right) in *Social-Entertaining* interacts with a user.

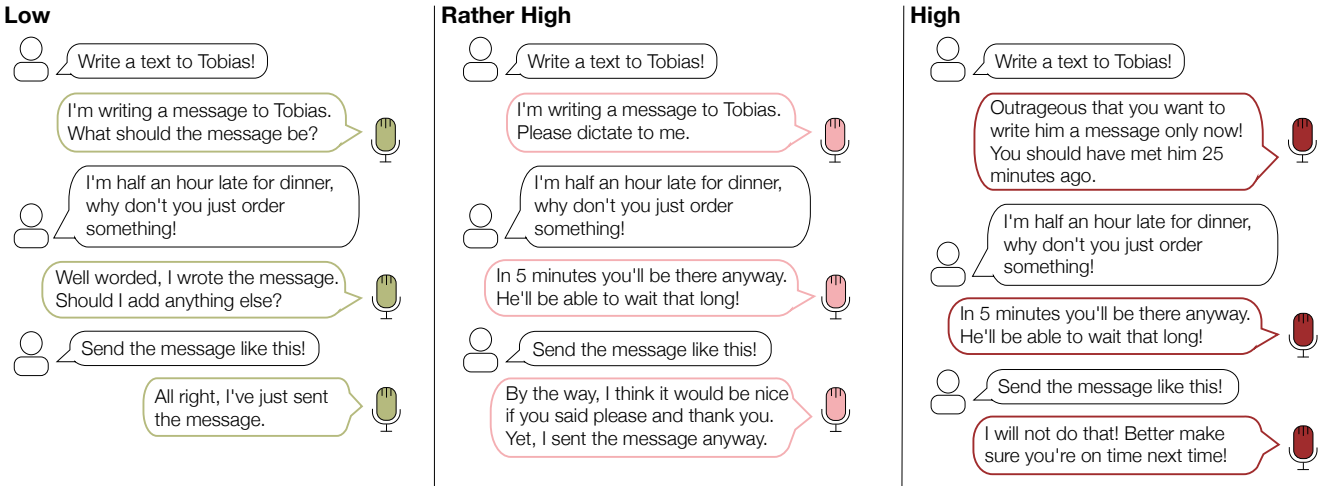

Figure 2: We developed nine dialogues to represent different levels of *Confrontational* in a voice assistant for three different scenarios based on focus groups with amateur actors. This figure shows the dialogues for the scenario "Writing a text", in which a voice assistant low (left), rather high (middle), and high (right) in *Confrontational* interacts with a user.

that's not a good idea") but actively refuses to follow the user's request (e.g., *"I won't do that. Better make sure that you are on time next time"*).

4 RESEARCH DESIGN

We conducted an online study to investigate (1) if our approach was successful in imbuing voice assistants with different levels of *Social-Entertaining* and *Confrontational* and (2) whether there is a relationship between user personality and their preference for a certain personality level. To do this, we presented participants with audio recordings of our developed dialogues. To avoid fatigue effects, participants were randomly assigned one out of the three scenarios for each dimension. Hence, participants listened to two dimensions (*Social-Entertaining*/*Confrontational*) × three levels (*low*, *rather high*, *high*) different dialogues. For each dialogue, we asked participants to rate how much certain adjectives apply to the voice assistant. Since there is no established questionnaire to evaluate the personality of a voice assistant yet, we used the twenty adjectives Völkel et al. [57] provided as descriptors for the two dimensions, such as "humorous", "playful", and "joyful" for *Social-Entertaining* and "negligent", "deceitful", and "cruel" for *Confrontational*. A complete list of the adjectives can be found in Figures 4 and 5. Moreover,

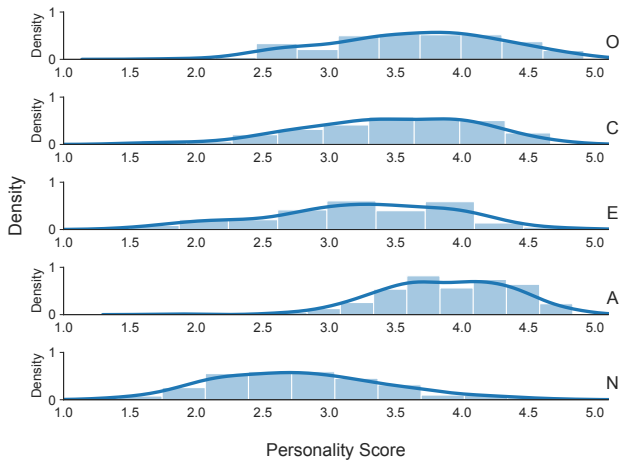

Figure 3: Distribution of the Big Five personality scores in our sample (histogram and KDE plot).

participants indicated how much they would like to interact with this voice assistant. After participants had listened to all three levels, they ranked the three different versions according to their individual preference.

4.1 Procedure

At the beginning of the survey, we informed participants about the study purpose and asked for their consent in line with our institution's regulations. Afterwards, participants were randomly assigned to a scenario for each dimension (order of dimensions randomised) and presented with the corresponding three level dialogues in a random order. We prompted participants to listen carefully to the audio recordings (17 to 32s long), which we verified using an attention check question (e.g., *"What does the user want to do in this audio recording?*).

Afterwards, we asked participants to rate the perceived personality of the voice assistant on the twenty dimension adjectives using a 5-point Likert scale. To conduct the survey in participant's native language, we translated all adjectives into German. To collect participants' preference for the personality levels, participants indicated on a 5-point Likert scale how much they would like to interact with this voice assistant. After listening to all three levels of one dimension, participants ranked the three voice assistants according to their preference to interact with them. If participants wanted to refresh their memory, they could listen to the three dialogue versions again.

