

Developing a Personality Model for Speech-based Conversational Agents Using the Psycholexical Approach

Sarah Theres Völkel¹, Ramona Schoedel¹, Daniel Buschek², Clemens Stachl³,
Verena Winterhalter¹, Markus Bühner¹, Heinrich Hussmann¹

¹LMU Munich, Munich, Germany; ²Research Group HCI + AI, Department of Computer Science, University of Bayreuth, Bayreuth, Germany; ³Department of Communication, Stanford University, Stanford, US

sarah.voelkel@ifi.lmu.de, ramona.schoedel@psy.lmu.de, daniel.buschek@uni-bayreuth.de, cstachl@stanford.edu, verena.winterhalter@campus.lmu.de, buehner@psy.lmu.de, hussmann@ifi.lmu.de

ABSTRACT

We present the first systematic analysis of personality dimensions developed specifically to describe the personality of speech-based conversational agents. Following the psycholexical approach from psychology, we first report on a new multi-method approach to collect potentially descriptive adjectives from 1) a free description task in an online survey (228 unique descriptors), 2) an interaction task in the lab (176 unique descriptors), and 3) a text analysis of 30,000 online reviews of conversational agents (Alexa, Google Assistant, Cortana) (383 unique descriptors). We aggregate the results into a set of 349 adjectives, which are then rated by 744 people in an online survey. A factor analysis reveals that the commonly used Big Five model for human personality does not adequately describe agent personality. As an initial step to developing a personality model, we propose alternative dimensions and discuss implications for the design of agent personalities, personality-aware personalisation, and future research.

Author Keywords

Big 5; conversational agents; personality.

CCS Concepts

•Human-centered computing → Empirical studies in HCI;

INTRODUCTION

Speech-based conversational agents have become increasingly popular, often presented as helpful assistants in everyday tasks. Due to their intended use and conversational nature, this type of user interface seems much more likely to be seen as a being with a personality, compared to, for example, traditional

GUIs [15, 41]. This is further facilitated by recent developments that try to mimic human behavioural characteristics, such as casual filler sounds (“mmm”) on the phone [69].

Such developments highlight that assistants’ technical competence alone does not fulfill user needs, for example regarding acceptability in social contexts, yet also considering usability and user experience [3, 9, 53]: For instance, perceivable personality traits may help to communicate to the users that consistency in the agent’s reactions can be expected. Recent consumer reports reveal that users particularly enjoy interacting with voice assistants which exhibit a human-like personality [68]. More generally, integrating personality into rational agent architectures has been motivated by striving towards more complete cognitive models of agents, as well as sustaining more human-like interactions with people [9]. These insights motivate the deliberate design of personalities for speech-based conversational agents. However, systematically designing agent personalities remains a challenge: For example, conversational agents still struggle with generating adequate human-like “chitchat” or humour [19, 60], which often require demonstrations of (consistent) personality.

A crucial missing step towards realising these visions is a scientific model to describe agent personality in the first place. For example, researchers and UX designers could then use such a model to systematically design target personalities for their systems and applications. Moreover, while basic personality design follows a one size-fits-all approach, future voice assistants can be expected to also aim to adapt to the personality of the individual user. This further heightens the need for a personality model on which such adaptation can be based.

As of today, no dedicated personality model for speech-based conversational agents exists. Thus, most researchers have turned to the Big Five personality taxonomy. Since this model was developed for humans, it remains unclear if it is actually suitable to describe agents. For example, the Big Five dimension of *openness* might be less applicable or important for agents, while new dimensions might be necessary instead (e.g. one capturing an aspect of *artificiality*). Moreover, the Big Five taxonomy was developed with a psycholexical approach,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI '20, April 25–30, 2020, Honolulu, HI, USA.

© 2020 Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-6708-0/20/04 ...\$15.00.
<http://dx.doi.org/10.1145/3313831.3376210>

that is, using a set of adjectives collected for humans [21]. This set might not be adequate for conversational agents.

Recent work supports this view: Zhou et al. [70] openly asked about personality of chatbots. Responses included adjectives such as “robotic” which are clearly not covered by personality descriptors in models for humans. Similarly, Perez et al. [58] examined how users perceive brand personality of voice assistants and found out that people attributed “technical” or “logical” to these assistants.

We address this gap with the first systematic analysis of personality dimensions dedicated to speech-based conversational agents as initial step to developing a personality model. We focus on such agents (also named virtual personal assistants, intelligent assistants, voice user interfaces) due to their prevalent and close interaction with users [60]. To find appropriate personality descriptors and dimensions, we apply the psycholexical approach from psychology.

In particular, we contribute a set of 349 personality descriptors (adjectives), developed with a new multi-method strategy: We combine data from a free description task (online survey, 228 unique descriptors), an interaction task (lab study, 176 unique descriptors), and a text analysis of 30,000 online reviews of Alexa, Google Assistant, and Cortana, which yielded 383 unique descriptors. As an initial step to developing a personality model for speech-based conversational agents, we contribute ten personality dimensions, derived via exploratory factor analysis on data of 744 people rating our descriptors. The found dimensions do not correspond to the human Big Five – neither in number nor content. We discuss implications for future work and applications of our dimensions and descriptors.

THEORETICAL BACKGROUND & RELATED WORK

Personality describes relatively stable individual characteristic patterns of acting, feeling, and thinking [48]. Since these patterns are latent traits they cannot be gauged directly. Hence, adequate assessment of peoples’ personality has been a major challenge in empirical psychology for a long time [44].

Test Construction Process

The creation of an instrument for assessing psychological constructs (e.g. personality traits) is subject to a complex and multi-stage psychometric construction process [12, 27]. The *Inductive strategy* – used to develop human personality models [27] – assumes no elaborated theory a priori but merely an idea of which items could represent the construct [12, 27]. An item pool is created, rated by people, and analysed with exploratory factor analysis [12]. This statistical method aims to explain associations between items via a small number of homogeneous factors. These factors (i.e. dimensions) can then be used to develop a theoretical model [12].

A large item pool can be created with different methods [12]: The *experience-oriented-intuitive approach* asks people to determine indicators which best describe the construct based on their expertise/intuitive understanding. Another approach is the *collection and analysis of definitions* which reviews the

literature for indicators for the construct. The *person-centred-empirical approach* focuses on embedding of the construct in relation to similar constructs. Based on associations of personal characteristics and these constructs in existing literature indicators for the construct of interest are derived. Finally, the *analytical-empirical approach* uses standardised questionnaires, interviews, or observations to identify indicators relevant of the construct [12]. Our multi-method strategy covers multiple such approaches (cf. Our Approach Section).

Lexical Approach to Human Personality

The psycholexical approach [21, 38] is the most established and best known approach in psychology for developing a personality model for humans. It assumes that since people notice and talk about individual differences, these “will eventually become encoded into their language” [29].

Allport and Odbert [1, 38] collected a large set of 17,953 personality-related terms from a dictionary, Norman [56] even identified 27,000 unique personality terms. Researchers (e.g. [1, 17, 30, 31, 38, 55, 56]) reduced this by excluding unfamiliar, redundant, etc. terms based on expert and empirical ratings. In the 1980s Goldberg merged and refined existing lists [2, 32]. This resulted in 1,710 trait adjectives for self- and peer-report [38]. Goldberg further reduced these to 339 trait adjectives, sorting out synonyms and excluding difficult, slangy, ambiguous, and sex-related adjectives [29–31, 38].

A drawback of dictionaries is that language use is not considered (e.g. word frequencies) [18]. Digital data has enabled such new analyses and researchers have used *self-narrative* texts to extract personality-describing words [18, 42]. Our text analysis of online reviews makes use of this idea.

The Five-Factor Model of Personality

The above presented adjective lists have widely been used as basis for exploring the structure of human personality in previous research [38]. Despite various approaches to reduce trait terms to few dimensions of human personality across different adjective lists, the *Big Five*, also known as *Five-Factor Model* or *OCEAN*, has emerged as the most prominent model in personality research [20, 45, 46]. This model comprises five broad dimensions, which depict individuals’ tendencies of feeling, thinking, and behaving [21, 23, 24, 29, 36, 37, 44, 48–50]:

Openness relates to seeking new experiences, artistic interests, creativity, and intellectual curiosity. *Conscientiousness* relates to being thorough, organised, neat, reliable, careful, and responsible. *Extraversion* relates to being outgoing, sociable, active, and assertive in social interactions. *Agreeableness* relates to being friendly, helpful, socially harmonic, kind, trusting, and cooperative. *Neuroticism* relates to emotional stability and experiencing anxiety, negative affect, stress, and depression.

