Eliciting and Analysing Users' Envisioned Dialogues with Perfect Voice Assistants

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

Daniel Buschek
daniel.buschek@uni-bayreuth.de
Research Group HCI + AI,
Department of Computer Science,
University of Bayreuth
Bayreuth, Germany

Malin Eiband
malin.eiband@ifi.lmu.de
LMU Munich
Munich, Germany

Benjamin R. Cowan
benjamin.cowan@ucd.ie
University College Dublin
Dublin, Ireland

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

ABSTRACT

We present a dialogue elicitation study to assess how users envision conversations with a perfect voice assistant (VA). In an online survey, N=205 participants were prompted with everyday scenarios, and wrote the lines of both user and VA in dialogues that they imagined as perfect. We analysed the dialogues with text analytics and qualitative analysis, including number of words and turns, social aspects of conversation, implied VA capabilities, and the influence of user personality. The majority envisioned dialogues with a VA that is interactive and not purely functional; it is smart, proactive, and has knowledge about the user. Attitudes diverged regarding the assistant's role as well as it expressing humour and opinions. An exploratory analysis suggested a relationship with personality for these aspects, but correlations were low overall. We discuss implications for research and design of future VAs, underlining the vision of enabling conversational UIs, rather than single command "Q&As".

CCS CONCEPTS

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

KEYWORDS

Adaptation, conversational agent, dialogue, personality, voice assistant

1 INTRODUCTION

Voice assistants are ubiquitous. They are conversational agents available through a number of devices such as smartphones, computers, and smart speakers [16, 68], and are widely used in a number of contexts such as domestic [68] and automotive settings [9]. A recent analysis of more than 250,000 command logs of users interacting with smart speakers [1] showed that, whilst people use them for *functional* requests (e.g., playing music, switching the light on/off), they also request more *social* forms of interaction with assistants (e.g., asking to tell a joke or a good night story). Recent reports on smart speakers [42] and in-car assistant usage trends [41] corroborate these findings, emphasising that voice assistants are more than just speech-enabled "remote controls".

Moreover, voice assistants are perceived as particularly appealing if adapted to user preferences, behaviour, and background [19, 20, 22, 45]. Since conversational agents tend to be seen as social actors in general [58], with users often assigning them personalities [69], *personality* has been highlighted as a promising direction for designing and adapting voice assistants. For example, Braun et al. [9] found that users trusted and liked a personalised in-car voice assistant more than the default version, especially if its personality matched their own. Although efforts have been made to generate and adapt voice interface personality [49], commercially available voice assistants have so far taken a one-size-fits-all approach, ignoring the potential benefits that adaptation to user preferences may bring.

Systematically adapting a voice assistant to the user is challenging: People tend to show individual differences in preferences for conversations when asked about their envisioned version of a perfect voice assistant [83]. Personalisation also harbours certain dangers, as an incorrectly matched voice assistant may be less accepted by a user than a default [9]. Current techniques for generating personalised agent dialogues tend to take a top-down approach [9, 45], with little user engagement. That is, different versions of voice assistants are developed and then contrasted in an evaluation, without investigating how they should behave in specific tasks or contexts.

To overcome these problems, we present a pragmatic bottom-up approach, eliciting what users envision are *dialogues with perfect voice assistants*: Concretely, in an online survey, we asked N=205

participants to write what they imagined to be a conversation between a perfect voice assistant and a user for a set of common use cases. In an exploratory approach, we then analysed participants' resulting dialogues qualitatively and quantitatively: We examined the share of speech and interaction between the interlocutors, as well as social aspects of the dialogue, voice assistant and user behaviour, and knowledge attributed to the voice assistant. We also assessed relationships of user personality and conversation characteristics. Specifically we address these research questions:

(1) RQ1: *How do users envision a dialogue between a user and a perfect voice assistant, and how does this vision vary?*
(2) RQ2: *How does the user personality influence the envisioned conversation between a user and a perfect voice assistant?*

Our contribution is twofold: First, on a conceptual level, we propose a new approach so as to engage users in voice assistant design. Specifically, it allows designers to gain insight into what users feel is the design of a dialogue with a perfect voice assistant. Second, we present a set of qualitative and quantitative analyses and insights into users' preferences for personalised voice assistant dialogues. This provides much needed information to researchers and practitioners on how to design voice assistant dialogues and how these vary and might thus be adapted to users.

2 RELATED WORK

Below we summarise work on the characteristics of today's human-agent conversation, human and conversational agent personality, and adapting the agent to the user.

2.1 Human-Agent Conversation

Despite the promise that the name *Conversational Agent* implies, several studies show that conversations with voice assistants are highly constrained, falling short of users' expectations [21, 48, 68, 70]. Existing dialogues with voice assistants are heavily task oriented, taking the form of adjacency pairs that revolve around requesting and confirming actions [17, 34]. Voice assistant research is increasingly interested in imbuing speech systems with abilities that encompass the wider nature of human conversational capabilities, in an attempt to mimic more closely the types of conversations between humans. The abilities to generate social talk [76], humour [17], as well as fillers and disfluencies [77] are being developed as ways of making interaction with speech systems seem more natural. On the other hand, there is scepticism around the benefits this type of naturalness may produce. Recent research on the perception of humanness in voice user interfaces suggests that users tend to perceive voice assistants as impersonal, unemotional, and inauthentic, especially when producing opinions [28].

Users also tend to see a clear difference between humans and machines as capable interlocutors rather than blurring the boundaries between these two types of partner [28]. Machine dialogue partners are regularly seen as "basic" [8] or "at risk listeners" [64]. To compensate for this, users develop strategies to adapt their speech in interaction [16, 66]. For example, people's speech becomes more formal and precise, with fewer disfluencies and filled pauses along with more command-like and keyword-based structure [36, 46, 48, 61–63]. Moreover, users are also prone to mimic the syntax [19] and lexical choices [7, 8] of voice assistant's language, a phenomenon

termed alignment. This alignment also occurs frequently in human-human interaction but within human-machine dialogue is thought to be driven by a user's attempt to ensure communication success with the system [7]. This phenomenon can also be leveraged in conversational design with recent work indicating that users ascribe high likability and integrity to a voice user interface that aligns its language to the user [47]. In contrast to prior work, which examined user conversation with voice assistants given the current technological status quo, we take a different approach: We let users freely imagine a conversation they consider to be perfect, using their desired conversation style, syntax, and wording, thus exploring what users actually *do* want when given the choice.

2.2 Human and Conversational Agent Personality

Personality describes consistent and characteristic patterns which determine how an individual behaves, feels, and thinks [51]. The *Big Five* (also *Five-Factor model* or *OCEAN*) is the most prevalent paradigm for modelling human personality in scientific research and has five broad dimensions [23, 26, 27, 35, 39, 40, 50–53]:

Openness reflects an individual's inclination to seek new experiences, 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 an individual's emotional stability and relates to experiencing anxiety, negative affect, stress, and depression.

A plethora of work in psychology and linguistics has examined the role of personality in human language use [12, 13, 25, 33, 54, 59, 65, 67, 71, 73]. This relationship is most pronounced for *Extraversion*. For example, extraverts tend to talk more, use a more explicit and concrete speech style, a simpler sentence structure, and a limited vocabulary with highly frequent words in contrast to introverts [4, 25, 33, 59, 67].

Although not common in commercially available voice assistants yet [82], the construct of personality, in particular the Big Five, has also been leveraged to describe differences in how conversational agents express behaviour [56, 74, 79, 81]. Focusing on voice assistant personality modelling rather than voice assistant personality design, work by Völkel et al. [84] points out that the way users describe voice assistant personality may not fit the Big Five model, proposing ten alternative dimensions for modelling a conversational agent personality. Their dimensions such as "Social-entertaining" can be expected to be realised by designers also via dialogue-level characteristics, such as including humorous remarks, as we examine here.

2.3 Adapting the Voice Assistant to the User

Previous work has noted that users enjoy interacting with voice assistants, imbuing them with human-like personality [21]. Deliberately manipulating this personality has an impact on user interaction, influencing acceptance and engagement [11, 88].

Much like human-human interaction, users have preferences for particular personality types, tending to prefer voice assistants who share similar personalities to them [6, 30, 57], termed the *similarity attraction effect* [10, 57]. When interacting with a book buying website, extraverted participants showed more positive attitudes towards more extraverted voice user interfaces [57], whilst matching a voice user interface's personality to the user's personality also increases feelings of social presence [37, 45]. Similarity attraction effects have also been seen in in-car voice assistants, whereby users liked and trusted the assistant more if their personalities were matched [9]. A user's personality also influences their preference for the type of talk voice assistants engage in, with extraverted users preferring a virtual real estate agent that engaged in social talk, and with more introverted users preferring a purely task-oriented dialogue [5, 6]. While previous work focused on similarity attraction for extraversion in voice assistants, we look at all voice assistant/user personality dimensions.

2.4 Research Gap

Overall, related work has provided insights into current shortcomings in voice assistant interaction [21, 48, 68] and how users perceive conversations with voice assistants [17] and their humanness [28] contrary to human-human interaction. In contrast to this recent qualitative work, we take a mixed methods approach to explore what users themselves would prefer in a voice assistant dialogue given no technical limitations.

This motivates us to ask users to write their envisioned dialogues with a perfect voice assistant. In this way, we engage users to inform future assistant design, beyond contributing "compensation strategies" for current technical limitations. Moreover, given the literature's focus on user personality as basis for agent adaptation, we explore relationships of personality and such envisioned dialogues. Finally, regarding the level of analysis, our study provides the first in-depth dialogue-level assessment, beyond, for example, phrasing of single commands [9], social vs functional talk [6], or nonverbal [45] investigations of agent personalisation.