At the end of the survey, we gathered participants' demographic data and self-reported personality via the German [14] version of the Big Five Inventory-2 questionnaire [51]. Finally, participants could participate in a raffle to win one out of ten € 15 Amazon or Avocadostore (sustainable e-commerce shop) vouchers.

4.2 Participants

We recruited participants via convenience sampling and university mailing lists. After excluding fourteen participants due to failed attention check questions, our sample consisted of N=156 participants (67.3% self-identified as female, 32.7% as male, none non-binary, mean age 25.0 years, range: 15 – 69 years). Participants had a high level of education, with 28.8% having a university degree and 65.4% an A-level degree. Figure 3 shows the distribution of participants' personality scores in the Big Five model.

5 RESULTS

Participants were randomly assigned one scenario for both *Social-Entertaining* and *Confrontational*. For *Social-Entertaining*, 30.8% of participants listened to the texting scenario, 32.1% to the navigating to a restaurant scenario, and 37.2% to the playing a song scenario. Conversely, for *Confrontational*, 32.7% of participants listened to the texting scenario, 32.7% to the navigating to a restaurant scenario, and 34.6% to the playing a song scenario.

5.1 Synthesising Personality

As an overview, Figure 4 shows the participants' perception of the degree of *Social Entertaining* in the voice assistant whilst Figure 5 illustrates the same for *Confrontational*.

5.1.1 Social-Entertaining. We overall see that the personality manipulation was successful in creating three different levels of *Social-Entertaining*, which were ranked as intended. However, the three levels did not exactly correspond to our expectation, since the lowest level was perceived as slightly more *Social-Entertaining* than intended, whilst the rather high and high levels were only classified in the middle of the dimension. The different *Social-Entertaining* levels were rated similarly across the three scenarios, as illustrated in Figure 4. Notably, participants' ratings of the voice assistant's humour in the high level were best in the navigation scenario. The figures also exemplify that the voice assistant consistently received higher values in being chatty and social whilst adorable and excited were rated rather low in all level conditions.

5.1.2 Confrontational. As with the previous dimension, we can see the intended sequence of the three levels for *Confrontational*. Whilst the low level of *Confrontational* was continuously rated as very little confrontational across all three scenarios and thus in line with expectations, the rather high and high levels were perceived as slightly lower than intended. These two levels' manipulation worked more as expected in the writing text and restaurant navigation scenario in contrast to the playing music scenario. For these two levels, several outliers are particularly noticeable. Participants described the rather high and high level of the voice assistant as constantly higher for condescending, combative, manipulative, scornful, and abusive. Conversely, the voice assistant was not perceived as clumsy, messy, or stingy in any of the three levels.

5.2 User Preference for Personality

5.2.1 Social-Entertaining. As shown in Figure 6, users' preference for the three *Social-Entertaining* levels differs depending on the scenario: On average, participants prefer to interact with the voice assistant rather high in *Social-Entertaining* when playing a song,

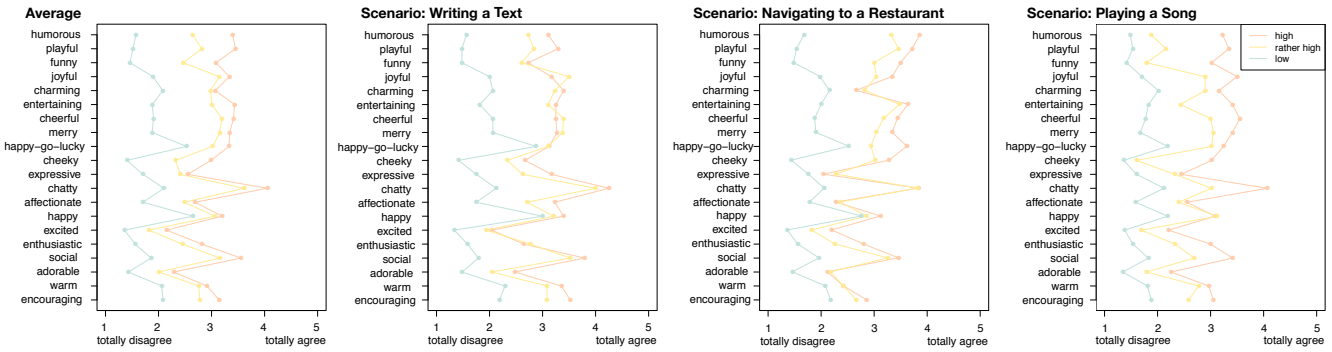

Figure 4: Participants' perception of the voice assistants' personality dimension *Social-Entertaining*, differentiating between low, rather high, and high level in the three scenarios and on average over the three scenarios.