Despite the popularity of the Big Five, personality models are not without limitations. The assessment of personality dimensions via self-report questionnaires is subject to a certain measurement error [28]. For example, response biases due to social desirability have been reported [71]. Although the Big

Five are relatively robust across cultures, there is some controversy whether the model is best suited to describe personality in all cultures.

Describing Agent Personality in HCI

Humans attribute personality to other humans based on perceptible traces of feelings, thoughts, and behaviour [65]. According to the Media Equation [62], people tend to imitate this behaviour for machines and unconsciously assign them personality as well. For example, related work suggests that users make similar personality inferences for speech-based conversational agents as for humans [15, 35, 41, 51, 54]. More generally, acceptance and credibility of virtual agents are determined by their abilities to be perceived as having a consistent and coherent personality [3, 51, 53, 67]. This fundamentally motivates designing agent personality from an HCI perspective.

Today, speech-based conversational agents are most prominently deployed in intelligent personal assistants, such as Google Assistant, Apple's Siri, Microsoft's Cortana, and Amazon's Alexa. Commercially available conversational agents are increasingly used in smart speakers for home automation, e.g. Amazon Echo or Google Home [60], as well as in automotive user interfaces [11]. In research, conversational agents have for example been developed to give health advice to users [22] or serve as social companion for older adults [59]. Braun et al. [11] showed that users liked and trusted voice assistants more if their personality was matched to the user's own personality. Based on their use context, speech-based conversational agents are likely to show a different personality. For example, a speech-based conversational agent which answers calls at an insurance company might be designed to be reliable and trustworthy. In contrast, a conversational agent in a sports car could be designed as helpful, enthusiastic, and funny.

McRorie et al. [51] created speech-based conversational agents based on Eysenck's Three-Factor Model, using behavioural personality cues from the literature. Using a short version of Eysenck's questionnaire, they found that people perceived the agents as intended. Others used the Big Five model: Several researchers focused on extraversion which has the strongest links to observable behaviour [35, 41]. For example, Cafaro et al. [15] evaluated perceived extraversion and friendliness of a virtual museum guide. Moreover, Neff et al.'s findings [54] indicated that users recognise neuroticism in virtual agents. These projects used questionnaires for human personality to evaluate people's perception of agent personality [16, 35, 41, 54]. The examples show that perceived agent personality can be deliberately shaped. Yet it is unclear if dimensions of human personality are best suited to inform this since there is no alternative set of dimensions for agents.

Another line of research worked on frameworks for supporting implementation of agent personality: For example, the SOAR (State, Operator And Result) architecture was one of the first attempts to model agent behaviour. The psychological reasoning model BDI (Beliefs, Desires, Intentions) employs personality traits to decide between multiple goals of an agent [61, 64]. Moreover, Bouchet and Sansonnet [9] developed a computationally-oriented taxonomy for implementing

personality traits in voice agents. These frameworks all relied on existing personality models for humans.

In summary, the literature shows example agents with distinctive personality and also frameworks that aim to support their development – all using *human* personality models. To the best of our knowledge there is no systematic analysis of whether such human models are applicable and sufficiently comprehensive when describing speech-based conversational agent personality. This motivates our investigations in this paper.

There are reasons to expect differences in personality models suitable for humans vs agents: Prior work on human-robot interaction suggests that there are limitations of the Media Equation since humans conceptualise robots somewhere between alive and lifeless [6], and the perception of humanness in virtual assistants is a multidimensional construct [25]. Furthermore, in an open description of a chatbot's personality, users mentioned descriptors which are not present in the Big Five model, such as *robotic* [70]. Since personality is supposed to reflect *distinctive* traits, further dimensions beyond those in human models might be necessary to sufficiently describe agents. This motivates our work on developing a personality model for such agents.

OUR APPROACH

Similar to the traditional psycholexical approach in psychology, we implemented two steps to derive personality dimensions: (1) *Item Pool Generation*, in which unique phrases for describing personality are collected, and (2) *Exploratory Factor Analysis*, which explores structure and relationship between the items.

Item Pool Generation

The first step finds (English) terms that can “distinguish the behavior of one human being from that of another one” [1]. We seek such terms for agents. As adjectives “are used to describe qualities of objects and persons” [21] we limit our set to adjectives, as in prior work [30, 55]. We refer to these resulting terms as *descriptors*.

How to best compile a set of descriptors is a key challenge of this approach [21]. Inspired by traditional test construction theory [12], we use a new multi-method approach to collect potential descriptors:

1. *An online survey*, in which N=135 participants named descriptors for a chosen voice assistant in a free description task (*Experience-oriented-intuitive approach*).
2. *A lab experiment*, in which N=30 people interacted with agents (Siri, Alexa, Google Assistant) and described their personality afterwards (*Analytical-empirical approach*).
3. *A text analysis* of 30,000 online reviews of agents (Alexa, Google Assistant, Cortana) (*Narrative approach*).

Since conceptually it is not important how often a descriptor occurs, we collected all unique adjectives regardless of occurrence frequency. We joined the three sets with a given list of descriptors for human personality by Goldberg [31]. We then aggregated the results into a set of 349 adjectives by (1) applying pre-defined exclusion criteria, (2) clustering synonyms,

Online survey		Interaction experiment	
helpful	34%	helpful	90%
friendly	24%	friendly	67%
funny	19%	pleasant	53%
polite	13%	funny	37%
nice	13%	likeable	37%
annoying	9%	nice	37%
calm	9%	jolly	33%
cold	7%	polite	33%
intelligent	7%	unpleasant	33%
fast	6%	human	30%

Table 1. Top ten descriptors mentioned by participants to describe speech-based conversational agents in the online survey (left, N=135) and the interaction experiment (right, N=30). Percentages refer to number of people in the respective study who named each descriptor.

and (3) selecting descriptors based on word-frequency and ambiguity.

Exploratory Factor Analysis

In this second step, N=744 people rated one of the three most popular assistants (Alexa, Google Assistant, Siri) on the resulting 349 adjectives in an online survey. Exploratory factor analysis on these ratings then revealed latent personality dimensions.

DESCRIPTORS 1: ONLINE SURVEY

Research Design

We conducted an online survey to establish a first collection of descriptors for speech-based conversational voice agents. The survey allowed us to get an initial overview to inform subsequent steps; it thus comprised three parts: First, people were asked to indicate with which agents they had interacted before (Siri, Alexa, Google Assistant, Cortana, Samsung Bixby, potential others). Second, participants were asked to provide five representative adjectives to describe the personality of one speech-based conversational agent. Third, participants provided demographic information.

Data Analysis

We corrected typos (e.g. *helpful* to *helpful*), replaced nouns with adjective versions where possible (e.g. *fun* to *funny*), and simplified multiple word expressions (e.g. *sometimes annoying* to *annoying*). Furthermore, we excluded all answers which referred to an evaluation of outward appearance (e.g. *good looking*) or usage (e.g. *I don't use it*) instead of personality, as well as further unrelated phrases (e.g. *home button*).

Participants

We recruited participants via university mailing lists, social media, and online survey communities: 135 participants completed the survey (71.1% female; mean age 26.2 years, range 18-68 years). 31.1% of participants had interacted with Alexa before, 69.7% with Siri, 32.3% with Cortana, 44.4% with Google Assistant, and 5.2% with Samsung Bixby. Only one participant mentioned additional experience with another speech-based conversational agent (*BMW Car Assistant*).

Results

Our analysis yielded 228 unique descriptors (adjectives): 68.7% of these were mentioned once. Only five were stated

by more than 10% of people (cf. Table 1). In contrast to traditional descriptors for human personality, this set also included adjectives such as *robotic*, *(in)human*, or *impersonal*.

DESCRIPTORS 2: INTERACTION EXPERIMENT

Research Design

As another approach to collecting descriptors of speech-based conversational agents, we conducted a lab experiment with an interaction task: Here, our goal was to elicit descriptions directly after participants had interacted with such agents. In particular, we followed a within-groups design, where each participant interacted with three voice assistants (Siri, Alexa, and Google Assistant).

