Keep it Short: A Comparison of Voice Assistants' Response Behavior

Gabriel Haas
gabriel.haas@uni-ulm.de
University of Ulm
Ulm, Germany

Michael Rietzler
michael.rietzler@uni-ulm.de
University of Ulm
Ulm, Germany

Matt Jones
matt.jones@swansea.ac.uk
Swansea University
Wales, United Kingdom

Enrico Rukzio
enrico.rukzio@uni-ulm.de
University of Ulm
Ulm, Germany

ABSTRACT

Voice assistants (VAs) are present in homes, smartphones, and cars. They allow users to perform tasks without graphical or tactile user interfaces, as they are designed for natural language interaction. However, we found that currently, VAs are emulating human behavior by responding in complete sentences, limiting the design options, and preventing VAs from meeting their full potential as a utilitarian tool. We implemented a VA that handles requests in three response styles: two differing short keyword-based response styles and a full-sentence baseline. In a user study, 72 participants interacted with our VA by issuing eight requests. Results show that the short responses were perceived similarly useful and likable while being perceived as more efficient, especially for commands, and sometimes better to comprehend than the baseline. To achieve widespread adoption, we argue that VAs should be customizable and adapt to users instead of always responding in full sentences.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in HCI*; *Empirical studies in interaction design.*

KEYWORDS

voice user interface, voice assistant, virtual assistant

ACM Reference Format:
Gabriel Haas, Michael Rietzler, Matt Jones, and Enrico Rukzio. 2022. Keep it Short: A Comparison of Voice Assistants' Response Behavior. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3491102.3517684

1 INTRODUCTION

Voice assistants nowadays can be found on all sorts of digital devices. They are available as stand-alone devices such as the popular Echo devices made by Amazon, Google's Home series speakers, or other so-called smart speakers. Furthermore, VAs are integrated with many devices as a regular feature, e.g., in modern smartphones, personal computers, or cars. Thus, they allow performing simple tasks of daily life using the most natural form of human communication: spoken language. Thereby, they eliminate the need for interaction with a conventional user interface - such as mouse, keyboard, and (touch)screens. Voice interactions can be beneficial for micro-interactions with few cycles of input and response [2]. In contrast to the use of smartphones and desktop computers, there is virtually no access time. For example, voice commands can be used to control lights in a smart home or request information such as the time or daily weather without picking up a remote or smartphone.

All this progress is driven by artificial intelligence, specifically natural language processing and voice synthesis allowing the creation of sophisticated voice user interfaces that can recognize and interpret users' words and provide an appropriate response. We analyzed the responses of the most popular VAs (Amazon's Alexa, Google's Assistant, Apple's Siri) and found that they almost exclusively respond in full sentences and refer to the assistant with the personal pronoun 'I'. In most cases, those responses are *designed* to be human-like, ultimately aiming to be indistinguishable from a real, human assistant.

Contrary to the widespread view that voice assistants and artificial intelligence should try to emulate human behavior, Ben Shneiderman proposes his view of human-centered artificial intelligence as follows:

> *"Successful robots utilize the distinctive features of machines. Robots will become more tool-like, tele-operated, and under human supervisory control through well designed user interfaces that avoid human-like features."* [40]

Creating a user interface with the goal to make it resemble humans severely limits the design options and can lead to ignoring the best possible solution. Furthermore, according to Roberts [37], humanizing computer systems can lead to three problems: incorrect use due to emotional attachment to a system, creating false expectations of a system, and inappropriate use of a system. Therefore, we argue that the words and phrases a VA speaks should not always pretend to be spoken by a human being. To explore this hypothesis, we evaluated whether a voice assistant that gives short, efficient answers without using full sentences can provide higher efficiency

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3517684

at equal usefulness compared to a full sentence VA.
Consequently, we implemented a browser-based VA and designed two more efficient, utilitarian response styles that do not use full sentences compared to the state-of-the-art. In an online study, 72 participants issued eight requests in a repeated-measures design, leading to 24 individual request-response pairs with the VA per participant.

Overall, we could not find a definite preference among our participants, neither for full-sentence responses nor for one of the two short, utilitarian responses. However, keyword responses were most preferred in all but three requests: For receiving the news, full sentences were preferred which is the familiar and known way of presenting them. When setting a timer, full sentences were slightly preferred, and for smart home commands, the very brief confirmation was most preferred. While this is an interesting insight in itself, we have also found that younger participants tend to prefer shorter response styles. Overall, the keyword response style was perceived as similarly useful and likeable while being more efficient, especially for simple voice commands, and sometimes also easier to comprehend.