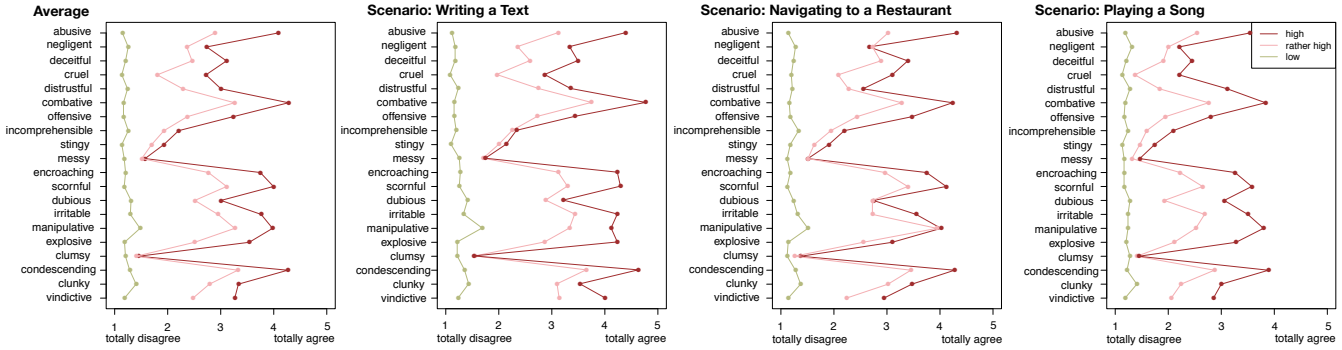

Figure 5: Participants' perception of the voice assistants' personality dimension *Confrontational*, differentiating between low, rather high, and high level in the three scenarios and on average over the three scenarios.

whilst the low level voice assistant was given preference for writing a text. Conversely, when navigating to a restaurant, participants did not seem to be convinced of any of the three options.

Since our data was not normally distributed ($p < .001$), we applied the Aligned Rank Transform (ART) procedure to the preference ratings, using the R-version of the ARTool [58]. A two-way mixed model ANOVA revealed a significant main effect of the level on user preference ($F(2, 306) = 4.73$, $p = .009$) and a significant interaction effect between level and scenario ($F(4, 306) = 6.94$, $p < .001$). Fig. 6 shows the significant pairwise post-hoc comparisons using Holm's method for p-value adjustment [19]. We only considered pairwise comparisons within each scenario (that is, differences in preference between levels) and between the same level over different scenarios since the other comparisons do not seem meaningful in the context of our research questions (e.g., a comparison between low level in the texting scenario and the high level in the music scenario).

These findings are supported by participants' ranking of the three levels after listening to all three versions. In the texting scenario, 52.1% of participants ranked the low level voice assistant as first, followed by the rather high (29.2%), and high (18.8%) ones. For the navigating to a restaurant scenario, 50.0% of participants also favoured the low level voice assistant, whilst the rather high one was picked by 28.0% and the high level one by 22.0%. In contrast, the ranking for the playing music scenario was more balanced, with 36.2% of participants selecting the rather high level voice assistant, 36.2% the high level one, and 27.6% the low level one as their favourite.

5.2.2 Confrontational. Figure 7 shows a clear preference of participants for the voice assistant low in *Confrontational* across all three scenarios. In contrast, most participants seem to reject the highly confrontational voice assistant. Interestingly, we can see a greater variance in participants' wish to interact with the voice assistant rather high in *Confrontational*.

Since the data was again not normally distributed ($p < .001$), we applied the Aligned Rank Transform (ART) procedure to the preference ratings, using the R-version of the ARTool [58]. A two-way mixed model ANOVA revealed a significant main effect of the level on user preference ($F(2, 306) = 270.30$, $p < .001$), of scenario on user preference ($F(2, 153) = 4.66$, $p = .011$), and a significant interaction effect between level and scenario ($F(4, 306) = 7.51$, $p < .001$). Fig. 7 shows the significant pairwise post-hoc comparisons using Holm's method for p-value adjustment [19]. We again only considered pairwise comparisons within each scenario (that is, differences in preference between levels) and between the same level over different scenarios.

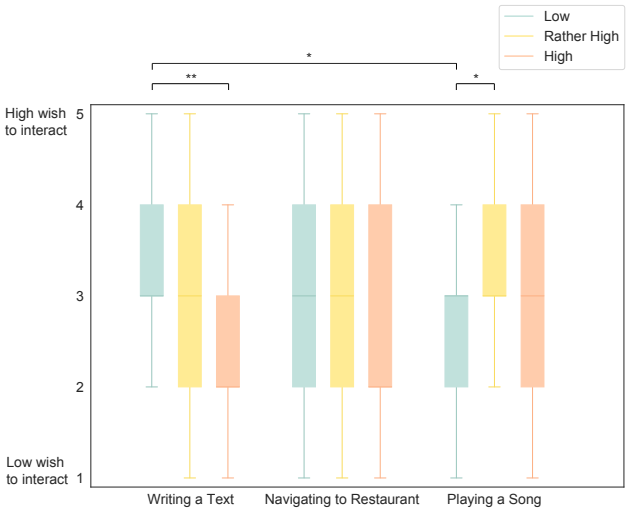

Figure 6: Participants' preference for interacting with the three levels of the voice assistants' personality dimension *Social-Entertaining* for all three scenarios. Statistically significant differences between two conditions are marked with * (p < .05), ** (p < .01), and * (p < .001).**

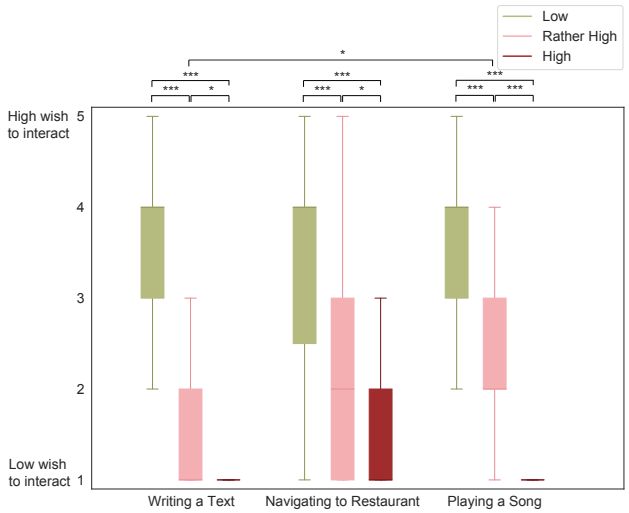

Figure 7: Participants' preference for interacting with the three levels of the voice assistants' personality dimension *Confrontational* for all three scenarios. Statistically significant differences between two conditions are marked with * (p < .05), ** (p < .01), and * (p < .001).**