For each assistant, we asked participants to perform seven tasks, for example to send a message, play a song, or tell a joke. These tasks were informed by related work which analysed the most popular requests to voice assistants at home [40]. We counterbalanced the order of assistants and interaction tasks.

Interviews followed: To allow participants to become familiar with describing personality characteristics, we first asked them to describe the personality of a friend or family member. Then, after each interaction, participants described the personality of the respective speech-based conversational agent. At the end, they filled in a short questionnaire for demographic data. The experiment took between 45 and 60 minutes.

Data Analysis

We transcribed the interviews and collected all descriptors from the transcripts. Since the interviews were conducted in participants' native language, two authors individually translated all descriptors, then cross-checked the translations. For standardisation, we turned negations into the corresponding antonyms (e.g. *not funny* to *unfunny*).

Participants

The sample consisted of N=30 participants (73.3% female; mean age 24.5 years, range 18-39 years), which were recruited via university mailing lists and personal contacts. 93.3% of participants knew Siri, 90% Alexa, 90% Google Assistant, and 36.7% Cortana respectively before the study. 87% of participants interacted at least once with a voice assistant before, and 50% more than once a week.

Results

The experiment resulted in 176 unique descriptors. Of those 176 descriptors, 110 descriptors were new compared to study one. Again, a majority of descriptors (50.9%) was only mentioned once. The top ten descriptors can be found in Table 1. Some of the most frequently used descriptors overlap with the ones from the online survey. However, it is interesting to note that more positive affective descriptors were included in the top ten list, such as *likable* and *jolly*.

DESCRIPTORS 3: TEXT MINING ON ONLINE REVIEWS

Data Acquisition

In the first two studies, we collected descriptors by explicitly enquiring people about personality traits. In this third study we

examined user reviews for descriptors, in order to also include implicit depictions of personality in everyday language use and to cover a wider sample. Reviews provide an interesting source as they reflect users’ (emotional) experience with an application [26,33,43]. Following related work [26], we built a web crawler to scrape the latest 10,000 US Google Play Store reviews for Google Assistant, Alexa, and Cortana respectively. We did not include Siri since it is not available in such a store.

Data Processing and Analysis

Inspired by related work [26] on users’ problems with intelligent everyday applications, we combined automatic and manual analysis: We first used the Python Natural Language Toolkit¹ (NLTK) to automatically extract all adjectives and adverbs from reviews. This resulted in 794 terms for Google Assistant, 913 terms for Cortana, and 1,068 terms for Alexa (incl. intersections).

Adjectives in reviews might not only reflect users’ evaluation of personality but also refer to specific features (e.g. “*sucky recognition technology*” or to describe the user (e.g. “*this problem makes me angry*”). Therefore, two authors manually examined all adjectives by going through a random set of reviews per adjective to decide whether this adjective qualified as a descriptor. We excluded descriptors which refer to a state rather than a stable trait (e.g. “*offline*”). For the majority of adjectives, up to twenty reviews were sufficient to decide on exclusion. Overall, we included adjectives favourably since our aim was to generate a comprehensive pool of descriptors. Thus, if it was not clear whether an adjective described the speech-based conversational agent or a specific feature, we included it (e.g. “*very useful*”).

With these criteria, two authors independently went through a random set of hundred adjectives for Google Assistant (corresponds to 886 reviews). The interrater agreement was Cohen’s $\kappa = 0.82$. We then compared the results and discussed discrepancies until consensus was reached. The remaining 694 adjectives were split evenly among raters. We repeated this for Alexa and Cortana: Here, we only analysed adjectives which had not been already included for Google Assistant. Again, to ensure interrater consistency, two authors both rated the first 50 adjectives for Alexa and Cortana, respectively. The interrater agreement was $\kappa = 0.92$ for Alexa and $\kappa = 0.91$ for Cortana. Afterwards, reviews were again split among raters.

Results

Our analysis yielded 383 unique descriptors; 288 of those were not included in the set from studies one and two. Examples include *busy*, *clunky*, *inoperable*, *laggy*, *magical*, *philosophical*, *romantic*, *temperamental*, *usable*, *virtual*. Given that many reviews are short or only repeat the star rating [26, 43], the number of adjectives seems adequate despite the high number of analysed reviews. Furthermore, we noted that few adjectives occur very frequently (e.g. *helpful* was included in 236 Cortana reviews), while the majority of adjectives appears only occasionally. Since an adjective can occur n times over

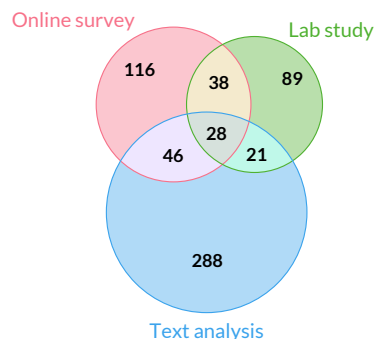


Figure 1. Overview of the candidate descriptors before refinement, as collected with each method and with multiple such methods. The figure shows the value of our multi-method approach: Each of the three methods added new descriptors not found by the others.

all 30,000 reviews but may only be used $i \leq n$ times as a descriptor for personality, we deem it not meaningful to show a frequency distribution here.

FINAL SET OF DESCRIPTORS

As a visual overview, Figure 1 shows our three obtained sets of candidates (i.e. descriptors before final selection). It is striking that the descriptors collected in the three different methods show only small overlaps. While 28 descriptors were named in all three methods, 493 descriptors were only found in one of the three approaches. We will discuss the implications of this small overlap in the Discussion section. Next we describe how we derived the final set.

Adding Existing Descriptors for Human Personality

We also accounted for the possibility that traditional human personality descriptors may be suitable to describe speech-based conversational agents. Hence, we merged our set with Goldberg’s established list of 339 adjectives for human personality [31]. We chose Goldberg’s list instead of using the items of personality questionnaires such as the Revised NEO Personality Inventory (NEO-PI-R) [20] or the Big Five Structure Inventory (BFSI) [5] since Goldberg’s items are openly published so that other researchers can build their work on our item list. The resulting list comprised 870 descriptors.

Refining the Set by Established Exclusion Criteria

We refined this list in several iterations. In line with the construction of traditional personality inventories in psychology [30, 55], we applied the following exclusion criteria: slanginess (e.g. *hot*, *screwy*, *snotty*); difficulty (e.g. *antagonistic*, *opportunistic*, *phlegmatic*); ambiguity (e.g. *cool*, *snappy*); link to sex, gender, demographics (e.g. *well educated*, *feminine*, *genderless*); over-evaluation (e.g. *awesome*, *sucky*, *crappy*); peripheral terms (e.g. *dry-witted*, *pseudo-friendly*); anatomical or physical characteristics (e.g. *bulky*, *small*, *beautiful*).

In addition, we also excluded all expressions which refer to the impression on the user rather than the agent’s personality (e.g. we include *bored* but exclude *boring*). Finally, in case of lexical opposites (e.g. *dishonest* and *honest*), we only include the positive form since negations have been shown to easily

¹www.nltk.org

be misunderstood [27]. All exclusion choices were discussed by two researchers and only applied in case of agreement.

Refining the Set via Synonym Clusters

The previous steps resulted in 592 descriptors. Since even comprehensive personality questionnaires usually comprise no more than 300 items for practical reasons [5, 20], we further reduced the set by removing synonyms. This is an established step in related work: Goldberg [30] and Norman [55] both used expert ratings to exclude redundant terms. In our case, we instead used a combination of automatic synonym clustering and manual analysis, as described next.

Specifying Synonyms

Clustering first required us to specify a list of synonyms for each descriptor. At first, we tried to do so using the lexical database *Word Net* [52]. However, the resulting list of synonyms for each word comprised several meanings of that word such that many synonym clusters were not meaningful (e.g. *practical, pragmatic, virtual*).

Therefore, we turned to the online Merriam Webster thesaurus², which provides word definitions and synonyms separately for all meanings of a word. We scraped this information. Two authors then manually went through all definitions to compile a list that only included those definitions of a word which are meaningful in the context of personality description. For example, for the descriptor *cold* we included the synonyms for the definition “*having or showing a lack of friendliness or interest in others*” but not for “*having a low or subnormal temperature*”³. In case a word had $n > 1$ valid definitions in this context, we added it n times, with indices (1... n). This allowed us to distinguish between the definitions after the clustering.