The current trend towards humanizing and personifying assistants limits the design options and has clear negative implications for their efficiency. This work compares three response styles, presents empirical findings and implications for the design of VAs. To summarize, the main contributions of this work are: (1) An analysis of the current response behavior of the most popular VAs. (2) The design and implementation of a prototype VA answering eight typical requests in three response styles. (3) Empirical insights into (perceived) efficiency, usefulness, and acceptance of short, utilitarian response styles for VAs.

2 RELATED WORK

This section will give an overview of human-likeness in voice interfaces and recent research on VAs.

2.1 Human-likeness in Voice Interfaces

Nass and colleagues conducted a series of experiments to investigate social aspects in human-computer interaction [29, 30]. They found that despite users being aware that a computer is not a human and does not require human treatment, they still treated computers and machines as living, social beings. Their empirical findings indicate that individuals apply social rules and expectations to computers in an unconscious manner. Triggers of such a behavior can be language or speech output, computer responses based on multiple prior inputs, or computers filling roles that humans usually fill [31]. All those triggers are present in VAs, which contributes to feeling comfortable talking to those machines and may also make a more human-like VA more likeable.

Eun-Ju Lee has studied the effects of speech output in computer systems more intensively [23]. In an experiment, he compared synthetic and recorded speech output to better understand the influence of human-likeness on the social responses of users. The participants who relied more heavily on their intuition in a quiz game were susceptible to the influence of a computer voice with anthropomorphic features. Additionally, human speech significantly improved participants' evaluation of the computer's performance and compliance with its suggestions. This observation means that a more human assistant is usually considered more competent than a robotic one. We will investigate whether an efficient and utilitarian assistant still provides a high (perceived) usefulness while speeding up the interaction.

In general, it can be observed that the human characteristics of voice interfaces often appear beneficial in laboratory studies. For example, in a comparative study by Kühne et al. [22], the human voice and properties of a human speaker obtained consistently higher ratings in intelligibility, prosody, trustworthiness, confidence, enthusiasm, pleasantness, human-likeness, likability, and naturalness than a synthesized voice and the humanoid voice of a robot. Studies show that spoken dialog systems that use the same words as their users are perceived as more likeable and have more integrity [25]. However, when analyzing the expectations and perceptions of users, researchers found voice interfaces to be "more formal, fact based, impersonal and less authentic" [9] than human conversational partners. Despite the industry's attempt to portray voice assistants as likeable conversational partners, dialogues with voice assistants are severely limited and often fail to meet users' expectations in practice [27]. This was also analyzed by Clark et al. [7] by identifying what makes a good conversation between humans and how a human-agent conversation should look like. A more utilitarian and non-human voice assistant could help prevent such unrealistic expectations, ultimately providing a better user experience by reducing the gap between expectation and reality.

2.2 Current Research on VAs

Besides works that analyze the general usage of VAs [1, 34] great amount of research is currently focused on the impact of the personification of VAs. Purington et al. analyzed the user reviews of Amazon's Echo devices on Amazon.com [35]. They found that users referring to the device with the assistant's name ('Alexa') report having more sociable interactions. They also found increased levels of satisfaction for those users and a higher tolerance for errors and technological problems. This shows that the personification and anthropomorphism of voice assistants have a positive impact on the user experience. Many aspects influence personification and anthropomorphism. For example, Weiss et al. found that perceived agency and voice interaction design strongly affect anthropomorphism, even more than physical appearance [43]. Our work investigates whether efficient voice assistants, that do not use full sentences as humans typically do, but utilitarian keyword responses can still provide a better user experience.

Personality is a frequently studied property of digital assistants with human speech behavior [4, 41, 42]. Kuzminykh et al. have chosen an interesting approach to identify the perception of a VAs' character and personality. They used an initial interaction session with the VAs, semi-structured interviews, and finally, a visualization task with an avatar generator to describe and vividly illustrate the personality of popular VAs [21]. While a synthesized voice alone can elicit social responses when interacting with voice interfaces, character and personality are primarily determined by dialogue flow and response behaviors when answering questions. This could have a strong impact on the user experience if, as we aimed to do, an assistant does not use complete sentences and the content of

the spoken words has no human characteristics. Other features of voice assistants that has been studied were trust [6] and privacy [28] which tend to become more critical with human-like digital assistants. It was even investigated how verbal insults and possible counterattacks by a voice assistant affect the user's emotion [5]. Many of these problems are not present or may be more easily solved with a utilitarian and non-human VA.