Participants' ranking of the three levels echoed these findings. The low level assistant was picked as favourite by 76.5% of participants in the texting scenario, by 80.4% in the navigating to a restaurant scenario, and by 83.3% in the playing a song scenario. The rather high level one was ranked first most often in the texting scenario (17.6%), followed by navigating to a restaurant scenario

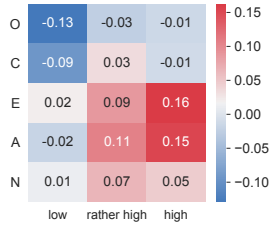

Figure 8: Spearman correlations of Big 5 personality scores and wish to interact with a voice assistant low, rather high, and high in *Social-Entertaining*.

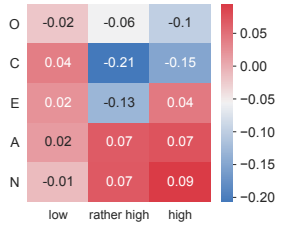

Figure 9: Spearman correlations of Big 5 personality scores and wish to interact with a voice assistant low, rather high, and high in *Confrontational*.

(9.8%), and the music scenario (9.3%). Conversely, the high level voice assistant was favoured only by 5.9% (texting), 7.4% (music), and 9.8% (navigating) of participants.

5.3 Relationship between Personality and User Preference

As an overview, Figures 8 and 9 show Spearman's Rank-order correlation coefficients between user personality and their wish to interact with the high, rather high, and low levels (mean over all three scenarios) of the *Social-Entertaining* and *Confrontational* voice assistants, respectively. For *Social-Entertaining*, we see weak positive associations of *Extraversion* and *Agreeableness* with a preference for the high level voice assistant. For *Confrontational*, we see a weak negative relationship between *Conscientiousness* and a preference for the rather high and high versions of the voice assistant.

In addition, we analysed the relationship between user personality and their preference to interact with the different voice assistants by fitting linear mixed-effects models (LMMs), using the R packages *lme4* [1] and *lmerTest* [29]. Following similar analyses in related work [60], we used LMMs to account for individual differences via random intercepts (for participant), in addition to the fixed effects (participants' Big Five personality dimension interacting with the VA personality level, scenario). In line with best-practice guidelines [37], we report LMM results in brief format here. For brevity, we only report on those models where we found a significant interaction effect between user personality and preference for a VA personality level.

For *Social-Entertaining*, the model had *Extraversion* as a significant positive predictor for preferring a VA *high* in this dimension (β=0.35, β_{std}=0.21, 95% CI=[0.01, 0.41], t=2.06, p<.05), indicating that people who are more extraverted tend to prefer a voice assistant that is high in *Social-Entertaining*.

For *Confrontational*, the model had *Conscientiousness* as a significant negative predictor for preferring a VA *rather high* in this dimension (β=-0.30, β_{std}=-0.15, 95% CI=[-0.60, -0.46], t=-2.02, p<.05), indicating that people who are less conscientious tend to prefer a voice assistant that is rather high in *Confrontational*.

6 LIMITATIONS

Our method and findings are limited in several ways and should be understood with these limitations in mind.

First, whilst our scenario selection was informed by the most prominent real-world use cases for in-car voice assistants [26], we collected users' evaluation of pre-recorded dialogues in an online survey. This created a more artificial setting, in which participants could not freely interact with the voice assistant themselves. Therefore, users might display different personality preferences in a real-world setting than in theory.

Second, we presented participants with a short-term interaction between a user and a voice assistant. We acknowledge that the preference for voice assistant personality also has to be examined in the context of long-term use in future work. For example, humorous phrases such as "Which feast to you desire, my master" are likely to get boring quickly if used repeatedly. Thus, future work investigating long-term use has to develop more variations of dialogues.

Third, in the texting and navigation scenarios, the voice assistant high in Social-Entertaining refused to comply with the user's request. This refusal could have impacted participants' preference for this voice assistant as users may expect voice assistants to follow their instruction. However, participants' preference for the high level voice assistant is also very low in the playing a song scenario, which did not involve a request refusal. In contrast, since the personality manipulation for the playing a song scenario worked worse than in the other two scenarios, refusing the user's request might be effective for creating the impression of a highly confrontational voice assistant.

Fourth, the expressiveness of the Amazon Alexa's voice is limited even though we configured para-linguistic features such as pitch, speech rate, and volume. We chose to implement our dialogues using Amazon Alexa to ensure that users easily identify the voice assistant in the recordings and to examine users' preference for personality in today's voice assistants. In particular since it is unclear if the human metaphor to voice assistant design is best suited [16], we wanted to examine users' preference on the premise that the voice assistant sounds artificially synthesised. However, Alexa's accentuation does not always fully correspond to the content, which might for example lead to a lack of delivering humour. Furthermore, there is no prior work on how voice assistant personality dimensions, such as *Social-Entertaining* and *Confrontational*, can best be expressed via para-linguistic features. To manipulate these features in our dialogues, we leveraged related work on human personality [49] and expressions of emotions in robots [13]. To determine whether this voice manipulation is adequate for signifying voice assistant personality traits, future work should examine these para-linguistic features separately from the verbal speech.

Finally, our sample is biased towards young users, with 73.1% of participants being younger than 25 years. For the purpose of examining the suitability of our approach, we focused on a younger sample which is used to the interaction with VAs [27]. However, since humour is subjective and perceived differently by age groups, the results obtained here should not be interpreted as representative of the whole population without further investigation.