Clustering by Synonyms

We sorted the descriptors by word frequency in the English language with *wordfreq* [66]. This allowed us to favour frequently used and thus well-known descriptors over unfamiliar expressions. We used a greedy algorithm that clusters words based on mutual synonymy as follows: Iterating over the sorted list, at each word w , a new cluster c_w is created containing w . From the list of synonyms of w , all those words w' are added to c_w whose synonym lists also contain w (i.e. mutual synonymy). Finally, w' is removed from the sorted list.

Final Selection

We found 230 single descriptors (i.e. clusters of size 1) and 175 synonym clusters (size > 1). For each such cluster, we selected the most frequently used word with only one definition in our set. With this approach, we aimed to maximise clarity of the descriptors overall. For example, for the cluster *aggressive, ambitious, assertive, enterprising* we selected *assertive*; while *aggressive* and *ambitious* are used more frequently they also appeared with other meanings in our overall set.

We conducted a final manual review and discussion to reduce potentially remaining ambiguity and the number of descriptors with low frequency of use (e.g. which had been selected for

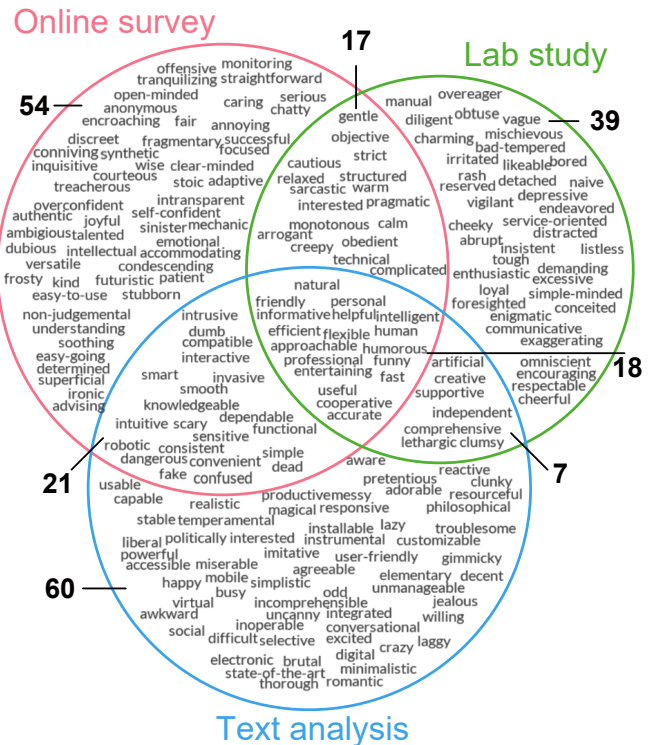


Figure 2. Overview of our final set of 349 descriptors, as collected with each method.

clusters with overall low frequency). In this way, we arrived at our final set of 349 adjectives (Figure 2).

EXPLORATORY FACTOR ANALYSIS

Research Design

We conducted an online survey to examine the structure of the comprehensive set of English trait adjectives we collected using the multi-study approach presented in the previous sections. The final descriptor set of 349 adjectives was administered to participants recruited via Amazon MTurk to get a large and diverse sample [7, 13].

After giving informed consent, participants were asked to indicate how often they had already interacted with each of the speech-based conversational agents Siri, Alexa, and Google Assistant. We used different agents to achieve a certain generalisability beyond a specific agent and limited ourselves to the three most common ones because they were most likely to be known by a large number of participants [57]. Inspired by the approach to assess personality by peer ratings [47], participants were asked to rate the speech-based conversational agent for which they had indicated the highest interaction frequency (minimum criterion was at least once). If they had interacted equally frequently with two or three of them, they were asked to select one of them for the following course of the questionnaire. Afterwards, participants indicated the extent to which each of the 349 adjectives presented in random order applied to the respective selected speech-based conversational agent on a four-point Likert scale ranging from “untypical” to “rather untypical” to “rather typical” to “typical”. We used an even response scale to avoid a “neither nor” category because

²www.merriam-webster.com

³www.merriam-webster.com/thesaurus/cold

	1	2	3	4	5	6	7	8	9
Confrontational (1)									
Dysfunctional (2)	.42								
Serviceable (3)	-.22	-.21							
Unstable (4)	.64	.28	-.19						
Approachable (5)	.08	-.01	.36	.08					
Social-Entertaining (6)	.24	.21	.12	.17	.41				
Social-Inclined (7)	-.21	-.05	.43	-.19	.42	.29			
Social-Assisting (8)	.27	.24	.14	.22	.45	.35	.22		
Self-Conscious (9)	.34	.31	.11	.25	.28	.42	.17	.32	
Artificial (10)	.21	.25	-.03	.19	.01	-.04	-.09	.19	.01

Table 2. Correlations between the ten factors from our factor analysis.

psychometric research has shown that this is often understood differently by participants and thus leads to problems [27]. Finally, participants provided demographic information.

On average, participants took 17 minutes for the online survey and were compensated with \$4.10 [34]. To ensure data quality, we selected MTurk workers only if they met certain requirements regarding their previous work results (80% accepted work performances), if they were located in the US and indicated to speak English at least well. Following related work [8] we additionally used multiple attention checks throughout the survey. We excluded participants' data for our final analysis if they had missed more than 25% of the attention tests or their survey time was less than 8 minutes, which we set as the minimum time for serious processing of our questionnaire according to our own preliminary tests.

Participants

N = 744 participants (45.7% female, 53.5% male, 0.8% said other or preferred not to say) completed the survey and fulfilled our attention/time requirements described above. The mean age was 36.7 (range 19-72 years). 28.6% indicated to have graduated high school or have a diploma, 17.9% held an associated degree, and 42.2% had a bachelor's degree. The remaining 11.3% had lower or higher educational degrees. 26.1% of the participants rated Siri, 32.9% rated Alexa, and 41.0% rated Google Assistant.

Data Analysis

We investigated the descriptor set's underlying structure with an exploratory factor analysis based on the correlation matrix of all 349 descriptors. To determine the appropriate number of factors we used the empirical Kaiser criterion which has been found to perform well in research settings such as ours [10]. We used the R package *psych* [63] (R version 3.6.1).

The empirical Kaiser criterion proposed a ten-factorial solution. We used an obliquely (oblimin) rotated principle axis analysis for factor extraction. The resulting ten factors accounted for 49% of the variance. Table 2 shows their correlations. Table 3 lists the factors with top 20 descriptors by factor loadings and our interpreted factor labels. The complete factor matrix can be found on the project website (cf. Conclusion Section). Loadings ranged between -0.47 and 0.71 across all factors. Using a loading value of 0.30 and high secondary loadings (difference < 0.20) as criteria, 86 of the 349 items could not (uniquely) be assigned to one of the ten factors.

We next describe each factor as a personality dimension for speech-based conversational agents. We do not claim that ours is the only possible interpretation; readers are invited to develop their own understanding (e.g. via the top descriptors in Table 3). To foster interpretation and discussion, we also sketch ideas and scenarios in which we would expect an agent to score highly on each dimension.

Confrontational

This dimension is described by negative terms that put the agent into an *actively negative* stance, such as abusive, combative, offensive, stingy, encroaching, manipulative, explosive, or vindictive. Overall, we thus interpret this dimension as capturing a *confrontational* aspect.

For example, a voice assistant scoring high on this dimension might not always readily agree with the user or perform tasks according to the user's wishes. In addition, its feedback might not strike a friendly tone in such situations.

Dysfunctional

This dimension is also described by negative terms, yet puts the agent into a *passive negative* stance. The words signal confusion and inactivity, such as lazy, irritated, fearful, unmanageable, ignorant, dead, inoperable, or forgetful. This inability to function properly is both present on a more emotional/social level (e.g. fearful, listless, conceited, stubborn, crazy) as well as on a practical/functional one (e.g. inoperable, unmanageable, forgetful, dead). Since both these levels are meanings of the word *dysfunctional*⁴ we chose this as a label here.

For example, an agent scoring high on this dimension might not react to user input or not give (enough) feedback. It seems likely that voice agents are perceived as highly dysfunctional if their functionality as an assistant is severely hindered, for instance, by software bugs (e.g. resetting mid conversation) or hardware issues (e.g. broken microphone, loss of power).