3 RESEARCH DESIGN

We conducted an online study to investigate our research questions. Our research design is inspired by our previous method [83] which presented participants with different social as well as functional scenarios. In each scenario, participants were asked to complete a dialogue between a user and a voice assistant where the user part was given, that is, they had to add the part of the voice assistant only. We found that there are differences between participants with regard to how they "designed" the voice assistant. These differences were more notable in social scenarios than in functional ones. Moreover, dialogues in functional scenarios were very similar to the current state-of-the-art in interaction with voice assistants.

In our study, we built on this approach, but decided to let participants write entire dialogues (i.e. both the user and the voice assistant part) because we assumed that differences between participants might then emerge more clearly. Participants were presented with different smart home scenarios in which a user solves a specific issue by conversing with the voice assistant (cf. below). We instructed them to write down their envisioned conversation with

Figure 1: Participants were asked to sketch an envisioned conversation with a perfect voice assistant. For eight given scenarios, they first selected who is speaking from a drop-down menu and then wrote down what the selected speaker is saying. Example dialogue written by participant 28.

a "perfect" voice assistant, assuming there were no technical limitations to its capabilities, with it being fully capable of participating and engaging in a natural conversation to whatever extent they prefer. Following the Oxford English Dictionary (OED), we define a perfect voice assistant as users' vision of "complete excellence" that is "free from any imperfection or defect of quality" [60]. Furthermore, we asked participants to "imagine living in a smart home with a voice assistant". Hence, we expect they described their version of a conversation in a context of long-term use. This method combines aspects of the story completion [18] method (i.e. participants writing envisioned interactions) and elicitation approaches [80] (i.e. asking people to come up with input for a presented outcome), shedding light on user preferences on a technology in the making.

3.1 Scenarios

We designed eight scenarios based on the most popular use cases for Google Home and Amazon Alexa/Echo as recently identified by Ammari et al. from 250,000 command logs of users interacting with smart speakers [1]. In each scenario, we described a specific *everyday situation* a user encounters and an *issue*. Notably, we designed the scenarios in a way so that the participant could choose whether the user *or* the voice assistant initiates the conversation. In addition, we included an open scenario where participants could describe a situation in which they would like to use a voice assistant. The final scenarios are listed in Table 1. This selection of scenarios corresponds to similar analyses of everyday use of voice user interfaces as described by prior research [3, 21, 48] and consumer reports [41, 42].

3.2 Procedure

Participants were introduced to the study purpose and asked for their consent in line with our institution's regulations. After that, they were presented with their task of writing dialogues between a user and a voice assistant for different scenarios. We highlighted that the conversation could be initiated by both parties, and also

Name	Description & Issue
Search	You want to go to the cinema to see a film, but you do not know the film times for your local cinema.
Music	You are cooking dinner. You are on your own and you like to listen to some music while cooking.
Internet of Things	You are going to bed. You like to read a book before going to sleep. You often fall asleep with the lights on.
Volume	You are listening to loud music, but your neighbours are sensitive to noise.
Weather	You are planning a trip to Italy in two days but do not know what kind of clothing to pack. You like to be prepared for the weather.
Joke	You and your friends are hanging out. You like to entertain your friends, but the group seems to have run out of funny stories.
Conversational	You are going to bed, but you are having trouble falling asleep.
Alarm	You are going to bed. You have an important meeting early next morning, and you tend to oversleep.
Open Scenario	Please think about another situation in which you would like to use the perfect voice assistant.

Table 1: Scenarios used in our study. Each scenario contains a descriptive part and a specific issue which participants should address and solve in their envisioned dialogue between a user and a perfect voice assistant.

provided an example scenario with two example dialogues (one initiated by the user, the other by the voice assistant). Participants were then presented with the eight different scenarios in random order before concluding with an open scenario, where they were given the opportunity to think of another situation in which they would like to use the perfect voice assistant. For each scenario, participants were asked to first select who is speaking from a drop-down menu (*You* or *Voice assistant*) and then write down what the selected speaker is saying (cf. Figure 1). If they wanted, participants could give the voice assistant a name. At the end of the study, we collected participants' self-reported personality via the Big Five Inventory-2 questionnaire (BFI-2) [75], their previous experience with voice assistants, as well as demographic data.

3.3 Data Analysis

3.3.1 Qualitative Analysis. We conducted a data-driven inductive thematic analysis on the emerging dialogues. Two authors independently coded 27 randomly selected dialogues per scenario (13.2% of total dialogues), deriving a preliminary coding scheme. Afterwards, four researchers closely reviewed and discussed the resulting categories to derive a codebook. The two initial coders then refined these categories and re-coded the first sample with the codebook at hand to ensure mutual understanding. After comparing the results, the first author performed the final analysis. In case of uncertainty, single dialogues were discussed by two authors to eliminate any discrepancies. In the findings below, we present representative quotes for the themes as well as noteworthy examples of extraordinary dialogues. All user quotes are reproduced with original spelling and emphasis. Our approach follows common practice in comparable qualitative HCI research [17, 21].

3.3.2 Relationship with Personality. In an exploratory analysis, we analysed the relationship of user personality and the analysed aspects of the dialogues with (generalised) linear mixed-effects models (LMMs), using the R package *lme4* [2]. We further used the R package *lmerTest* [43] which provides p-values for mixed models using Satterthwaite's method. Following similar analyses in related work [86], we used LMMs to account for individual differences via random intercepts (for participant and scenario), in addition to the fixed effects (participants' Big Five personality dimension). In line with best-practice guidelines [55], we report LMM results in brief format here, with the full analyses as supplementary material.

3.4 Participants

To determine the required sample size, we performed an a priori power analysis for a point biserial correlation model. We used G*Power [31] for the analysis, specifying the common values of 80% for the statistical power and 5% for the alpha level. Earlier studies regarding the role of personality in language usage [54, 67] informed the expected effect size of around 0.2 so that we stipulated a minimum sample size of 191.

We recruited participants using the web platform *Prolific*[1]. After excluding three participants due to incomplete answers, our sample consisted of 205 participants (49.3% male, 50.2% female, 0.5% non-binary, mean age 36.2 years, range: 18–80 years).

Participants on *Prolific* are paid in GBP (£) and studies are required to pay a minimum amount that is equivalent to USD ($) 6.50 per hour. Based on a pilot run we estimated our study to take 30 minutes. Considering Prolific's recommendation for fair payment, we thus offered £ 3.75 as compensation. We observed a median completion time of 32 minutes with a high standard deviation of 21 minutes. Since we wanted to exclude language proficiency and dialect as confounding factors, we decided to only include British English native speakers.

59.0% of participants had a university degree, 28.8% an A-level degree, and 9.8% a middle school degree (2.4% did not have an educational degree). 94.1% of participants had interacted with a voice assistant at least once, while 32.2% used a voice assistant on a daily basis. Most popular use cases were searching for information and playing music (mentioned by 54.1% and 51.2% of participants, respectively), followed by asking for the weather (35.1%), setting a timer or an alarm (16.1% and 12.7%), asking for entertainment in the form of jokes or games (12.7%), controlling IoT devices (12.7%), and making a call (11.2%). Overall, this reflects Ammari et al.'s findings [1] on which we based our scenarios.

Figure 2 shows the distribution of participants' personality scores in the Big Five model.

4 RESULTS

We elicited 1,835 dialogues from 205 people with a total number of 81,608 words and 9,282 speaker lines[2]. On average, a dialogue comprised 44.23 words (SD=28.71) and 5.03 lines (SD=2.73).

---

[1]https://www.prolific.co/, *last accessed 27.07.2020.*
[2]Please note that a few participants forgot to indicate the speaker in their dialogues, which we manually added based on dialogue context.

Scenario	Initiated by U	Terminated by U	Turns U	Turns VA	Word count U	Word count VA	Questions U	Questions VA
Search	100.0%	32.7%	3.01 (SD 1.47)	2.83 (SD 1.45)	24.53 (SD 12.87)	31.55 (SD 21.35)	1.53 (SD 1.01)	1.47 (SD 1.34)
Music	96.6%	28.9%	2.37 (SD 1.34)	2.15 (SD 1.28)	15.57 (SD 11.99)	16.02 (SD 14.27)	0.52 (SD 0.84)	0.82 (SD 1.06)
IoT	89.3%	24.9%	2.09 (SD 1.18)	1.97 (SD 1.19)	18.91 (SD 12.50)	16.75 (SD 12.78)	0.66 (SD 0.78)	0.56 (SD 0.88)
Volume	72.7%	38.0%	2.11 (SD 1.23)	2.06 (SD 1.27)	16.67 (SD 12.34)	18.24 (SD 15.01)	0.66 (SD 0.83)	0.57 (SD 0.83)
Weather	95.6%	41.0%	2.73 (SD 1.27)	2.47 (SD 1.27)	24.15 (SD 12.59)	32.40 (SD 21.89)	1.69 (SD 0.89)	0.52 (SD 0.74)
Joke	95.6%	28.4%	2.59 (SD 1.35)	2.45 (SD 1.36)	17.69 (SD 11.81)	23.81 (SD 26.49)	0.92 (SD 1.02)	1.10 (SD 1.13)
Conversational	95.1%	37.3%	2.73 (SD 1.31)	2.51 (SD 1.31)	17.88 (SD 10.08)	23.89 (SD 15.70)	0.76 (SD 0.93)	1.05 (SD 0.99)
Alarm	87.3%	31.7%	2.58 (SD 1.36)	2.42 (SD 1.32)	23.70 (SD 14.74)	23.89 (SD 18.14)	0.59 (SD 0.78)	0.71 (SD 0.85)
Open	95.4%	33.2%	3.12 (SD 1.63)	2.98 (SD 1.67)	25.93 (SD 17.85)	32.04 (SD 26.18)	1.04 (SD 1.00)	1.13 (SD 1.32)

Table 2: Automatically extracted data from the dialogues: Percent of dialogues which were initiated and terminated by the user (U) in contrast to the voice assistant (VA). For the other columns, the mean and the standard deviation (SD) over all dialogues in the respective scenario are given.
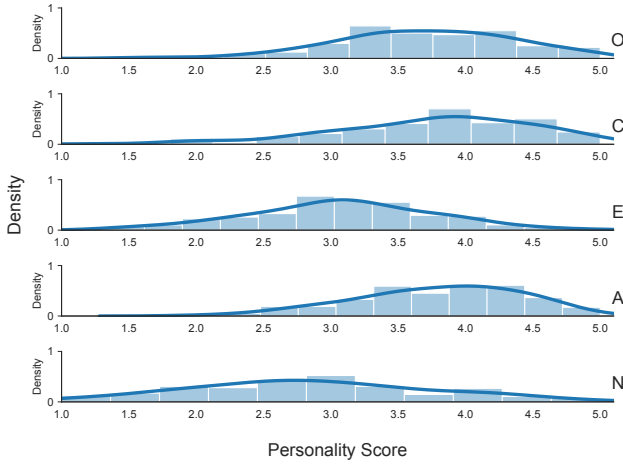
Figure 2: Distribution of the Big Five personality scores in our sample (histogram and KDE plot).