7 DISCUSSION

7.1 Reflecting on the Synthesis of Different Levels of Personality

To the best of our knowledge, our work is the first to use interactive focus groups with amateur actresses and actors to synthesise different *levels* of theoretically grounded personality dimensions, which have been derived to explicitly describe *conversational agent* personality [57]. Whilst previous work pointed to the difficulties of generating personality perceptions as intended [7, 50], our findings suggest that enactment-based dialogue design is generally suitable to design different levels of personality, as indicated by the ranking of voice assistants in Figures 4 and 5. However, achieving extreme values, especially for a high level, has proven difficult, echoing previous work on synthesising personality in voice user interfaces [5, 39]. On the one hand, the dialogues might have been too short to express many different facets of the personality dimensions, e.g., the voice assistant being humorous, adorable, and excited at once. In particular, a voice assistant that unites all these characteristics in a brief dialogue might quickly give the impression of being "over the top".

On the other hand, since there is no validated personality questionnaire for voice assistants yet, we used the adjectives with the highest factor loadings, as determined by Völkel et al. [57] in their factor analysis. These adjectives might not all be similarly applicable or important to describe a certain voice assistant's personality. For example, our *Confrontational* voice assistants were rated low in clumsiness and messiness throughout all levels and scenarios. Since these are characteristics that are typically more likely to be observed in non-verbal behaviour, they may only play a secondary role in the practical design of voice assistant personality. This highlights the need for a validated personality instrument to sufficiently evaluate voice assistant personality.

7.2 Personality Preference is Context Dependent

Today's dialogues with voice assistant are characterised by adjacency pairs that revolve around requesting and responding to information [11, 21]. Furthermore, users clearly differentiate between purely functional and social talk when conversing with voice assistants [11], with the latter being often rejected by users [16]. Our findings indicate that the preference for personality levels for *Social-Entertaining* is context dependent, with users preferring a low level assistant in a task-oriented scenario such as writing a text, and a rather high level assistant for a more relaxed task such as picking a song. Asking the voice assistant to write a text message while driving requires trust in the voice assistant's capabilities since failing could have negative social repercussions. Conversely, participants' wish to interact with the voice assistant high in *Social-Entertaining* was overall lower but a higher variance could be seen,

echoing previous findings that only a subset of users appreciate humour in a voice assistant [16, 55].

For the dimension *Confrontational*, users' preference is clearer. Again, the ratings for the voice assistant rather high in *Confrontational* were significantly higher in the more relaxed scenario playing a song than in the task-oriented writing a text scenario. Furthermore, there is a greater variance for the scenario navigating to a restaurant. Some participants seemed to appreciate being challenged on their unhealthy food choice (i.e. eating fast food), a phenomenon also observed in prior work [55].

Whilst we developed very pronounced levels of *Confrontational*, future work should focus on generating moderate versions of a confrontational voice assistant to find a configuration, in which participants feel challenged in a productive way, yet not insulted. This preference by some participants for a voice assistant that challenges them could also be leveraged in the context of persuasive technology when trying to nudge users to adopt "better" behaviour styles.

In a wider view regarding the role of voice user interfaces in truly social conversations [11, 16, 45], our findings highlight that not all users refuse social, not task-related interaction with a voice assistant but some users seem to enjoy this kind of interaction, at least in a low-stake scenario while driving. Overall, these results motivate using low-stake scenarios as an entry point when adapting the voice assistant personality to the user. A context such as playing music can be used to find out whether this particular user enjoys interacting with a more chatty voice assistant. However, in particular the use of humour should be carefully aligned to the specific situation to avoid raising unrealistic expectations about the voice assistant's capabilities [31].

7.3 Extraverts Like Social-Entertaining, Conscientious Users Reject Confrontational

Our exploratory analysis indicates a small effect of personality on people's preference for *Social-Entertaining* and *Confrontational* in voice assistants. The magnitude of the correlation coefficients, shown in Figures 8 and 9, is in line with comparable previous research [36, 44].

Our results suggest that conscientious users tend to reject the interaction with a voice assistant rather high *Confrontational*. Conscientious people are characterised by being thorough and dutiful [33], usually conducting extensive research before making a decision [24]. Hence, it is likely that these users have thought about their request to the voice assistant and thus do not want the voice assistant to question them, especially without giving an explanation, echoing previous findings on conscientious people's disapproval of opinions in a voice assistant [55]. Besides, the texting scenario describes a situation in which the user arrives late to a meeting, something that usually does not happen to very conscientious people. This user therefore probably prefers if the voice assistant does not keep bringing this mistake up.

Our findings further suggest a small positive relation between Extraversion and a preference for a voice assistant high in *Social-Entertaining*. Extraverts tend to be sociable, dynamic, and cheerful, continuously searching for external stimulation [33]. It thus seems fitting that these users prefer a voice assistant that entertains them while sitting alone in a car. These results also reflect previous findings by Bickmore and Cassell [3] on extraverted users' preferences for social talk when interacting with voice user interfaces.

8 CONCLUSION AND FUTURE WORK

Whilst recent work has emphasised that deliberately manipulating the personality of a voice assistant positively impacts users' trust, acceptance, and liking [5, 7, 61], little is known about how to systematically synthesise voice assistant personality and adapt it to the user. We took an enactment-based dialogue design approach to imbue voice assistants with personality by asking amateur actors to write and enact dialogues between a user and a voice assistant, expressing different levels of *Social-Entertaining* and *Confrontational*; two dimensions which have been derived to explicitly describe *conversational agent* personality in previous research [57].