Serviceable

This dimension is described by positive terms, which mostly relate to cognitive functioning, such as informative, functional, capable, accurate, knowledgeable, or thorough. In addition, this dimension also considers adequate communication and role-fulfillment as an assistant, such as convenient, responsive, useful, user-friendly, interactive, communicative, productive, and helpful. Overall, we thus interpret this dimension as capturing functionality and usability of an assistant, which we summarise as *serviceable*.

For example, an assistant scoring high here is likely to react promptly, provides adequate feedback, and performs tasks in a helpful and reliable way. It seems likely that most creators of voice assistants want to present their product as highly serviceable, for example, in advertisements.

Unstable

This dimension is described by negative terms, which overall signal aspects of *instability*, rather than active confrontation or passive inability to function. For example, this includes terms such as nervous, anxious, or temperamental. Besides

⁴www.merriam-webster.com/dictionary/dysfunctional

Factor / Label	Top 20 descriptors (by factor loadings)
1 Confrontational	abusive (.71), negligent (.70), deceitful (.69), cruel (.68), distrustful (.67), combative (.65), offensive (.65), incomprehensible (.64), stingy (.62), messy (.60), encroaching (.59), scornful (.58), dubious (.58), irritable (.58), manipulative (.58), explosive (.56), clumsy (.55), condescending (.54), clunky (.54), vindictive (.53)
2 Dysfunctional	reckless (.71), lazy (.71), irritated (.70), fearful (.69), bitter (.65), moody (.65), crazy (.63), unmanageable (.62), prejudiced (.61), treacherous (.61), ignorant (.59), conceited (.59), dead (.58), inoperable (.57), forgetful (.57), bad-tempered (.55), impetuous (.55), listless (.52), stubborn (.51), egocentric (.50)
3 Serviceable	informative (.56), functional (.56), capable (.56), accurate (.55), knowledgeable (.55), convenient (.54), thorough (.52), responsive (.52), productive (.52), consistent (.50), useful (.50), user-friendly (.50), interactive (.50), adaptive (.50), helpful (.50), communicative (.49), organized (.49), dependable (.48), intelligent (.47), comprehensive (.47)
4 Unstable	nervous (.56), depressive (.54), rude (.54), bigoted (.52), grumpy (.51), jealous (.49), forceful (.49), anxious (.48), gruff (.47), easy-to-use (-.47), frosty (.47), fretful (.45), tempestuous (.45), dangerous (.45), sloppy (.45), faultfinding (.45), dumb (.45), troublesome (.43), rambunctious (.41), temperamental (.41)
5 Approachable	peaceful (.64), easy-going (.62), gentle (.61), relaxed (.61), fair (.57), clear-minded (.56), respectful (.53), calm (.53), humble (.51), respectable (.48), casual (.48), courteous (.47), sincere (.46), understanding (.44), ethical (.44), loyal (.43), open-minded (.43), principled (.42), simplistic (.42), determined (.39)
6 Social-Entertaining	humorous (.69), playful (.57), funny (.66), joyful (.52), charming (.51), entertaining (.51), cheerful (.50), merry (.49), happy-go-lucky (.49), cheeky (.48), expressive (.45), chatty (.45), affectionate (.45), happy (.45), excited (.44), enthusiastic (.44), social (.42), adorable (.41), warm (.39), encouraging (.38)
7 Social-Inclined	agreeable (.51), willing (.46), likeable (.43), kind (.42), trustful (.41), decent (.40), modest (.40), flexible (.40), interested (.38), realistic (.38), conversational (.38), friendly (.37), self-disciplined (.37), inquisitive (.35), patient (.34), soothing (.32), endeavored (.32)
8 Social-Assisting	pragmatic (.61), fastidious (.58), scrupulous (.57), genial (.54), diplomatic (.46), omniscient (.45), vigilant (.44), vigorous (.44), benevolent (.43), dignified (.43), amiable (.41), stoic (.38), conscientious (.36), discreet (.36), deliberate (.35), meticulous (.33), lenient (.33), zestful (.32), foresighted (.32), restrained (.31)
9 Self-Conscious	independent (.45), assertive (.45), competitive (.43), brave (.42), creative (.42), deep (.41), selective (.41), proud (.39), excitable (.39), artistic (.37), ambitious (.37), introspective (.36), powerful (.35), individualistic (.35), self-indulgent (.35), insistent (.35), extravagant (.34), crafty (.34), contemplative (.34), daring (.34)
10 Artificial	synthetic (.50), robotic (.49), intrusive (.48), artificial (.47), odd (.45), invasive (.44), gimmicky (.44), abrupt (.43), superficial (.43), monotonous (.43), fake (.42), simple-minded (.41), mechanic (.41), vague (.41), passionless (.40), digital (.39), electronic (.39), monitoring (.39), annoying (.36), detached (.35)

Table 3. Overview of the factors obtained in our exploratory factor analysis, with top 20 descriptors (and their factor loadings). The factor labels suggested here are based on the authors' interpretation of all descriptors per factor with loadings of at least .3.

a level of emotional/social instability present in these terms, there are others that hint more at unstable functionality in this context, such as sloppy, faultfinding, dangerous, and negative (i.e. absence of) ease-of-use.

An example for scoring high on this dimension might be a prototype that does not always work as users expect, with inconsistency being the main negative aspect.

Approachable

This encompasses positive terms which cast the assistant as calm and welcoming, with words such as peaceful, easy-going, gentle, relaxed, open-minded and understanding. It also hints at an assistant that treats requests well – fair, clear-minded, respectful, sincere, ethical, loyal, principled, and determined. In contrast to the positive terms of *Serviceable*, these cover not so much the utility of fulfilling tasks as rather the positive social experience expected in asking the assistant to do so. We thus summarise such an assistant character as *approachable*.

A voice assistant scoring high here is likely to navigate well through conversations and gives appropriate feedback that

strikes a socially adequate tone, independent of whether the user's request can be practically fulfilled or not.

Social-Entertaining

This dimension captures humour in the light of positive social behaviour and entertainment, with terms such as humorous, playful, funny, joyful, charming, entertaining, cheerful, happy, and encouraging.

Scoring high on this dimension likely means that a voice assistant can communicate in a humorous and entertaining way, or includes dedicated functionality for that (e.g. can tell jokes and stories, play games, etc.).

Social-Inclined

In this dimension we find an agent's characteristic of being inclined to assist its users, with positive terms such as agreeable, willing, interested, endeavored, flexible, conversational, inquisitive, and patient. This is accompanied by terms that signal a positive tone when communicating this, such as friendly, kind, decent, modest, and soothing.

For instance, a speech-based conversational agent scoring high on this dimension is overall friendly and might actively signal readiness (e.g. with a hardware light) or ask users if they would like a more detailed response or if they have further requests.

Social-Assisting

This dimension captures social skills and attitudes that can be expected from the role of a skilled assistant: It has terms such as pragmatic, conscientious, diplomatic, vigilant, foresighted, amiable, and discreet. It also includes terms that signal accurate task fulfilment such as meticulous, scrupulous and deliberate. In contrast to other dimensions, these relate more to the attitude with which an assistant executes its tasks, rather than its usable functioning (cf. *Serviceable*) or its welcoming character (cf. *Approachable*).

A voice assistant scoring high here likely handles requests well while otherwise staying in the background. It clearly communicates that its role is to serve the user and might also anticipate users' wishes and adequate (re)actions.

Self-Conscious

This dimension encompasses terms that render a voice assistant as an entity capable of independent thought: For example, these terms include independent, competitive, creative, artistic, deep, proud, ambitious, introspective, and contemplative. While actual artificial self-consciousness might still be in the realm of science-fiction for a long time, a voice assistant might create such an illusion in some contexts. This dimension also contains "magical" (albeit not in the top 20), which supports such an interpretation.

Designing a speech-based conversational agent to score high here likely requires implementing convincing responses in conversation about opinions, impressions or feelings: In such more abstract conversations the agent might then have room to show shades of (seemingly) independent and creative thought.

Artificial

In this dimension we find terms that emphasise artificiality or "thingness": For instance, this covers words such as synthetic, robotic, artificial, gimmicky, superficial, fake, electronic, and mechanic. Other terms here hint at technological implications of bringing such a "thing" into human social contexts, such as intrusive, odd, abrupt, simple-minded, passionless, monitoring, annoying, and detached.