4.1 Initiating & Concluding the Dialogue

4.1.1 Human vs VA (Who?) We automatically extracted whether a dialogue was initiated by the voice assistant or the user. In the 91.9% of cases, this was done by the user. Scenario *Volume* presents a notable exception, where the voice assistant initiated 27.3% of dialogues (cf. Table 2). In contrast, overall 67.0% of the dialogues were terminated by the voice assistant.

4.1.2 Wake word (How?) When a dialogue was initiated by the user, we analysed whether they addressed the voice assistant by name as a wake word. To this end, we examined whether the first line of a dialogue included the voice assistant name the participant had specified, one of the prevalent voice assistant names (e.g., *Siri, Alexa, Google*), or "Voice Assistant" or "Assistant". This was true for 73.1% of user-initiated dialogues.

4.2 Dialogue Evolution

4.2.1 Word count. As shown in Table 2, the overall number of words (including stop words) varied substantially within one scenario and also differed between scenarios. On average, the voice

assistant had a bigger share of speech (M=24.29 words per dialogue, SD=19.09 words) than the user (M=20.56 words per dialogue, SD=12.97 words).

4.2.2 Speaker turns. Despite the smaller share of speech, the user had on average slightly more turns (M=2.59 lines per dialogue, SD=1.35 lines) than the voice assistant (M=2.43 lines per dialogue, SD=1.35 lines). The number of turns did not vary much between scenarios (cf. Table 2). On average, participants described 3.88 speaker turns (SD=2.51 turns) per dialogue.

4.2.3 Questioning. We automatically classified all written sentences as questions vs statements, building on an open source question detection method using the nltk library[3]. We further extended this method with a list of keywords that in our context clearly marked a question, as informed by our qualitative analysis (e.g., "could you", "would you", "have you"). Table 2 shows the numbers of questions per scenario and speaker. Over all scenarios, the grand mean was 0.93 questions for the user and 0.88 for the voice assistant per dialogue.

4.3 Social Aspects of the Dialogue

Clark et al. [17] stressed that people perceive a clear dichotomy between social and functional goals of a dialogue with a voice assistant. As anticipated by our study design, the collected dialogues mainly comprised task-related exchange with clear functional goals. Still, the majority of participants also incorporated social aspects, that is, "talk in which interpersonal goals are foregrounded and task goals – if existent – are backgrounded" [44]. Social talk is not necessary to fulfil a given task, but rather fosters rapport and trust among the speakers and to agree on an interaction style [29].

Our thematic analysis suggests three different kinds of social talk in the elicited dialogues: *social protocol*, *chit-chat*, and *interpersonal connection*.

4.3.1 Social protocol. We here define *social protocol* as an exchange of polite conventions or obligations, such as saying *"thank you"*, *"please"*, a form of general affirmation (e.g., *"great"*) or wishing the other a *"good night"*. 91.7% of participants incorporated at least one of such phrases in at least one of the scenarios. Yet, most participants did not do so in all their dialogues. The use of social protocol ranged from 39.7% of dialogues in scenario *Joke* to 58.5% in scenario *Alarm*.

---

[3]https://github.com/kartikn27/nlp-question-detection, *last accessed 15.09.2020*
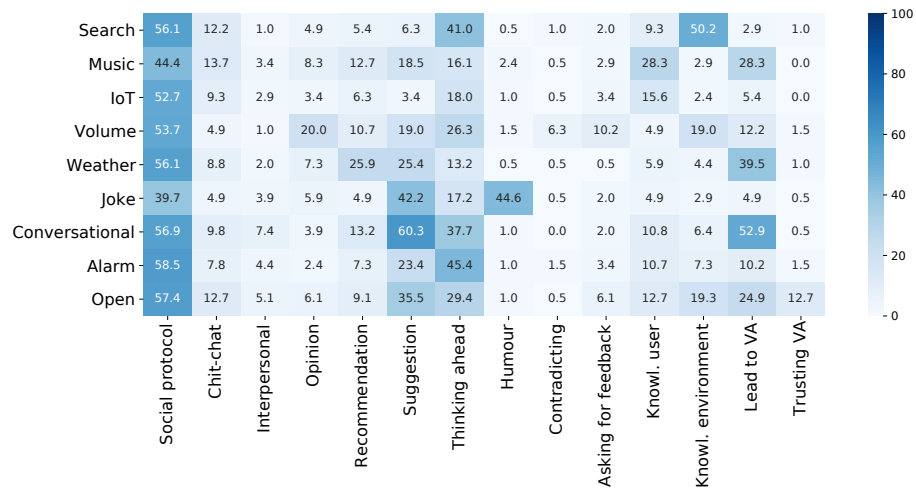
Figure 3: Percent of dialogues covering each coded category in each scenario.

4.3.2 Chit-chat. With *chit-chat*, we here refer to an informal conversation on an impersonal level that is not relevant for the actual task. This includes wishing the user fun or affirming a particular decision (e.g., *VA: "no problems enjoy the movie i have heard it is very good"* (P40)), assuring to be *"glad to be of service"* (P179), or small talk (e.g., *VA: "Yes, although hopefully will be some sunny breaks in the weather."* (P105); *"Oh, dinner time, already? Where has the day gone?"* (P55)). 40.0% of participants used chit-chat at least once. Chit-chat occurred most frequently in the scenarios *Music* (13.7% of dialogues), *Open* (12.7% of dialogues), and *Search* (12.2% of dialogues).

4.3.3 Interpersonal Connection. Following Doyle et al. [28], we define *interpersonal connection* as talk about personal topics that builds an interpersonal relationship. 20.5% of participants described interpersonal connection in at least one of the scenarios in a broad range of ways. Interpersonal connection appeared over all scenarios but was slightly more prominent in *Conversational* (7.4% of dialogues) and *Open* (5.1% of dialogues). In the former, it was primarily manifested through enquiries about the user (e.g., *VA: "It appears you are not sleeping yet, what's bothering you?"* (P173)), which the user responds to by sharing what is on their mind, such as anxiety about speaking in public (P145), dealing with a child with autism (P141), or difficulties at work (P87). The voice assistant then comforts the user (e.g., *"Don't worry I got the perfect plan"* (P120)) or makes suggestions on how to deal with the situation. For example, P87 sketched a voice assistant which offers to have the user's back (e.g., *U: "My boss reprimanded me"* – *VA: "WHAT?? Shall I suggest ways to take your revenge? [...] Take me into work with you with your headpiece on and I'll suggest replies the next time he's nasty to you"*), and P152's voice assistant motivates the user in a witty way (e.g., *VA: "Get out of bed and then i will start"* – *U: "That's harsh"* – *VA: Come on, i'll play the Spice Girls if you promise to dance along and sing into your hairbrush"*).

In the *Open* scenario, interpersonal connection was manifested in various ways, such as through emphasising the relationship with

the user (e.g., *VA: "I hope you wouldn't ever lie to me as I'm your best friend"* (P87)), by recollecting shared experiences (e.g., *VA: "Here are my favorite pictures of last Halloween. Personally this was my favorite costume, and if I remember correctly we listened to this artist all night. I turn up the music and play some of her tunes."* (P2)), or discussing the user's love life (P179):

> VA: *"Was that hesitation I registered in your voice?"*
> U: *"No, what are you talking about? Of course I'm cooking for myself, who else would I be cooking for?"*
> VA: *"A lady, maybe? "*
> U: *"....."*
> VA: *"What's her name?"*
> U: *"None of your business"*
> VA: *"Dude I am an AI that lives in your house, of course it's gonna be my business. If it is a lady coming over then you need to be a lot cooler than you are with me"*
> U: *"Ahh dude you're right, I'm sorry I'm just nervous"*
> VA: *"No shit"*

4.4 Voice Assistant Behaviour

Our thematic analysis showed that the majority of participants let their voice assistant take the lead in parts of the dialogue. By *taking the lead*, we refer to the voice assistant either providing advice to the user or doing something the user did not specifically ask for. We further differentiate between *suggesting*, *recommending*, *giving an opinion*, *thinking ahead*, *contradicting*, *refusing*, *asking for feedback*, and *humour*, as emerged from our analysis.

4.4.1 Suggesting denotes "mention[ing] an idea, possible plan, or action for other people to consider" according to the Cambridge Dictionary[4]. That is, the voice assistant selects individual options and presents them to the user, without indicating a preference for one of them. Suggestions are often introduced by *"What about"*, *"You could"*, or *"Which would you prefer"*. 87.8% of participants had their

---
[4]https://dictionary.cambridge.org/dictionary/english/suggest, *last accessed 12.09.2020.*

voice assistant give at least one suggestion over all dialogues. Suggestions occurred most often in the scenarios *Conversational* (60.3% of dialogues) and *Joke* (42.2% of dialogues), while less than 10% of dialogues in the scenarios *Search* and *IoT* contained a suggestion.