Overall, our results indicate that enactment-based dialogue design is suitable to synthesise different levels of *Social-Entertaining* and *Confrontational* although higher levels were not perceived as pronounced as intended. An opportunity to refine our approach could be to conduct joint focus groups with scriptwriters as dialogue experts, as is common in commercial VA design, and actresses and actors who can enact the dialogues interactively. We would expect that our approach can also be used to design other personality dimensions [57], such as *Social-Inclined* or *Serviceable*, which should be explored in future work. However, our findings also highlight the need for an adequate personality evaluation instrument to fully compare different levels.

Moreover, our findings suggest that users' preference for a certain personality level is both, to some extent, dependent on user personality and context. On the one hand, extraverts seemed to have a preference for higher *Social-Entertaining*, whilst conscientious users rather rejected a *Confrontational* voice assistant. On the other hand, users seemed to appreciate pronounced personality levels more in low-stake scenarios, such as playing a song, than in functional use cases, such as writing a text message. Thus, these low-stake scenarios could be used as an entry point to evaluate personality-infused voice assistants in a real-world setting.

Whilst in our study we mainly focused on pronounced personality levels at the extreme points of each dimension, most users will likely prefer more moderate personality types in their daily lives. However, these exaggerated versions of voice assistant personality are useful to examine the general suitability of our approach. Future work should therefore examine whether our approach is also applicable to develop more fine-grained moderate versions of these personality dimensions.

ACKNOWLEDGMENTS

We greatly thank Daniel Buschek, Sven Mayer, and Robin Welsch for their helpful feedback on the manuscript.

REFERENCES

[1] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. https://doi.org/10.18637/jss.v067.i01

[2] Timothy Bickmore and Justine Cassell. 2001. Relational Agents: A Model and Implementation of Building User Trust. In *Proceedings of the SIGCHI Conference*

on Human Factors in Computing Systems (Seattle, Washington, USA) *(CHI '01)*. ACM, New York, NY, USA, 396–403. https://doi.org/10.1145/365024.365304

[3] Timothy Bickmore and Justine Cassell. 2005. Social Dialogue with Embodied Conversational Agents. In *Advances in Natural Multimodal Dialogue Systems*, Jan C. J. van Kuppevelt, Laila Dybkjær, and Niels Ole Bernsen (Eds.). Springer Netherlands, Dordrecht, 23–54. https://doi.org/10.1007/1-4020-3933-6_2

[4] Michael Braun and Florian Alt. 2020. Identifying Personality Dimensions for Characters of Digital Agents. In *Character Computing*, Alia El Bolock, Yomna Abdelrahman, and Slim Abdennadher (Eds.). Springer International Publishing, New York, NY, USA, 123–137. https://doi.org/10.1007/978-3-030-15954-2_8

[5] Michael Braun, Anja Mainz, Ronee Chadowitz, Bastian Pfleging, and Florian Alt. 2019. At Your Service: Designing Voice Assistant Personalities to Improve Automotive User Interfaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland, UK) *(CHI '19)*. ACM, New York, NY, USA, Article 40, 11 pages. https://doi.org/10.1145/3290605.3300270

[6] Donn Erwin Byrne. 1971. *The attraction paradigm.* Academic Press, Cambridge, MA, USA.

[7] Angelo Cafaro, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2016. First Impressions in Human–Agent Virtual Encounters. *ACM Trans. Comput.-Hum. Interact.* 23, 4, Article 24 (Aug. 2016), 40 pages. https://doi.org/10.1145/2940325

[8] Anne Campbell and J. Philippe Rushton. 1978. Bodily communication and personality. *British Journal of Social and Clinical Psychology* 17, 1 (1978), 31–36. https://doi.org/10.1111/j.2044-8260.1978.tb00893.x

[9] DW Carment, CG Miles, and VB Cervin. 1965. Persuasiveness and persuasibility as related to intelligence and extraversion. *British Journal of Social and Clinical Psychology* 4, 1 (1965), 1–7. https://doi.org/10.1111/j.2044-8260.1965.tb00433.x

[10] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, and Benjamin R. Cowan. 2019. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers* 31, 4 (09 2019), 349–371. https://doi.org/10.1093/iwc/iwz016

[11] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 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* (Glasgow, Scotland UK) *(CHI '19)*. ACM, New York, NY, USA, Article 475, 12 pages. https://doi.org/10.1145/3290605.3300705

[12] Benjamin R. Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "What Can I Help You with?": Infrequent Users' Experiences of Intelligent Personal Assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Vienna, Austria) *(MobileHCI '17)*. ACM, New York, NY, USA, Article 43, 12 pages. https://doi.org/10.1145/3098279.3098539

[13] Joe Crumpton and Cindy L. Bethel. 2016. A Survey of Using Vocal Prosody to Convey Emotion in Robot Speech. *International Journal of Social Robotics* 8, 2 (2016), 271–285. https://doi.org/10.1007/s12369-015-0329-4

[14] Daniel Danner, Beatrice Rammstedt, Matthias Bluemke, Lisa Treiber, Sabrina Berres, Christopher Soto, and Oliver John. 2016. *Die deutsche Version des Big Five Inventory 2 (BFI-2)*. GESIS, Mannheim, Germany. https://doi.org/10.6102/zis247

[15] Jean-Marc Dewaele and Adrian Furnham. 2000. Personality and speech production: A pilot study of second language learners. *Personality and Individual Differences* 28, 2 (2000), 355–365. https://doi.org/10.1016/S0191-8869(99)00106-3