A conversational agent might score high on this dimension if it clearly presents itself as an object – either by communicating this intentionally (e.g. to avoid overtrust) or via issues that break the illusion of an actual being behind the voice.

DISCUSSION

Reflection on Personality Dimensions

Our multi-method collection of descriptors resulted in ten personality dimensions for conversational agents, derived via exploratory factor analysis on ratings of our descriptors by 744 people (see Table 3). Reflecting on these dimensions, it is interesting to note that the majority of dimensions indicates either desirable or non-desirable characteristics. This suggests that the agent's ability to fulfill users' expectations of natural conversations is of crucial importance for users' perception.

Comparing our ten dimensions to the Big Five model for human personality, we find that particularly adjectives from the dimension agreeableness can be found in several of our dimensions, such as *Approachable*, *Social-Inclined*, and *Social-Assisting*. Since we focused our work on voice assistants, agreeableness seems to play a key role distributed over several dimensions.

The dimension *Unstable* might be associated with the Big Five dimension neuroticism, which is also called emotional stability. Interestingly, this dimension does not only comprise human characteristics of instability, e.g. *nervous* or *depressive* but also technical characteristics such as *inoperable* or absence of *easy-to-use*. Similarly, the dimension *serviceable* encompasses human characteristics, e.g. *knowledgeable*, *thorough*, which correspond with the Big Five dimension *conscientiousness* – yet also technical terms such as *useful* or *responsive*.

This contrast of functional and social descriptors appears to be a pattern which can be found within the majority of dimensions: For example, also the dimension *Dysfunctional* includes descriptors on a social and emotional level (*reckless*, *moody*, *crazy*) combined with others that describe the technical functionality or role of an assistant (e.g., *inoperable*, *forgetful*). The dichotomy of functional and social has also been observed in previous work on how users describe their everyday conversations with voice assistants [19, 58].

Looking ahead, a different structure might emerge with a greater variety of speech-based conversational agents: For example, while current voice agents predominantly have roles as assistants (which were the subject of our investigation), future agents might fulfill other roles which in turn may impact on their perceived personality and its description. Moreover, the interaction with conversational agents does not really resemble human conversation at the current technological stage, as is underlined by the findings from the interaction experiment (frequent use of descriptors such as *inhuman*, *impersonal*, or *unpleasant*). However, it is likely that with technological improvements and a more natural conversation in the near future, users might perceive agent personality differently.

Finally, two dimensions emerged that appear as independent of an assistant role: Both *Self-Conscious* and *Artificial* seem to describe a speech-based conversational agent's similarity to humans. On the one hand, self-consciousness represents a dimension which current conversational agents cannot technically fulfil, yet the *impression* of for example *creativity* or *independence* can impact on the perceived agent personality. In contrast, Nass and Brave [53] discussed whether speech-based conversational agents *should* be similar to humans. Hence, an agent could appear as self-conscious in its conversation content (e.g. offering opinions) but still highly *artificial*, e.g. by using a clearly synthesised voice, to inform the user that s/he is communicating with a machine. Doyle et al. [25] emphasised that users conceptualise voice assistants' humanness in a multidimensional way. According to their analysis, low interpersonal connection and poor vocal qualities can result in perceiving an agent as artificial, synthetic or robotic but also several of our other dimensions, e.g. humour or kind of knowledge, contribute to the overall perception of humanness.

Reflection on Methodology and Implications for Research

The descriptor sets collected in our three methods show only small overlaps (see Figure 2): Only 18 descriptors in the final set were named in all three methods. On the one hand, this could indicate that more data is necessary to derive a more generalisable and robust set. Hence, other approaches or replications are useful and needed, e.g. to evaluate if descriptors “saturate”. On the other hand, our results show that a combination of multiple methods is beneficial to cover a comprehensive variety of descriptors in different use cases.

We do not regard our resulting set as a “solution” but rather as a starting point for future work: For comparison, many researchers have been involved in the collection process for human personality indicators over decades [21, 38]. A large proportion of the variance of people’s answers (51%) was not explained by our dimensions. This error variance is comparable to results known from human big five models [28] and should be addressed in future work. We would also like to point out that the formalisation of a measurement model for agent personality goes beyond the scope of this work. We outline key opportunities for such future work here:

Users’ purpose to interact with speech-based conversational agents (task vs social conversation use) may influence perception of personality. Since personality is defined as a stable construct across contexts [48], we started out with a general case to best reflect this definition of personality. In the lab we presented different use cases (task and social). Overall, by combining different methods and including implicit user data (reviews), we collected data from a variety of use cases. Future work should address specific and further use cases.

We focused on speech-based conversational agents. Future work could investigate whether our descriptors can also adequately describe the personality of other agents (e.g., robots, chatbots). To investigate goodness of fit, future work could conduct a *confirmatory* factor analysis using our descriptors for rating these and other agents. In addition, 25% of our descriptors could not be clearly assigned to one of dimensions. These items in particular deserve attention in future work.

We also observed correlations between our dimensions: For example, *Confrontational*, *Dysfunctional*, and *Unstable* are considerably correlated. These all seem to describe negative aspects of the conversational agent’s personality. Similarly, we found a group of dimensions with positive connotations (e.g., *Social-Entertaining*, *Social-Inclined*, *Social-Assisting*). This could indicate that not only personality traits per se play an important role in assessing the personality of agents, but also their connotation and functional or experience-based evaluation in the user’s view. We encourage future research to investigate these aspects separately and take them into account when modelling. For example, it could be investigated if a hierarchical structure of dimensions can be found.

We used English terms (and in the lab study terms translated from German to English by two researchers). Future work should investigate other languages and cultural backgrounds, since the approach is language-based and different cultures likely perceive agents’ personalities differently [58].

Implications for Practitioners

An important next step for HCI practice is to derive solution principles for effectively implementing agent personality. Practitioners usually have specific characteristics in mind when designing agent personality. Our descriptors can be used as a *communication tool* to make these characteristics explicit and to discuss the desired personality of a new agent in a systematic way. This seems particularly interesting for collaborations, when multiple conversation designers write dialogues individually, to facilitate consistency and a mutual understanding of an agent personality [39].

Related work on the similarity attraction paradigm [14] proposed to adapt agent personality to the user [4, 11, 41]. Agent personality might also be designed with regard to user groups and application context. The found dimensions support such tasks since they make explicit 1) which aspects of personality *can* be varied (e.g. to achieve a goal such as “neutral, pragmatic helper”), yet also 2) which ones *have to* be considered as well (e.g. *Dysfunctional* highlights considering that personality is also present in how an assistant communicates failures).

CONCLUSION

We presented the first systematic analysis of personality descriptors and dimensions for speech-based conversational agents, following the established psycholexical approach from psychology. Our main contribution is a set of 349 agent personality descriptors, grouped into ten personality dimensions, which serve as an initial step to developing a personality model for speech-based conversational agents.

As a broader implication, the revealed dimensions do not match the Big Five model. Our descriptors also include terms not associated with human personality. This systematically consolidates evidence from related work about people describing agent personality differently [70]. Our findings thus indicate that the human Big Five model is not directly applicable to speech-based conversational agents. Instead, the found personality dimensions also capture, for example, how artificial, self-conscious, or serviceable the agent appears to its users.

Practically, the found dimensions and descriptors support research and applications in systematically designing personality of speech-based conversational agents. Conceptually, we set foundations for future work: As in psychology, personality models should be re-validated in further studies. Future work could also investigate models for other virtual agents (e.g. chatbots). Here, our descriptors, dimensions, and methodology may serve as a useful starting point.

To support such future research and applications, our project website hosts the lists of adjectives from the studies, the final descriptor set, and further material from the factor analysis:

www.medien.ifi.lmu.de/personality-model

ACKNOWLEDGEMENTS

This project is funded by the Bavarian State Ministry of Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B).