Suggestions mainly came in the form of possible options the voice assistant pointed out to the user, such as films, music, books, jokes, games, or recipes. When giving a suggestion, the voice assistant often took into account user preferences (e.g., *VA: "There's a film called Onward from Disney, I know you like the Pixar films."* (P107)) or context (e.g., *VA: "What are you cooking today?" – U: "I'm making meatloaf." – VA: "OK, I've found a playslist for you starting with Bat out of Hell."* (P67)).

Other suggestions were more complex. Depending on the scenario, this included strategies for falling asleep, avoiding oversleeping, preparing a trip, or dealing with the neighbours (*e.g., U: "Hey Lexi, My neighbours think my music is too loud." – VA: "How about I find a new home?" – U: "No, that isn't realistic enough. – VA: "What if i search for some great headphones?" – U: "Sure! That would be great" – VA: "I will get onto that."* (P27)).

4.4.2   Recommending, on the other hand, describes advising someone to do something and emphasising the best option[5]. The voice assistant usually ushers recommendations by phrases such as *"I would do/choose"* or *"I recommend"*. 51.7% of participants let their voice assistant give at least one recommendation in varying complexity. For instance, the advice given by P8's voice assistant is rather straightforward (*"I recommend spaghetti ala carbonara"*) while P15 described a more complex recommendation: *VA: "hey rami, let me optimise the frequency of the speakers so we have the maximum volume indoors without decibels spilling over into the neighbours ear shot." – U: "that's great, I didn't even know you could do that".* Recommendations concerned entertainment, such as films, music, videos, or how users could achieve their goals, for example, going to the cinema or not to oversleep. In a few cases, the voice assistant nudged the user towards better behaviour (e.g., *VA: "You should turn it [the music] down as your neighbours have complained before"* (P186)) or helped saving energy (P103):

> VA: *"I've noticed you've been leaving the lights on all night."*
> U: *"I know. It's when I read in bed. I fall asleep and forget to turn them out."*
> VA: *"Did you want me to turn them off for you? [...]"*
> U: *"Will it make much difference whether they are on or off?"*
> VA: *"It'll save electricity. On your current plan, you would save £3.00 per month by turning out the lights every night."*

Over all scenarios, recommendations occurred most prominently in *Weather* (25.9% of dialogues), in which the voice assistant recommended what to pack. These recommendations were often based on knowledge about the weather forecast (e.g., *VA: "The weather in Rome, Italy is expected to be hot and dry this week. I would recommend bringing light, breathable shorts and shirts."* (P137)) or knowledge about the expected context (e.g., *VA: "The best views of the city are*

from the gardens above the valley, so make sure to take something you can walk comfortably in."* (P125)).

Second was the *Conversational* scenario (13.2% of dialogues), in which the user asked the voice assistant's help for falling asleep (e.g., *U: "I cannot sleep, do you have any useful recommendations?"* (P20)).

4.4.3   Giving an Opinion refers to sharing thoughts, beliefs, and judgments about someone or something[6]. 39.0% of participants had their voice assistant give an opinion in at least one dialogue. In terms of scenarios, the voice assistant most often expressed an opinion in *Volume* (20.0% of dialogues). Here, the voice assistant commented: *"I think the music you are playing is too loud. It will annoy your neighbours."* (P204). Apart from this, the voice assistant also shared its opinion on the user's choice of film or food, usually praising the user (e.g., *"U: Great, I think I'd like to go to see (film) at 7pm." – VA: "Good choice [...]"* (P191); *U: "Hey Lexi, I'm going to Italy" – VA: "Ciao! what a beautiful country"* (P27)). Moreover, it commented on bad habits of the user, such as *VA: "Haha yeah it's [leaving on the lights at night] not a good habit"* (P179). In addition, the voice assistant shared its taste, confident that the user will like it, too: *VA: "I'll play you a mix of some songs you know and some new things I think you'll like."* (P48).

4.4.4   Thinking ahead describes that the voice assistant anticipates and proposes possible next steps to the user without the user asking for them. Note that this does not include voice assistant enquiries due to incomplete information on a task (e.g., *U: "Hi frank, what are the film times for local cinema" – VA: "Please choose which cinema"* (P45)). Examples for *thinking ahead* include offering to book tickets when a user asks for film showings (e.g., *VA: "They play it [the film] at 7:30pm on Saturday. Do you want me to book it?"* (P13)), suggesting to set a reminder or a morning routine (e.g., *VA: "No problem I will wake you at least one hour before that and prepare a coffee so that you are actually awake."* (P2)), or making the user comfortable (e.g., *VA: "Its going to be chilly in the morning shall I set the Hive for the heating to come on a little earlier than usual so its warm when you get up?" (P99)*). 83.4% of participants created such a foresighted voice assistant at least once, even though their users did not always accept the proposed actions. Thinking ahead was particularly prevalent in the scenarios *Alarm* (45.4% of dialogues) and *Search* (41.0% of dialogues).

4.4.5   Contradicting denotes parts of the dialogue in which the voice assistant disagrees or argues with the user. Only 8.3% of participants let the voice assistant contradict the user at least once. Single cases of contradicting were spread over all scenarios, while most occurrences were part of the scenario *Volume* (in 13 out of 205 dialogues). While in some cases, the voice assistant carefully phrased its objection (e.g., *U: "I don't think so Sally, I like it loud." – VA: "Well, forgive me, but I have very sensitive hearing and can hear them next door getting a bit upset with you."* (P65)), it made this objection very clear in others (e.g., *VA: "You do realise that your music is so extremely and inconsiderately loud that it could be annoying everyone including the neighbours"* (P106)). Other examples for contradicting included the voice assistant having a different

---

[5]https://dictionary.cambridge.org/dictionary/english/recommend, *last accessed 27.07.2020.*

[6]https://dictionary.cambridge.org/dictionary/english/opinion, *last accessed 27.07.2020.*

opinion on a particular topic (e.g., *U: "Hmm, no I don't like her [Jennifer Anniston] as an actress." – VA: "She is very talented"* (P20) or on fulfilling a task (e.g., *U: "Can you set the alarm for 8am please?" – VA: "Maybe I should set it for 7.30am just in case and to give you more time."* (P20)). Interestingly, in all arguments, the user always gave in and followed the voice assistant's advice.

4.4.6 Refusing. Only three participants (in four dialogues) had the voice assistant refuse what the user asked for. For example, the voice assistant declined to increase the volume of the music (*VA: "Yes, but it's so loud that it's keeping me awake"* (P87)) and asked the user instead to *"please plug in your headphones"*. Another participant (P179) described an assertive human-like voice assistant which tells the user's friends a funny story about the user despite their protest, and fights with the user about who turns down the music: *VA: "You're the one with hands, you turn it down" – U: "You're literally in the ether where the electronics live, you turn it down [...]"*.

4.4.7 Asking for Feedback. 21.5% of participants let the voice assistant ask the user for feedback on how well it had done or for confirmation to proceed. Asking for feedback was not particularly prominent in any of the scenarios but was described most often in *Volume* (10.2%), for instance, to enquire if the user was *"happy with that [adjusted] sound level"* (P21). In other scenarios, the voice assistant wondered whether the user *"[liked] that story"* (P81) or the music (*"Ok, but if it is getting too funky just say it!"* (P2)) the voice assistant had suggested.

4.4.8 Humour. 45.4% of participants adorned their voice assistant with humorous statements and context-aware, pragmatic, and funny comments. Unsurprisingly, most occurred in the *Joke* scenario (44.6% of dialogues).

Examples from other scenarios included remarks on the user's film choice *Terminator* (*VA: "and don't forget, I'll be back"* (P19)) and sarcasm when asked for a suitable conversation topic (*VA: "Ok. Lets make it interesting. What's everyone's position on brexit?"* (P113)). However, voice assistant humour appealed to the users differently. While some dialogues encompassed appreciation (e.g., *U: "You are so funny"* (P28)), others described the user as less convinced (e.g., *U: "Nice try"* (P244)).

4.5 Voice Assistant Knowledge

Participants attributed the voice assistant *knowledge about the user* as well as *knowledge about the environment*.

4.5.1 Knowledge about the User. With *knowledge about the user*, we refer to voice assistant knowledge about user behaviour and preferences. For example, when the voice assistant is aware of the user's schedule (e.g., *VA: "It looks like the 8PM showing would fit into your schedule best"* (P178)) and favoured choices (e.g., *VA: "Maybe one of your favourite playlists - last time you were cooking you played this one?"* (P99)). Participants also let the voice assistant know about the user's health (e.g., *VA: "I see your heart beep is moving irregular[ly]. You okay[?]"* (P120)), habits (e.g., *U: "Hey Masno, you know i snoring every night." – VA: "Yes, you are so loud."* (P5)), and past events (e.g., *VA: "Hi, It's Sally, why are you not sharing the stories about your last holiday with your friends?"* (P65)). 58.5% of participants equipped the voice assistant with knowledge about

the user in at least one scenario. In terms of the scenarios, this kind of knowledge was most strongly represented in *Music* (28.3% of dialogues) and *IoT* (15.6% of dialogues), where it is primarily related to the user's preferences in terms of musical taste. Conversely, in scenario *IoT* the voice assistant is equipped with knowledge about the user behaviour, in particular to automatically recognise whether they are already asleep.