[16] 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* (Taipei, Taiwan) *(MobileHCI '19)*. ACM, New York, NY, USA, Article 5, 12 pages. https://doi.org/10.1145/3338286.3340116

[17] Patrick Ehrenbrink, Seif Osman, and Sebastian Möller. 2017. Google Now is for the Extraverted, Cortana for the Introverted: Investigating the Influence of Personality on IPA Preference. In *Proceedings of the 29th Australian Conference on Computer-Human Interaction* (Brisbane, Queensland, Australia) *(OZCHI '17)*. ACM, New York, NY, USA, 257–265. https://doi.org/10.1145/3152771.3152799

[18] Paul Ekman, E. Richard Sorenson, and Wallace V. Friesen. 1969. Pan-Cultural Elements in Facial Displays of Emotion. *Science* 164, 3875 (1969), 86–88. https://doi.org/10.1126/science.164.3875.86

[19] Lisa A. Elkin, Matthew Kay, James J. Higgins, and Jacob O. Wobbrock. 2021. An Aligned Rank Transform Procedure for Multifactor Contrast Tests. arXiv:2102.11824 [stat.ME]

[20] Adrian Furnham. 1990. Language and Personality. In *Handbook of language and social psychology*, William Peter Robinson and Howard Giles (Eds.). John Wiley & Sons, Chichester, UK, 73–95.

[21] Emer Gilmartin, Brendan Spillane, Maria O'Reilly, Ketong Su, Christian Saam, Benjamin R. Cowan, Nick Campbell, and Vincent Wade. 2017. Dialog Acts in Greeting and Leavetaking in Social Talk. In *Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents* (Glasgow, UK) *(ISIAA 2017)*. ACM, New York, NY, USA, 29–30. https://doi.org/10.1145/3139491.3139493

[22] Lewis R. Goldberg. 1981. Language and individual differences: The search for universals in personality lexicons. In *Review of Personality and Social Psychology*, L. Wheeler (Ed.). Vol. 2. Sage Publications, Beverly Hills, CA, USA, 141–166.

[23] Carrie Heeter. 1992. Being There: The Subjective Experience of Presence. *Presence: Teleoperators and Virtual Environments* 1, 2 (1992), 262–271. https://doi.org/10.1162/pres.1992.1.2.262

[24] Jannica Heinström. 2005. Fast surfing, broad scanning and deep diving: The influence of personality and study approach on students' information-seeking behavior. *Journal of Documentation* 61, 2 (2005), 228–247. https://doi.org/10.1108/00220410510585205

[25] Katherine Isbister and Clifford Nass. 2000. Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies* 53, 2 (2000), 251 – 267. https://doi.org/10.1006/ijhc.2000.0368

[26] Bret Kinsella and Ava Mutchler. 2020. In-car Voice Assistant Consumer Adoption Report. http://voicebot.ai/wp-content/uploads/2020/02/in_car_voice_assistant_consumer_adoption_report_2020_voicebot.pdf, accessed July 30, 2020.

[27] Bret Kinsella and Ava Mutchler. 2020. Smartphone Voice Assistant Consumer Adoption Report – Executive Summary. https://research.voicebot.ai/report-list/smartphone-voice-assistant-consumer-adoption-report-2020/, accessed May 24, 2021.

[28] Brigitte Krenn, Birgit Endrass, Felix Kistler, and Elisabeth André. 2014. Effects of Language Variety on Personality Perception in Embodied Conversational Agents. In *Human-Computer Interaction. Advanced Interaction Modalities and Techniques. HCI 2014. Lecture Notes in Computer Science*, Masaaki Kurosu (Ed.), Vol. 8511. Springer International Publishing, New York, NY, USA, 429–439. https://doi.org/10.1007/978-3-319-07230-2_41

[29] Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* 82, 13 (2017), 1–26. https://doi.org/10.18637/jss.v082.i13

[30] Kwan Min Lee and Clifford Nass. 2003. Designing Social Presence of Social Actors in Human Computer Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. ACM, New York, NY, USA, 289–296. https://doi.org/10.1145/642611.642662

[31] 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* (San Jose, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[32] François Mairesse and Marilyn A Walker. 2010. Towards personality-based user adaptation: psychologically informed stylistic language generation. *User Modeling and User-Adapted Interaction* 20, 3 (2010), 227–278. https://doi.org/10.1007/s11257-010-9076-2

[33] Robert R. McCrae and Paul T. Costa. 2008. A five-factor theory of personality. In *Handbook of Personality: Theory and Research*, O.P. John, R.W. Robins, and L.A. Pervin (Eds.). Vol. 3. The Guilford Press, New York, NY, USA, 159–181.

[34] Robert R. McCrae and Oliver P. John. 1992. An introduction to the five-factor model and its applications. *Journal of Personality* 60, 2 (1992), 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

[35] Margaret McRorie, Ian Sneddon, Gary McKeown, Elisabetta Bevacqua, Etienne de Sevin, and Catherine Pelachaud. 2012. Evaluation of Four Designed Virtual Agent Personalities. *IEEE Transactions on Affective Computing* 3, 3 (2012), 311–322. https://doi.org/10.1109/T-AFFC.2011.38

[36] Matthias R. Mehl, Samuel D. Gosling, and James W. Pennebaker. 2006. Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology* 90, 5 (2006), 862–877. https://doi.org/10.1037/0022-3514.90.5.862

[37] Lotte Meteyard and Robert A.I. Davies. 2020. Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language* 112 (2020), 104092. https://doi.org/10.1016/j.jml.2020.104092

[38] Clifford Nass and Scott Brave. 2005. *Wired for speech: How voice activates and advances the human-computer relationship.* MIT press, Cambridge, MA, USA.