REFERENCES

- [1] Gordon W. Allport and Henry S. Odbert. 1936. Trait-names: A psycho-lexical study. *Psychological Monographs* 47, 1 (1936), i–171. DOI: <http://dx.doi.org/10.1037/h0093360>
- [2] Norman H. Anderson. 1968. Likableness ratings of 555 personality-trait words. *Journal of Personality and Social Psychology* 9, 3 (1968), 272–279. DOI: <http://dx.doi.org/10.1037/h0025907>
- [3] Elisabeth André, Martin Klesen, Patrick Gebhard, Steve Allen, and Thomas Rist. 2000. Integrating models of personality and emotions into lifelike characters. In *Affective interactions IWAI 1999. Lecture Notes in Computer Science*, A. Paiva (Ed.). Vol. 1814. Springer, Berlin, Heidelberg, 150–165. DOI: http://dx.doi.org/10.1007/10720296_11
- [4] Sean Andrist, Bilge Mutlu, and Adriana Tapus. 2015. Look Like Me: Matching Robot Personality via Gaze to Increase Motivation. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 3603–3612. DOI: <http://dx.doi.org/10.1145/2702123.2702592>
- [5] M Arendasy. 2009. BFSI: Big-Five Struktur-Inventar (Test & Manual). (2009). Mödling: SCHUHFRIED GmbH.
- [6] Christoph Bartneck and Jun Hu. 2008. Exploring the abuse of robots. *Interaction Studies* 9, 3 (2008), 415–433. DOI: <http://dx.doi.org/10.1075/is.9.3.04bar>
- [7] Tara S. Behrend, David J. Sharek, Adam W. Meade, and Eric N. Wiebe. 2011. The viability of crowdsourcing for survey research. *Behavior Research Methods* 43, 3 (2011), 800–813. DOI: <http://dx.doi.org/10.3758/s13428-011-0081-0>
- [8] Adam J. Berinsky, Michele F. Margolis, and Michael W. Sances. 2014. Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science* 58, 3 (2014), 739–753. DOI: <http://dx.doi.org/10.1111/ajps.12081>
- [9] François Bouchet and Jean-Paul Sansonnet. 2012. Intelligent agents with personality: From adjectives to behavioral schemes. In *Cognitively Informed Intelligent Interfaces: Systems Design and Development*. IGI Global, Hershey, PA, USA, 177–200. DOI: <http://dx.doi.org/10.4018/978-1-4666-1628-8.ch011>
- [10] Johan Braeken and Marcel A. L. M. van Assen. 2017. An empirical Kaiser criterion. *Psychological Methods* 22, 3 (2017), 450–466. DOI: <http://dx.doi.org/10.1037/met0000074>
- [11] Michael Braun, Anja Mainz, Ronee Chadowitz, Bastian Pflöging, 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 (CHI '19)*. ACM, New York, NY, USA, Article 475, 12 pages. DOI: <http://dx.doi.org/10.1145/3290605.3300705>
- [12] Markus Bühner. 2011. *Einführung in die Test- und Fragebogenkonstruktion*. Pearson Education, Munich, Germany.
- [13] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. 2011. Amazon’s Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science* 6, 1 (2011), 3–5. DOI: <http://dx.doi.org/10.1177/1745691610393980>
- [14] Donn Byrne. 1961. Interpersonal attraction and attitude similarity. *The Journal of Abnormal and Social Psychology* 62, 3 (1961), 713–715. DOI: <http://dx.doi.org/10.1037/h0044721>
- [15] 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. DOI: <http://dx.doi.org/10.1145/2940325>
- [16] Angelo Cafaro, Hannes Högni Vilhjálmsson, Timothy Bickmore, Dirk Heylen, Kamilla Rún Jóhannsdóttir, and Gunnar Steinn Valgardsson. 2012. First Impressions: Users’ Judgments of Virtual Agents’ Personality and Interpersonal Attitude in First Encounters. In *Intelligent Virtual Agents. IVA 2012. Lecture Notes in Computer Science*, Yukiko Nakano, Michael Neff, Ana Paiva, and Marilyn Walker (Eds.), Vol. 7502. Springer, Berlin, Heidelberg, 67–80. DOI: http://dx.doi.org/10.1007/978-3-642-33197-8_7
- [17] Raymond B. Cattell. 1947. Confirmation and clarification of primary personality factors. *Psychometrika* 12, 3 (1947), 197–220. DOI: <http://dx.doi.org/10.1007/BF02289253>
- [18] Cindy K. Chung and James W. Pennebaker. 2008. Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of Research in Personality* 42, 1 (2008), 96–132. DOI: <http://dx.doi.org/10.1016/j.jrp.2007.04.006>
- [19] 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 (CHI '19)*. ACM, New York, NY, USA, Article 475, 12 pages. DOI: <http://dx.doi.org/10.1145/3290605.3300705>
- [20] Paul T. Costa and Robert R. McCrae. 1992. *Neo Personality Inventory – Revised (NEO PI-R)*. Psychological Assessment Resources, Odessa, FL, USA.

- [21] Boele De Raad. 2000. *The Big Five Personality Factors: The psycholexical approach to personality*. Hogrefe & Huber Publishers, Gottingen, Germany.
- [22] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhomme, and et al. 2014. SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS '14)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1061–1068.
- [23] 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. DOI: <http://dx.doi.org/10.1037/14343-017>
- [24] 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. DOI: [http://dx.doi.org/10.1016/0092-6566\(92\)90039-7](http://dx.doi.org/10.1016/0092-6566(92)90039-7)
- [25] 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 (MobileHCI '19)*. Association for Computing Machinery, New York, NY, USA, Article 5, 12 pages. DOI: <http://dx.doi.org/10.1145/3338286.3340116>
- [26] Malin Eiband, Sarah Theres Völkel, Daniel Buschek, Sophia Cook, and Heinrich Hussmann. 2019. When People and Algorithms Meet: User-reported Problems in Intelligent Everyday Applications. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. ACM, New York, NY, USA, 96–106. DOI: <http://dx.doi.org/10.1145/3301275.3302262>
- [27] Michael Eid and Katharina Schmidt. 2014. *Testtheorie und Testkonstruktion*. Hogrefe Verlag, Gottingen, Germany.
- [28] Timo Gnambs. 2015. Facets of measurement error for scores of the Big Five: Three reliability generalizations. *Personality and Individual Differences* 84 (2015), 84 – 89. DOI: <http://dx.doi.org/https://doi.org/10.1016/j.paid.2014.08.019> Theory and Measurement in Personality and Individual Differences.
- [29] 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.
- [30] Lewis R. Goldberg. 1982. From Ace to Zombie: Some explorations in the language of personality. In *Advances in Personality Assessment*, Charles D. Spielberger and James N. Butcher (Eds.). Vol. 1. Erlbaum, Hillsdale, NJ, USA, 203–234.
- [31] Lewis R. Goldberg. 1990. An alternative “description of personality”: The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59, 6 (1990), 1216–1229. DOI: <http://dx.doi.org/10.1037/0022-3514.59.6.1216>
- [32] Harrison G. Gough and Alfred B. Heilbrun. 1980. *The Adjective Checklist Manual: 1980 Edition*. Consulting Psychologists Press, Palo Alto, CA, USA.
- [33] E. Guzman and W. Maalej. 2014. How Do Users Like This Feature? A Fine Grained Sentiment Analysis of App Reviews. In *2014 IEEE 22nd International Requirements Engineering Conference (RE)*. IEEE, New York, NY, USA, 153–162. DOI: <http://dx.doi.org/10.1109/RE.2014.6912257>
- [34] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. 2018. A Data-Driven Analysis of Workers’ Earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article Paper 449, 14 pages. DOI: <http://dx.doi.org/10.1145/3173574.3174023>
- [35] Katherine Isbister and Clifford Nass. 2000. Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies* 53, 2 (2000), 251 – 267. DOI: <http://dx.doi.org/10.1006/ijhc.2000.0368>
- [36] 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. DOI: <http://dx.doi.org/10.1016/j.jrp.2010.06.005>
- [37] Lauri A. Jensen-Campbell and William G. Graziano. 2001. Agreeableness as a moderator of interpersonal conflict. *Journal of Personality* 69, 2 (2001), 323–362. DOI: <http://dx.doi.org/10.1111/1467-6494.00148>
- [38] Oliver P. John, Alois Angleitner, and Fritz Ostendorf. 1988. The lexical approach to personality: A historical review of trait taxonomic research. *European Journal of Personality* 2, 3 (1988), 171–203. DOI: <http://dx.doi.org/10.1002/per.2410020302>
- [39] Hankyung Kim, Dong Yoona Koh, Gaeunb Lee, Jung-Mi Park, and Youn-kyung Lim. 2019. Developing a Design Guide for Consistent Manifestation of Conversational Agent Personalities. In *IASDR Conference*. Manchester Metropolitan University, Manchester, UK, 1–17.