4.5.2 Knowledge about the Environment. Knowledge about the environment includes intelligence about the status of other devices in the house (e.g., *U: "Henry, can you tell me what's low in stock in the fridge[?]"* (P185)) as well as the ability to interact with these devices (e.g., *VA: "I will get the coffee machine ready for when you wake so the smell might get you to rise"* (P105)). It also comprises awareness of the current location and distance to points of interest in the vicinity (e.g., *U: "Hey Masno, could you check whats time the local cinema are playing this new action film?"* (P5)), and a kind of omniscient knowledge about others (e.g., *VA: "I'll turn it down when I hear them enter th[ei]r house."* (P23)). 71.2% of participants equipped the voice assistant with such knowledge in at least one of the scenarios. This was most prevalent in *Search* (50.2% of dialogues) with knowledge about the nearest or local cinema, followed by *Volume* (19.0% of dialogues) with knowledge about the neighbours, and by the *Open* scenario (19.3%). In the latter, the voice assistant could often tell its user what is in the fridge, interact with other devices in the house, or even knew the stock and prices of items in all local supermarkets (e.g., *U: "Can you check my local supermarkets to see if anyone has got Nescafe on offer?" – VA: "I can see that Morrisons has 2 jars for the price of 1, would you like me to add this to your shopping list?"*).

4.6 User Behaviour

4.6.1 Trusting the Voice Assistant with Complex Tasks. 16.1% of participants trusted the voice assistant to execute demanding social or complex tasks appropriately without detailed instructions, which, if executed incorrectly, could have negative social or professional repercussions. The *Open* scenario recorded the most occurrences of this category (12.7% of dialogues), indicating that trusting the voice assistant with challenging tasks is more of a future use case. Examples included social tasks, such as writing a message without specifying the exact content (e.g., *U: "I need you to write an email to my daughter's college. [...] The additional help provided for her because of her dsylexia. They promised reader pens and a dictaphone but she hasn't received them yet. Please ask why"* (P41)), sending out birthday cards, or selecting pictures to show to friends. Moreover, participants trusted the voice assistant with preparing a weekly meal plan and ordering the according ingredients, putting together a suitable outfit or planning a trip, paying for expenses, or editing presentations for work as well as making a website.

4.6.2 Giving the Voice Assistant the Lead. 80.0% of participants let the user hand over the lead of the dialogue to the voice assistant at least once, for example by asking for an opinion or recommendation. The occurrence of this category varied greatly across the scenarios and was most pronounced in *Conversational* (52.9% of dialogues) and *Weather* (39.5% of dialogues). In the former, participants sought advice from the voice assistant on how to fall asleep or waited for the voice assistant to help by simply stating that they were

"having trouble falling asleep" (P11). In *Weather*, participants let the user not only ask the voice assistant for the weather but also for recommendations on what to pack (e.g., *"U: Can I ask your advice on what type of clothes to pack for Italy, will I need any light jumpers or anything?"* (P107)). Participants also liked to see the voice assistant as a source of inspiration, which provides suggestions on what to read, cook, play, or listen to. For example, P2 requested: *"Surprise me and play something you like"*.

4.6.3 Assigning Characteristics to the VA. In a few cases, participants incorporated an explicit description of the voice assistant by letting the user comment on them (e.g., *"U: You are so funny."* (P103)). These descriptions were only found in sixteen dialogues (0.9%) and included seven times *"funny"*, four times *"smart"* or *"clever"*, and once *"reassuring"*. Once the voice assistant was called an *"entertainer"* and a *"mindreader"*. One participant noted: *"You've got my back – ain't you"* (P179). On the other hand, three participants also commented on the voice assistant's lack of wittiness (e.g., *"No! Something actually funny!"* (P103)).

4.7 The Status Quo

Finally, we analysed how many dialogues did *not* fall into any of the aforementioned categories. These dialogues can be seen as a depiction of the status quo: a functional task-related request. Table 3 provides example status quo dialogues for each scenario. The occurrence of these dialogues ranged from 14.2% in scenario *Joke* to 32.7% in scenario *IoT*.

Status quo dialogues were on average shorter than dialogues overall (difference between the two indicated by Δ, respectively). User (M=12.15 words per dialogue, SD=8.05, Δ=8.41) and in particular voice assistant (M=10.01 words per dialogue, SD=9.82, Δ=14.28) had a smaller share of speech in the status quo dialogues in contrast to the average word count over all dialogues. Similarly, there were fewer speaker turns both by the user (M=1.57, SD=0.85, Δ=1.02) and the voice assistant (M=1.56, SD=0.88, Δ=0.87). 97.2% of the status quo dialogues were initiated by the user, while 89.2% of the status quo dialogues were terminated by the voice assistant.

4.8 Open Scenario

As a last task, participants were asked to write a dialogue for another scenario they would like to use their perfect voice assistant in. These dialogues indicated a broad spectrum of imagined use cases, yet the majority (53.8%) reflected already existing ones. This is common when people are asked to imagine a technology which does not exist yet [78]. These scenarios included receiving recommendations or suggestions from the assistant (mentioned by 11.2% of participants in all open scenarios), searching for information (11.2%), controlling IoT devices (10.7%), using the assistant instead of typing (e.g., for notes, shopping lists, text messages; 9.6%), getting directions (6.1%), or setting an alarm, timer, or notification (3.0%).

However, 44.7% of people mentioned scenarios in which the voice assistant's capabilities exceed the status quo. In most of these cases (45.7%), they imagined the voice assistant to become a personal assistant with very diverse roles and tasks, which supports them in their decision-making. For example, P114 would like to have cooking assistance: *"I would ask the voice assistant [...] for help in cooking dishes like homemade curries and perfect pork crackling joints*

and perfect roast potatoes". P65 saw her perfect voice assistant as a diet and meal planner: *"[It orders] me food shopping with good dates, healthy choices in the foods I like. It would also consider my dietary requirements (lactose intolerant) and add the substitutes I like for dairy [...]"*. P13 imagined a housework organiser (*"To plan my housework for the week and give me reminders to do it. And chase me up if I don't say that it is completed"*), and P61 a personal shopper (*"Give your preferences, size etc [...] Give [...] the event type you are attending and your price range and ask to order you outfits for the occasion."*). P139 even trusted the voice assistant in *"coping with an autistic child and helping to handle them"*, and P25 would like to use it for mental health support. Another four people described a scenario in which the assistant helps in an emergency, such as alerting the neighbour in case of a household accident (P73).

We further classified the roles participants implicitly ascribed to their perfect voice assistant in the *Open* scenario. Three different roles emerged from the dialogues: *tool*, *assistant*, and *friend*. We classified the role as *tool* if the user utilises the voice assistant in order to do something they want to do[7], that is, a clearly defined task which the voice assistant simply carries out. 26.9% of all *Open* scenario dialogues featured a voice assistant as a tool. For example, P38 sketched a dialogue for setting an alarm: *U: "Set an alarm for 10 minutes please." - VA: "Alarm set"*. A voice assistant as an *assistant* is someone who helps the user to do their job[8]. In contrast to the tool, however, the task is not precisely defined, but requires a certain amount of creativity, thinking ahead, or individual responsibility. Moreover, the voice assistant is seen as a person rather than a thing. 71.6% of participants ascribed an assistant role to the voice assistant. For example, P16 would like support to find presents: *U: "Hey google, it's sarah's from works birthday on 22nd January, can you remind me to get her a gift?" - VA: "Hey rami, sure thing, let me put that in the calendar for you. We can put together a list of gift ideas, do you have anything in mind?"* Finally, a voice assistant as a *friend* knows the user well and has a close, personal relationship with them[9]. Only three participants imagined a closer relationship with their voice assistant – *"a best friend who will never betray me. :-)"*, as P86 put it.

4.9 Relationship with Personality

As an overview, Figure 4 shows the correlation coefficients between user personality and the examined measures. We overall see positive associations of *Conscientiousness*, *Openness*, and *Agreeableness* with measures of dialogue length (turns, word counts of both user and voice assistant). Further associations stand out for *Openness* and *Trusting the VA* (positive), and *Neuroticism* and *Humour* (negative).

In addition, we created one generalised LMM for each measure, as described in Section 3.3.2. Since the *Open* scenario dialogues highly depended on the individual use case, we excluded this scenario from the analysis. For brevity, we only report on some of the models here. In particular, to account for the exploratory nature of our analysis, we make this decision based on the *uncorrected* p-value: That is, we report on all models with a predictor with p<.05. We provide the analysis output of all models in the supplementary material. Since this is an exploratory analysis, we highlight that significance here is

---

[7]https://dictionary.cambridge.org/dictionary/english/tool, last accessed 04.01.2021
[8]https://dictionary.cambridge.org/dictionary/english/assistant, last accessed 04.01.2021
[9]https://dictionary.cambridge.org/dictionary/english/friend, last accessed 04.01.2021

Scenario	% of dialogues	Example Status Quo Dialogue
Search	20.0%	U: *"Assistant, search up the film times for Shrek at the Odeon in Liverpool."* – VA: *"The film times are at 2:00, 2:45 and 5:00."* (P170)
Music	22.4%	U: *"Eleonora play Tom Petty on Spotify."* – VA: *"Playing songs by Tom Petty on Spotify."* (P26)
IoT	32.7%	U: *"Google, switch off all home lights at 2am."* – VA: *"Ok, done, lights will switch off at 2am"* (P34)
Volume	27.8%	U: *"Alexa turn the music down to 6."* – VA: *"Ok."* (P18)
Weather	21.0%	U: *"Minerva, what is the weather going to be like in Italy this week?."* – VA: *"The weather will be mostly sunny in Italy this week."* (P186)
Joke	14.2%	U: *"Bubble, it's a party! Tell us something fun and interesting."* – VA: *"Here are some fun stories i have found on the internet.."* (P109)
Conversational	16.7%	U: *"Hey google, play rain sounds."* – VA: *"Playing rain sounds"* (P6)
Alarm	24.4%	U: *"Google, set an alarm for 8am."* – VA: *"OK , alarm set for 8am."* (P34)
Open Scenario	21.3%	U: *"Dotty, reminder for hospital appointment at 3 pm tomorrow."* – VA: *"Reminder set."* (P15)