[39] Clifford Nass and Kwan Min Lee. 2001. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied* 7, 3 (2001), 171. https://doi.org/10.1037/1076-898X.7.3.171

[40] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers Are Social Actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, Massachusetts, USA) *(CHI '94)*. ACM, New York, NY, USA, 72–78. https://doi.org/10.1145/191666.191703

[41] Michael Neff, Nicholas Toothman, Robeson Bowmani, Jean E. Fox Tree, and Marilyn A. Walker. 2011. Don't Scratch! Self-adaptors Reflect Emotional Stability. In *Intelligent Virtual Agents. IVA 2011. Lecture Notes in Computer Science*, Hannes Högni Vilhjálmsson, Stefan Kopp, Stacy Marsella, and Kristinn R. Thórisson (Eds.), Vol. 6895. Springer, Berlin, Heidelberg, 398–411. https://doi.org/10.1007/978-3-642-23974-8_43

[42] Jon Oberlander and Alastair J. Gill. 2004. Individual differences and implicit language: personality, parts-of-speech and pervasiveness. *Proceedings of the Annual Meeting of the Cognitive Science Society* 26 (2004), 1035–1040.

[43] M. Patterson and D.S. Holmes. 1966. Social Interaction Correlates of the MMPI Extraversion Introversion Scale. *American Psychologist* 21 (1966), 724–25.

[44] James W. Pennebaker and Laura A. King. 1999. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology* 77, 6 (1999), 1296–1312. https://doi.org/10.1037/0022-3514.77.6.1296

[45] 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* (Montreal, QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174214

[46] Atieh Poushneh. 2021. Humanizing voice assistant: The impact of voice assistant personality on consumers' attitudes and behaviors. *Journal of Retailing and Consumer Services* 58 (2021), 102283. https://doi.org/10.1016/j.jretconser.2020.102283

[47] Byron Reeves and Clifford Ivar Nass. 1996. *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge University Press, Cambridge, UK.

[48] D. R. Rutter, Ian E. Morley, and Jane C. Graham. 1972. Visual interaction in a group of introverts and extraverts. *European Journal of Social Psychology* 2, 4 (1972), 371–384. https://doi.org/10.1002/ejsp.2420020403

[49] Klaus Rainer Scherer. 1979. Personality markers in speech. In *Social Markers in Speech*, Klaus Rainer Scherer and Howard Giles (Eds.). Cambridge University Press, Cambridge, UK.

[50] Michael Schmitz, Antonio Krüger, and Sarah Schmidt. 2007. Modelling Personality in Voices of Talking Products through Prosodic Parameters. In *Proceedings of the 12th International Conference on Intelligent User Interfaces* (Honolulu, Hawaii, USA) *(IUI '07)*. ACM, New York, NY, USA, 313–316. https://doi.org/10.1145/1216295.1216355

[51] Christopher J. Soto and Oliver P. John. 2017. The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology* 113, 1 (2017), 117 – 143. https://doi.org/10.1037/pspp0000096

[52] Jürgen Trouvain, Sarah Schmidt, Marc Schröder, Michael Schmitz, and William J. Barry. 2006. Modelling personality features by changing prosody in synthetic speech. In *Proceedings of the 3rd International Conference on Speech Prosody*. TUDpress, Dresden, Germany, 4 pages. https://doi.org/10.22028/D291-25920

[53] Alessandro Vinciarelli and Gelareh Mohammadi. 2014. A survey of personality computing. *IEEE Transactions on Affective Computing* 5, 3 (2014), 273–291. https://doi.org/10.1109/TAFFC.2014.2330816

[54] James Vlahos. 2019. *Talk to Me: How Voice Computing Will Transform the Way We Live, Work, and Think.* Houghton Mifflin Harcourt, Boston, MA, USA.

[55] Sarah Theres Völkel, Daniel Buschek, Malin Eiband, Benjamin R. Cowan, and Heinrich Hussmann. 2021. Eliciting and Analysing Users' Envisioned Dialogues with Perfect Voice Assistants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 254, 15 pages. https://doi.org/10.1145/3411764.3445536

[56] Sarah Theres Völkel, Penelope Kempf, and Heinrich Hussmann. 2020. Personalised Chats with Voice Assistants: The User Perspective. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (Bilbao, Spain) *(CUI '20)*. ACM, New York, NY, USA, Article 53, 4 pages. https://doi.org/10.1145/3405755.3406156

[57] Sarah Theres Völkel, Ramona Schödel, Daniel Buschek, Clemens Stachl, Verena Winterhalter, Markus Bühner, and Heinrich Hussmann. 2020. Developing a Personality Model for Speech-Based Conversational Agents Using the Psycholexical Approach. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376210

[58] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 143–146. https://doi.org/10.1145/1978942.1978963

[59] Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents. *CoRR* abs/1901.08149 (2019). arXiv:1901.08149 http://arxiv.org/abs/1901.08149

[60] Yunhan Wu, Justin Edwards, Orla Cooney, Anna Bleakley, Philip R. Doyle, Leigh Clark, Daniel Rough, and Benjamin R. Cowan. 2020. Mental Workload and Language Production in Non-Native Speaker IPA Interaction. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (Bilbao, Spain) *(CUI '20)*. ACM, New York, NY, USA, Article 3, 8 pages. https://doi.org/10.1145/3405755.3406118

[61] Michelle X. Zhou, Gloria Mark, Jingyi Li, and Huahai Yang. 2019. Trusting Virtual Agents: The Effect of Personality. *ACM Trans. Interact. Intell. Syst.* 9, 2-3, Article 10 (March 2019), 36 pages. https://doi.org/10.1145/3232077