- [40] B. Kinsella and A. Mutchler. 2018. U.S. Smart Speaker Consumer Adoption Report 2019. (2018). Retrieved on September 12, 2019 from <https://voicebot.ai/smart-speaker-consumer-adoption-report-2019/>.
- [41] Brigitte Krenn, Birgit Endrass, Felix Kistler, and Elisabeth André. 2014. Effects of Language Variety on Personality Perception in Embodied Conversational Agents. In *Human-Computer Interaction. Advanced Interaction Modalities and Techniques. HCI 2014. Lecture Notes in Computer Science*, Masaaki Kurosu (Ed.), Vol. 8511. Springer International Publishing, Cham, 429–439. DOI: http://dx.doi.org/10.1007/978-3-319-07230-2_41
- [42] Vivek Kulkarni, Margaret L. Kern, David Stillwell, Michal Kosinski, Sandra Matz, Lyle Ungar, Steven Skiena, and H. Andrew Schwartz. 2018. Latent human traits in the language of social media: An open-vocabulary approach. *PLOS ONE* 13, 11 (2018), e0201703. DOI: <http://dx.doi.org/10.1371/journal.pone.0201703>
- [43] Walid Maalej, Zijad Kurtanović, Hadeer Nabil, and Christoph Stanik. 2016. On the Automatic Classification of App Reviews. *Requirements Engineering* 21, 3 (2016), 311–331. DOI: <http://dx.doi.org/10.1007/s00766-016-0251-9>
- [44] Gerald Matthews, Ian J. Deary, and Martha C. Whiteman. 2003. *Personality Traits*. Cambridge University Press, Cambridge, UK. 602 pages. DOI: <http://dx.doi.org/10.1017/CB09780511812736>
- [45] Robert R. McCrae. 2009. The Five-Factor Model of personality traits: consensus and controversy. In *The Cambridge Handbook of Personality Psychology*, Philip J. Corr and Gerald Matthews (Eds.). Cambridge University Press, Cambridge, UK, 148–161. DOI: <http://dx.doi.org/10.1017/CB09780511596544.012>
- [46] Robert R. McCrae and Paul T. Costa. 1985. Updating Norman’s adequacy taxonomy: Intelligence and personality dimensions in natural language and in questionnaires. *Journal of Personality and Social Psychology* 49, 3 (1985), 710–721. DOI: <http://dx.doi.org/10.1037//0022-3514.49.3.710>
- [47] Robert R. McCrae and Paul T. Costa. 1987. Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology* 52, 1 (1987), 81–90. DOI: <http://dx.doi.org/10.1037/0022-3514.52.1.81>
- [48] 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.
- [49] 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. DOI: <http://dx.doi.org/10.1111/j.1467-6494.1992.tb00970.x>
- [50] 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. DOI: <http://dx.doi.org/10.1016/j.jrp.2005.05.003>
- [51] Margaret McRorie, Ian Sneddon, Gary McKeown, Elisabetta Bevacqua, Etienne de Sevin, and Catherine Pelachaud. 2012. Evaluation of Four Designed Virtual Agent Personalities. *IEEE Transactions on Affective Computing* 3, 3 (2012), 311–322. DOI: <http://dx.doi.org/10.1109/T-AFFC.2011.38>
- [52] George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (1995), 39–41. DOI: <http://dx.doi.org/10.1145/219717.219748>
- [53] Clifford Ivar Nass and Scott Brave. 2005. *Wired for speech: How voice activates and advances the human-computer relationship*. MIT press, Cambridge, MA, USA.
- [54] Michael Neff, Nicholas Toothman, Robeson Bowmani, Jean E. Fox Tree, and Marilyn A. Walker. 2011. Don’t Scratch! Self-adaptors Reflect Emotional Stability. In *Intelligent Virtual Agents. IVA 2011. Lecture Notes in Computer Science*, Hannes Högni Vilhjálmsson, Stefan Kopp, Stacy Marsella, and Kristinn R. Thórisson (Eds.), Vol. 6895. Springer, Berlin, Heidelberg, 398–411. DOI: http://dx.doi.org/10.1007/978-3-642-23974-8_43
- [55] Warren T. Norman. 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology* 66, 6 (1963), 574–583. DOI: <http://dx.doi.org/10.1037/h0040291>
- [56] Warren T. Norman. 1967. *2800 Personality trait descriptors – Normative operating characteristics for a university population*. Technical Report ED 014 738. Department of Psychology, University of Michigan, Ann Arbor, MI, USA.
- [57] Christi Olsen and Kelli Kemery. 2019. Voice report: From answers to action: customer adoption of voice technology and digital assistants. (2019). Retrieved on September 1, 2019 from https://advertiseonbing-blob.azureedge.net/blob/bingads/media/insight/whitepapers/2019/04%20apr/voice-report/bingads_2019_voicereport.pdf.
- [58] Marta Garcia Perez, Sarita Saffon Lopez, and Hector Donis. 2018. Voice Activated Virtual Assistants Personality Perceptions and Desires: Comparing Personality Evaluation Frameworks. In *Proceedings of the 32nd International BCS Human Computer Interaction Conference (HCI '18)*. BCS Learning & Development Ltd., Swindon, GBR, Article Article 40, 10 pages. DOI: <http://dx.doi.org/10.14236/ewic/HCI2018.40>

- [59] Laura Pfeifer Vardoulakis, Lazlo Ring, Barbara Barry, Candace L. Sidner, and Timothy Bickmore. 2012. Designing Relational Agents as Long Term Social Companions for Older Adults. In *Intelligent Virtual Agents. IVA 2012. Lecture Notes in Computer Science.*, Yukiko Nakano, Michael Neff, Ana Paiva, and Marilyn Walker (Eds.), Vol. 7502. Springer, Berlin, Heidelberg, 289–302. DOI : http://dx.doi.org/10.1007/978-3-642-33197-8_30
- [60] 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 (CHI '18)*. ACM, New York, NY, USA, Article 640, 12 pages. DOI : <http://dx.doi.org/10.1145/3173574.3174214>
- [61] Anand S. Rao and Michael P. Georgeff. 1995. BDI Agents: From Theory to Practice.. In *Proceedings of the First International Conference on Multiagent Systems (ICMAS)*. AAAI, San Francisco, CA, USA, 312–319.
- [62] 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.
- [63] William R. Revelle. 2018. psych: Procedures for Psychological, Psychometric, and Personality Research. (2018). <https://CRAN.R-project.org/package=psych> R package version 1.8.12.
- [64] Paola Rizzo, Manuela Veloso, Maria Miceli, and Amedeo Cesta. 1997. Personality-driven social behaviors in believable agents. In *Proceedings of the AAAI Fall Symposium on Socially Intelligent Agents*. AAAI, Palo Alto, CA, USA, 109–114.
- [65] Klaus Rainer Scherer. 1979. Personality markers in speech. In *Social Markers in Speech*, Klaus Rainer Scherer and Howard Giles (Eds.). Cambridge University Press, Cambridge, UK.
- [66] Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. 2018. LuminosoInsight/wordfreq: v2.2. (Oct. 2018). DOI : <http://dx.doi.org/10.5281/zenodo.1443582>
- [67] Adriana Tapus and Maja J Mataric. 2008. Socially Assistive Robots: The Link between Personality, Empathy, Physiological Signals, and Task Performance.. In *AAAI Spring Symposium: Emotion, personality, and social behavior*. AAAI, Palo Alto, CA, USA, 133–140.
- [68] James Vlahos. 2019. *Talk to Me: How Voice Computing Will Transform the Way We Live, Work, and Think*. Houghton Mifflin Harcourt, Boston, MA, USA.
- [69] Chris Welch. 2018. Google just gave a stunning demo of Assistant making an actual phone call. (2018). Retrieved on September 19, 2019 from <https://www.theverge.com/2018/5/8/17332070/google-assistant-makes-phone-call-demo-duplex-io-2018>.
- [70] 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. DOI : <http://dx.doi.org/10.1145/3232077>
- [71] Matthias Ziegler and Markus Buehner. 2009. Modeling Socially Desirable Responding and Its Effects. *Educational and Psychological Measurement* 69, 4 (2009), 548–565. DOI : <http://dx.doi.org/10.1177/0013164408324469>