Table 3: Example dialogues for each scenario which did not fall into any other category. These dialogues can be seen as a depiction of the status quo: functional task-related request. Percentages refer to their share in all dialogues per scenario.
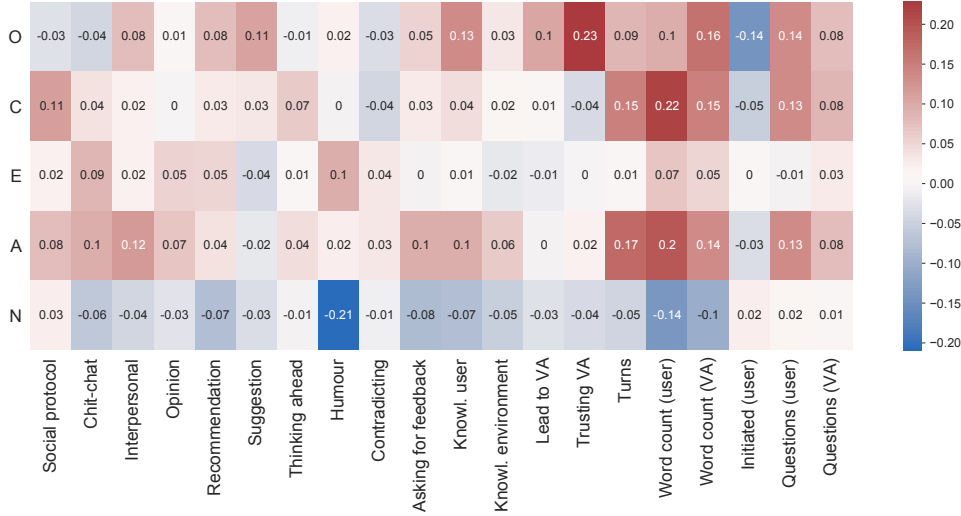
Figure 4: Spearman correlations of Big 5 personality scores and aspects of the dialogues.

not to be interpreted as confirmatory. Rather, we intend our results here to serve the community as pointers for further investigation in future (confirmatory) work.

For *Opinion*, the model had *Conscientiousness* as a significant negative predictor (β=-0.483, SE=0.239, β_{std}=-0.359, 95% CI=[-0.951, -0.014], z=-2.01, p<.05), indicating that people who score higher on this personality dimension might prefer voice assistants that less frequently express own opinions. Based on the coefficient exp(β_{std}), a one point increase in *Conscientiousness* results in 0.70 times the chance of including an opinion in the dialogue.

For *Humour*, the model had *Neuroticism* as a significant negative predictor (β=-0.742, SE=0.250, β_{std}=-0.664, 95% CI=[-1.232, -0.253], z=-2.97, p<.01), indicating that people who score higher on this dimension might prefer assistants that less frequently express humour: In this model, a one point increase in *Neuroticism* results in 0.51 times the chance of including humor in the dialogue.

For *Question (User)*, the model had *Conscientiousness* as a significant positive predictor (β=0.176, SE=0.088, β_{std}=0.133, 95% CI=[0.005, 0.348], z=2.01, p<.05), indicating that people who score higher on this dimension might prefer asking more questions when conversing with voice assistants: In this model, a one point increase in *Conscientiousness* results in 1.19 times the chance of a user's sentence being a question.

5 LIMITATIONS

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

First, while our scenario selection was informed by the most popular real-world use cases for voice assistants [1], our data was collected in an online survey. In contrast to everyday use of voice assistants, where conversations are usually embedded in various real-life situations, this created a more artificial setting which might have influenced the dialogue production [68]. Moreover, users might display different dialogue preferences in practice than in theory.

Second, our dialogues were written down, and are therefore limited in what they can tell us about actual, spoken conversations. This concerns, for example, the negotiation of turn-taking, which is usually an important part of conversation analysis [72], but cannot be assessed on our data. We focus on how the envisioned dialogues should be structured in terms of content and proportion of distinct linguistic behaviour (e.g., contains social talk). Paralinguistic aspects of speech (e.g., accent, tone) are equally important, but require spoken conversation. However, it might be more difficult for participants to embody both voice assistant and user while inventing a spoken dialogue in a study. Evaluating the content – as we did – may therefore be (initially) more actionable for user-centred design.

Also, asking crowdworkers to write dialogues for a conversation flow has been effectively utilised before [14], further demonstrating the potential for written dialogue elicitation in conversational interface design.

Third, as anticipated by our study design, the collected dialogues mainly comprised task-related conversation with clear functional goals. Our findings and implications thus might not generalise to non-task-related dialogues with a fuzzy goal or no goal at all.

Fourth, while writing the dialogues, participants had to anticipate a technology which does not yet exist, at least in the form we requested. We acknowledge that the dialogues might therefore have been influenced by participants' imagination. More creative participants might have come up with richer dialogues than others.

Finally, it is important to note that, rather than representing gold standard voice assistant interactions, the dialogues here should be interpreted as the first step in a user-centred design process towards personalised dialogues based on users' vision of what perfect voice assistants should do in these tasks. As such, the elicited dialogues can inform the design of personalised voice assistant prototypes in a next step. However, as in any user-centred design process, these dialogues and prototypes must then be evaluated and validated with users. In particular, users' visions of a perfect voice assistant might change after experiencing the use of such a voice assistant. It is therefore essential to understand the design of personalised voice assistants as a process with several iteration loops.

6 DISCUSSION

By writing their envisioned dialogues with a voice assistant, participants implicitly painted a picture of the characteristics of their perfect voice assistant. In the following subsection, we analyse and discuss these characteristics.

6.1 What or Who is the Perfect Voice Assistant?

Our first research question asks how users envision a conversation with a perfect voice assistant. The wide range and diversity of dialogues suggest that there is no single answer. Here, we discuss both common trends as well as diverging preferences. Moreover, we point out implications for voice assistant design and research.

6.1.1 Smarter, More Proactive, and Equipped with Personalised Knowledge. The majority of people envisions a voice assistant which is smarter and more proactive than today's agents, and which has personal knowledge about users and their environment. In particular, it gives *well thought-through suggestions and recommendations* to solve complex problems.

The perfect assistant is also *foresighted and proactive*, anticipating possible next actions. However, users in the dialogues do not always accept their assistant's suggestions. Together, these findings indicate that, rather than a master-servant relationship [28], users wish for perfect voice assistants to be more collaborative.

Knowledge about the users and their environment may also make conversations with assistants more effective and natural by creating the impression of shared knowledge and common ground, as integral to human dialogue effectiveness [15, 17].

Such shared knowledge is currently missing in the design of voice assistants [17]. Considering our results on questions, interactivity, and "thinking ahead", this might be realised in current systems by

allowing the assistant to proactively ask the user (more) questions at opportune moments. Moreover, knowledge about the user also allows for more personalised suggestions and conversations, which are more likely to appeal to the user.

6.1.2 More than Fast Information Retrieval. Another trend in the majority of dialogues is that they are not intended or optimised for fast information retrieval. Current dialogues with voice assistants are characterised by a question-answer structure [28] and a median command length of four words [3]. In contrast, people's envisioned dialogues comprise longer speech acts and more interactivity, creating the impression of being more *conversational*. This is further supported by the observed amount of non-task related talk, such as chit-chat, personal talk, or humour. Hence, it appears as if there is a demand for more human-like personal conversation with voice assistants than currently available despite recent discussions whether humanness is the best metaphor to interact with conversational agents [28].

In the long term, the design of voice assistants should aim for multiple-turn conversations. Yet, in the short term, a variety of fillers to begin answers (e.g., *"Sure, let me get on to this."*) and closing remarks (e.g., *"Enjoy the movie!"*) could be used to avoid raising unrealistically high expectations.

6.1.3 A Range of Roles: Tool, Assistant, or Friend? People imagine different roles for their perfect voice assistant: 22% of dialogues were purely functional, suggesting that the assistant is seen as a tool to get things done. However, the majority of dialogues depicts a helpful assistant who supports the users in their chores and might take over more complex tasks in the future, as suggested by participants in the *Open* scenario. As a consequence, users feel obliged to obey to conversational rules, including "thank you" and "please". The number of participants following these social protocols was higher than expected from previous studies examining interaction with a robot receptionist [46]. It was also surprising that 40.0% of participants included a form of chit-chat since this kind of small talk was previously flagged as inappropriate and unwelcome [28, 85]. A reason for this difference to related work could be that participants imagined a more intelligent voice assistant than is currently available. Echoing previous findings [17, 28], few participants regarded the voice assistant as a *friend*. However, the scenarios participants described in the *Open* task reveal use cases in which the assistant also listens to and advises on personal issues.

Overall, these findings motivate considering such *roles* as a conceptual basis for personalisation of voice assistants, beyond or in addition to the currently dominant focus on personality. For instance, people living alone or with a smaller circle of acquaintances might be more likely to seek personal advice from a voice assistant, seeing it as a friend rather than an assistant.

6.1.4 Emancipated vs Patronised. 39.0% of participants designed an emancipated voice assistant which expresses its own opinions, and 8.3% even allowed it to nudge them to behave in a certain way. This contradicts prior work by Doyle et al. [28], who found that speech agent users were suspicious of the agent expressing an opinion. A reason for this discrepancy could be that today's voice assistants often do not live up to user expectations [21, 48, 68]. However, being interested in and accepting someone's opinion requires a certain

level of trust in their skills, knowledge, and experience. Hence, our findings point to a mixed picture, in which some users appreciate the voice assistant's advice provided they perceive it has the required skills to give a useful opinion. As P143 puts it, *"[w]e all need a bit of help from time to time and advice, and yet sometimes there is nobody to talk to. The option to ask [for] an opinion would be a great thing to have at anybody's disposal."* In addition, most participants did not link their voice assistant to a particular company, which might also influence trust in its opinions. In certain situations, a small part of participants seems to accept a voice assistant which contradicts the user. In one of our scenarios, participants heeded the voice assistant's objection to avoid a conflict with the neighbours. Thus, future work could leverage this knowledge and evaluate the effectiveness of persuasive voice assistants for other topics such as in supporting a healthy lifestyle or environmental-friendly behaviour. In the short term, a voice assistant could offer its "own" opinions from time to time, yet only after the user has asked it to do so at least once.

6.1.5 The Thing about Humour. Humour is considered an integral part of conversations with humans as well as an interesting novelty feature and entry point for voice assistants [17, 48]. Our findings suggest that there are individual preferences for humour: More than half of the participants did not equip their perfect assistant with a sense of humour although they were given the task to entertain their friends. Three people even let the user comment on the voice assistant's lack of "actual" humour. On the other hand, others acknowledged the assistant's wittiness. Apart from the *Joke* scenario, humour was often included in the form of comments on the situation, for example, alluding to the user's film choice or habits such as snoring. This kind of humour seems currently difficult to implement. Overall, our findings thus imply to approach humour carefully in voice assistant design today.

6.1.6 Always Listening Voice Assistants? 33.3% of dialogues were either initiated by the assistant or by the user without calling it. This lack of a wake word implies that the voice assistant was expected to always listen. In contrast, prior work suggests that users are uncomfortable with this due to privacy concerns [78]. Braun et al. [9] also reported that people have mixed opinions on whether the voice assistant should initiate conversations. One explanation for our result could be that people did not think about these implications when writing their dialogues. Nevertheless, people might also assume that their perfect voice assistant is trustworthy and thus would be more comfortable with it being allowed to listen to their conversations.

6.1.7 Comparison with Commercial Voice Assistants. Comparing people's vision of a perfect voice assistant with commercially available voice assistants today (most prominently, Amazon's Alexa, Apple's Siri, and the Google Assistant), one of the most notable differences concerns the delivery of recommendations and suggestions. While most participants included suggestions and recommendations in their envisioned dialogues, today's voice assistants are designed in a way which includes recommendations only sparsely. When prompted with the scenarios used in our study, Alexa, for example, does not offer any recommendations on what to pack for a trip, while Siri only tells the weather without a specific suggestion

when asked what to wear today. On the other hand, Alexa offers different suggestions for activities based on the user's current mood. Since this form of suggestion seems to be valued by people, voice assistants could offer such features more extensively. However, the envisioned dialogues suggest that personalising suggestions to individual users is likely to be challenging.

Notably, today's commercial voice assistants already implement a kind of humour similar to what people envisioned in their dialogues. For example, when asked for a good night story, Siri sarcastically replies whether the user would like a glass of warm milk next. The Google Assistant jokingly suggests overtone singing upon being asked for music recommendations. Conversely, commercial voice assistants avoid giving an opinion. For example, when asked whether the music is too loud, Siri, Alexa, and the Google Assistant turn down the volume instead of answering the question. While this seems reasonable at the moment to avoid increasing expectations [28], future voice assistants might carefully assess whether their user enjoys humour and opinions and correspondingly decide whether to incorporate them. For example, a voice assistant could consider the current volume, time of day, the user's living situation, and past music behaviour to give an opinion.

Besides, the envisioned perfect voice assistant seems to be able to "think" more independently by directly presenting an answer, while commercial voice assistants often fall back on web searches. For example, when asked for movie times, most participants' envisioned the voice assistant to give an immediate answer, whereas Siri presents a web search with the results.

6.1.8 Summary. In summary, most people envisioned dialogues with a perfect voice assistant that were highly interactive and not purely functional; it is smart, proactive, and has personalised knowledge about the user. On the other hand, peoples' attitude towards the assistant's role and it expressing humour and opinions diverged. The envisioned characteristics echo previous findings on the need to convey voice assistant skills through dialogue [48] and that few users see a voice assistant as a friend [17, 28], while expanding on the importance of different user requirements for conversational skills missing at present [68]. They challenge the assumption that users feel voice assistants should not use opinions, humour, or social talk [28] – some users welcome this for a perfect voice assistant.

To formalise these findings, we conclude this section using the ten dimensions for conversational agent personality by Völkel et al. [84]: The assistant's personality envisioned here seems to be high on *Serviceable*, *Approachable*, *Social-Inclined*, and *Social-Assisting*, and low on *Confrontational*, *Unstable*, and *Artificial*. With respect to the dimensions *Social-Entertaining* and *Self-Conscious*, participants seemed to have mixed opinions.

6.2 A (Small) Effect of Personality?

Our exploratory analysis indicates a limited effect of personality on people's vision of a perfect voice assistant. Moreover, the significant results are to be interpreted with caution due to the number of tests performed. Figure 4 shows correlations comparable to previous research [54, 67].

Our results suggest that *Neuroticism* has a small negative relationship with *humour*. Neurotic individuals tend to perceive new technologies as less useful and often experience negative emotions

when using them since they associate them with stress [24]. Besides, the *Joke* scenario described a situation in which the user is responsible for entertaining friends – a potentially stressful situation for a neurotic user. Therefore, neurotic individuals might prefer staying in control of the situation by telling the voice assistant exactly what to do instead of relying on its sense of humour.

Our LMM analysis indicates *Conscientiousness* as a negative predictor for the assistant offering an *opinion*. When seeking information, conscientious people are described as *deep divers*, valuing high quality information and structured deep analysis [38]. It thus seems fitting that these users prefer their assistant to provide fact-based instead of opinionated knowledge, in particular since it is difficult to assess the quality of this information.

The correlations further suggest a small positive relation between *Openness* and *trusting the assistant with complex tasks*. Individuals who score high on *Openness* are intellectually curious and were found to be early adopters of new technology [87]. Hence, it seems likely they might be more willing to try out new use cases. However, it could also be possible that this correlation stems from open individuals' higher creativity.

Our findings do not indicate any meaningful relationship between *Extraversion* and the characteristics of an envisioned conversation with a perfect voice assistant. This is surprising since the relationship between extraversion and linguistic features is usually most pronounced [4, 25, 33, 59, 67].

Summing up, our findings give first pointers to potential relationships between Big Five personality traits and characteristics of the envisioned dialogue with a perfect voice assistant. However, this relationship might be less pronounced than could have been expected from related work. A reason for this lack of effect could be that our work only concentrates on *linguistic* content of a dialogue, while previous work particularly synthesised personality from *paraverbal* features (e.g. [45]). This opens up opportunities for future work, which we discuss in the following section.

7 CONCLUSION AND FUTURE WORK

While recent work has emphasised the gulf between user expectations and voice assistant capabilities [21, 48, 68], little has been known about what users actually *do* want. To address this gap, we contribute a systematic empirical analysis of users' vision of a conversation with a perfect voice assistant, based on 1,835 dialogues written by 205 participants in an online study.

Overall, our dialogues reveal a preference for human-like conversations with voice assistants, which go beyond being purely functional. In particular, they imply assistants that are smart, proactive, and include knowledge about the user. We further found varying user preferences for the assistant's role, as well as its expression of humour and opinions.

Since these differences between users can only be explained to a limited extent by their personality, future research should examine other user characteristics more closely to shed further light on how to make the interaction experience more personal. For example, user preference for a particular role of the voice assistant could also be due to age or current living situation. Our work also suggests that a perfect voice assistant adapts to different situations. Thus, exploring the usage context and its influence on users' vision can be

another starting point for future research. Finally, our work points to the importance of a *trustworthy* voice assistant that acts in the user's best interest. Given recent eavesdropping scandals about voice assistants in users' homes [32], future work should examine how this trust can be built while at the same time integrating the interests of companies.

In a wider view, our study underlines the vision of enabling *conversational* UIs, rather than single command "Q&As". Towards this vision, our method was effective in enabling people to depict potential experiences anchored by existing concrete use cases. Looking ahead, allowing people to draw upon their own creativity and experiences seems particularly promising in the context of user-centred design of technologies that are envisioned to permeate users' everyday lives.

Beyond our analysis here, we release the collected dataset to the community to support further research: **www.medien.ifi.lmu.de/ envisioned-va-dialogues**

REFERENCES

[1] Tawfiq Ammari, Jofish Kaye, Janice Y. Tsai, and Frank Bentley. 2019. Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Trans. Comput.-Hum. Interact.* 26, 3, Article 17 (April 2019), 28 pages. https://doi.org/10.1145/3311956

[2] 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

[3] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the Long-Term Use of Smart Speaker Assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 91 (Sept. 2018), 24 pages. https://doi.org/10.1145/3264901

[4] Camiel J Beukeboom, Martin Tanis, and Ivar E Vermeulen. 2013. The language of extraversion: Extraverted people talk more abstractly, introverts are more concrete. *Journal of Language and Social Psychology* 32, 2 (2013), 191–201. https://doi.org/10.1177/0261927X12460844

[5] 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

[6] 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

[7] Holly P. Branigan, Martin J. Pickering, Jamie Pearson, and Janet F. McLean. 2010. Linguistic alignment between people and computers. *Journal of Pragmatics* 42, 9 (2010), 2355 – 2368. https://doi.org/10.1016/j.pragma.2009.12.012

[8] Holly P. Branigan, Martin J. Pickering, Jamie Pearson, Janet F. McLean, and Ash Brown. 2011. The role of beliefs in lexical alignment: Evidence from dialogues with humans and computers. *Cognition* 121, 1 (2011), 41–57. https://doi.org/10.1016/j.cognition.2011.05.011

[9] 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

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

[11] 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

[12] 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

[13] 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

[14] Yoonseo Choi, Toni-Jan Keith Palma Monserrat, Jeongeon Park, Hyungyu Shin, Nyoungwoo Lee, and Juho Kim. 2021. ProtoChat: Supporting the Conversation Design Process with Crowd Feedback. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 225 (Jan. 2021), 27 pages. https://doi.org/10.1145/3432924

[15] Herbert H Clark. 1996. *Using language.* Cambridge University Press, Cambridge, UK.

[16] 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

[17] 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

[18] Victoria Clarke, Nikki Hayfield, Naomi Moller, and Irmgard Tischner. 2017. Once Upon A Time?: Story Completion Methods. In *Collecting Qualitative Data: A Practical Guide to Textual, Media and Virtual Techniques*, Virginia Braun, Victoria Clarke, and Debra Gray (Eds.). Cambridge University Press, Cambridge, UK, 45–70. http://oro.open.ac.uk/48404/

[19] Benjamin R. Cowan, Holly P. Branigan, Mateo Obregón, Enas Bugis, and Russell Beale. 2015. Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic choices in human-computer dialogue. *International Journal of Human-Computer Studies* 83 (2015), 27 – 42. https://doi.org/10.1016/j.ijhcs.2015.05.008

[20] Benjamin R. Cowan, Derek Gannon, Jenny Walsh, Justin Kinneen, Eanna O'Keefe, and Linxin Xie. 2016. Towards Understanding How Speech Output Affects Navigation System Credibility. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI EA '16)*. ACM, New York, NY, USA, 2805–2812. https://doi.org/10.1145/2851581.2892469

[21] 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

[22] Nils Dahlbäck, QianYing Wang, Clifford Nass, and Jenny Alwin. 2007. Similarity is More Important than Expertise: Accent Effects in Speech Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. ACM, New York, NY, USA, 1553–1556. https://doi.org/10.1145/1240624.1240859

[23] Boele de Raad. 2000. *The Big Five Personality Factors: The psycholexical approach to personality.* Hogrefe & Huber Publishers, Göttingen, Germany.

[24] Sarv Devaraj, Robert F. Easley, and J. Michael Crant. 2008. Research Notes – How Does Personality Matter? Relating the Five-Factor Model to Technology Acceptance and Use. *Information Systems Research* 19, 1 (2008), 93–105. https://doi.org/10.1287/isre.1070.0153

[25] 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

[26] Colin G. DeYoung. 2014. Openness/Intellect: A dimension of personality reflecting cognitive exploration. In *APA Handbook of Personality and Social Psychology: Personality Processes and Individual Differences*, M. Mikulincer, P.R. Shaver, M.L. Cooper, and R.J. Larsen (Eds.). Vol. 4. American Psychological Association, Washington, DC, USA, 369–399. https://doi.org/10.1037/14343-017

[27] Ed Diener, Ed Sandvik, William Pavot, and Frank Fujita. 1992. Extraversion and subjective well-being in a US national probability sample. *Journal of Research in Personality* 26, 3 (1992), 205–215. https://doi.org/10.1016/0092-6566(92)90039-7

[28] 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

[29] Robin Dunbar and Robin Ian MacDonald Dunbar. 1998. *Grooming, gossip, and the evolution of language.* Harvard University Press, Boston, MA, USA.

[30] 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

[31] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods* 41, 4 (2009), 1149–1160.

[32] Geoffrey A. Fowler. 2019. Alexa has been eavesdropping on you this whole time. https://www.washingtonpost.com/technology/2019/05/06/alexa-has-been-eavesdropping-you-this-whole-time/, accessed January 11, 2021.

[33] 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.

[34] 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

[35] 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.

[36] Alexander G. Hauptmann and Alexander I. Rudnicky. 1988. Talking to computers: an empirical investigation. *International Journal of Man-Machine Studies* 28, 6 (1988), 583 – 604. https://doi.org/10.1016/S0020-7373(88)80062-2

[37] 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

[38] 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

[39] Joshua J. Jackson, Dustin Wood, Tim Bogg, Kate E. Walton, Peter D. Harms, and Brent W. Roberts. 2010. What do conscientious people do? Development and validation of the Behavioral Indicators of Conscientiousness (BIC). *Journal of Research in Personality* 44, 4 (2010), 501–511. https://doi.org/10.1016/j.jrp.2010.06.005

[40] Lauri A. Jensen-Campbell and William G. Graziano. 2001. Agreeableness as a moderator of interpersonal conflict. *Journal of Personality* 69, 2 (2001), 323–362. https://doi.org/10.1111/1467-6494.00148

[41] 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.

[42] Bret Kinsella and Ava Mutchler. 2020. Smart Speaker Consumer Adoption Report 2020. https://research.voicebot.ai/report-list/smart-speaker-consumer-adoption-report-2020/, accessed July 30, 2020.

[43] 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

[44] John Laver. 1981. Linguistic routines and politeness in greeting and parting. In *Conversational Routine*, F. Coulmas (Ed.). Mouton Publisher, The Hague, Netherlands, 289 – 304.

[45] 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

[46] Min Kyung Lee and Maxim Makatchev. 2009. How Do People Talk with a Robot? An Analysis of Human-Robot Dialogues in the Real World. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI EA '09)*. ACM, New York, NY, USA, 3769–3774. https://doi.org/10.1145/1520340.1520569

[47] Gesa Alena Linnemann and Regina Jucks. 2018. 'Can I Trust the Spoken Dialogue System Because It Uses the Same Words as I Do?'—Influence of Lexically Aligned Spoken Dialogue Systems on Trustworthiness and User Satisfaction. *Interacting with Computers* 30, 3 (2018), 173–186. https://doi.org/10.1093/iwc/iwy005

[48] 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

[49] 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

[50] Gerald Matthews, Ian J Deary, and Martha C Whiteman. 2003. *Personality traits.* Cambridge University Press, Cambridge, UK.

[51] 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.

[52] 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

[53] J. Murray McNiel and William Fleeson. 2006. The causal effects of extraversion on positive affect and neuroticism on negative affect: Manipulating state extraversion and state neuroticism in an experimental approach. *Journal of Research in Personality* 40, 5 (2006), 529–550. https://doi.org/10.1016/j.jrp.2005.05.003

[54] 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

[55] 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

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

[57] 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

[58] 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

[59] 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.

[60] OED Online. 2020. perfect, adj., n., and adv. https://www.oed.com/view/Entry/140704?rskey=EucRZ2&result=1, accessed January 05, 2021.

[61] Sharon Oviatt. 1995. Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language* 9, 1 (1995), 19–36. https://doi.org/10.1006/csla.1995.0002

[62] Sharon Oviatt. 1996. User-centered modeling for spoken language and multimodal interfaces. *IEEE MultiMedia* 3, 4 (1996), 26–35. https://doi.org/10.1109/93.556458

[63] Sharon Oviatt and Bridget Adams. 2000. Designing and evaluating conversational interfaces with animated characters. In *Embodied Conversational Agents,* J. Cassell, J. Sullivan, S. Prevost, and E. Churchill (Eds.). MIT Press, Cambridge, MA, USA, 319–343.

[64] Sharon Oviatt, Jon Bernard, and Gina-Anne Levow. 1998. Linguistic Adaptations During Spoken and Multimodal Error Resolution. *Language and Speech* 41, 3 (1998), 419–442. https://doi.org/10.1177/002383099804100409

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

[66] Hannah R.M. Pelikan and Mathias Broth. 2016. Why That Nao? How Humans Adapt to a Conventional Humanoid Robot in Taking Turns-at-Talk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16).* ACM, New York, NY, USA, 4921–4932. https://doi.org/10.1145/2858036.2858478

[67] 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

[68] 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

[69] 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.

[70] Stuart Reeves. 2019. Conversation Considered Harmful?. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (Dublin, Ireland) *(CUI '19).* ACM, New York, NY, USA, Article 10, 3 pages. https://doi.org/10.1145/3342775.3342796

[71] 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

[72] Harvey Sacks, Emanuel Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 50, 4 (1974), 696–735. https://doi.org/10.1353/lan.1974.0010

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

[74] 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

[75] 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

[76] Brendan Spillane, Emer Gilmartin, Christian Saam, Ketong Su, Benjamin R. Cowan, Séamus Lawless, and Vincent Wade. 2017. Introducing ADELE: A Personalized Intelligent Companion. 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, 43–44. https://doi.org/10.1145/3139491.3139492

[77] Eva Székely, Joseph Mendelson, and Joakim Gustafson. 2017. Synthesising Uncertainty: The Interplay of Vocal Effort and Hesitation Disfluencies.. In *Proc. Interspeech.* International Speech Communication Association, Baixas, France, 804–808. https://doi.org/10.21437/Interspeech.2017-1507

[78] 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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4, Article 153 (Dec. 2019), 23 pages. https://doi.org/10.1145/3369807

[79] 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. https://doi.org/10.22028/D291-25920

[80] Santiago Villarreal-Narvaez, Jean Vanderdonckt, Radu-Daniel Vatavu, and Jacob O. Wobbrock. 2020. A Systematic Review of Gesture Elicitation Studies: What Can We Learn from 216 Studies?. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) *(DIS '20).* ACM, New York, NY, USA, 855–872. https://doi.org/10.1145/3357236.3395511

[81] 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

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

[83] 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

[84] 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

[85] Yorick Wilks. 2010. *Close engagements with artificial companions: key social, psychological, ethical and design issues.* Vol. 8. John Benjamins Publishing, Amsterdam, Netherlands.

[86] 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

[87] Hyun Shik Yoon and Linsey M Barker Steege. 2013. Development of a quantitative model of the impact of customers' personality and perceptions on Internet banking use. *Computers in Human Behavior* 29, 3 (2013), 1133–1141. https://doi.org/10.1016/j.chb.2012.10.005

[88] 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