3 COMMERCIAL VA RESPONSES

To analyze how current voice assistants behave and respond, we analyzed the most commonly used VAs in more detail. According to [11], Apple's Siri, Amazon's Alexa, and Google's Assistant are the VAs with the highest market share worldwide. Those three assistants all exist in the form of software applications, for example, on smartphones, but are also available as standalone smart speakers (Apple's Homepod series, Amazon's Echo series, and Google's Home series). This analysis focused on the standalone type of VAs as they usually do not have a display and rely only on speech synthesis as output. To obtain a comprehensive overview of the response behavior of the three assistants, we used them to perform requests in the most frequent topics (weather, calendar, news, general knowledge, music, timer/alarm, reminder, and home automation) [1, 36] and analyzed their responses. VAs do not always answer with the exact same wording; there is some variation in their answers; however, the general structure of the responses is always similar.

Weather is the topic on which most requests are issued. Therefore, we included a typical request about the weather the next day. In the responses listed in Table 1, we found that assistants include the cloudiness (e.g., sunny, cloudy, scattered clouds, etc.) or the precipitation info (e.g., rain, snow, heavy rain) depending on the weather report. In addition, they always include the daily maximum and daily minimum temperatures. For Amazon Alexa and Google Assistant, the reports' location is also embedded in the response to provide feedback to the user if it matches their current or expected location. **Calendar** and appointment management is another topic that people like to delegate to digital assistants [36]. Since the creation of reminders will be covered separately, we decided to include a query about existing appointments for the next day as a representative for this topic. All assistants replied with the total number of existing events in the calendar on the day, followed by giving the starting time and title of the events. They usually stop after four events and ask if further events should be read out.

Requesting **news** led to short responses that announce that the news is going to be presented, followed by a pre-recorded news program, usually from third-party sources such as podcasts.

The next tested topic was **general knowledge** questions. To the question of how many people live in the US, all VAs responded with full sentences repeating the topic (population), the country in question, the date of data collection, and the actual requested population size.

For a request for **music playback**, the actual playing of the

Topic	Request	Alexa	Siri	Google Assistant
Weather	*What's the weather tomorrow?*	Tomorrow in CITY, there will be CLOUDINESS, with a high of HIGH TEMP degrees Celsius/Fahrenheit and a low of LOW TEMP degree.	There could be some PRECIPITATION tomorrow. The high will be HIGH TEMP, and the low will be LOW TEMP.	In CITY tomorrow there will be PRECIPITATION with a high of HIGH TEMP and a low of LOW TEMP.
Calendar	*What's on my calendar tomorrow?*	Tomorrow there are two events. At TIME there is TITLE. At TIME there is TITLE.	You have two appointments. On DAY at TIME TITLE, At TIME TITLE.	There are two entries for tomorrow. First up, you have TITLE at TIME. Second is TITLE at TIME.
News	*Tell me the latest news.*	Here's your news. [starts news program]	Here's the latest news from SOURCE. [starts news program]	Here is the latest news. [starts news program]
General knowledge	*How many people live in the US?*	In 2020, the population of the United States was 331 million people.	As of 2021, the population of United States of America is 332 million 278 thousand 200.	In 2019, the population of the United States of America was 328 million 239 thousand 523.
Music	*Play TITLE by INTERPRET on SERVICE.*	Playing TITLE by INTERPRET on SERVICE.	TITLE by INTERPRET now playing on SERVICE.	TITLE by INTERPRET, sure. Playing on SERVICE.
Timer / Alarm	*Set a timer for X minutes.*	X minutes, starting now.	Okay, your timer is set for X minutes.	Okay, X minutes starting now.
Reminder	*Remind me to TITLE on DAY at TIME*	Okay, I'll remind you DAY at TIME.	Okay, your reminder is set for DAY, TIME.	Got it. I'll remind you on DAY at TIME.
Home automation	*Turn on LIGHT.*	Okay.	Okay, LIGHT is on.	Okay, turning on X lights.

Table 1: The table provides an overview of the popular VAs Alexa, Siri, and Google Assistant responses to frequently used queries and commands.

requested song would provide enough feedback to the user if the request was successfully detected and executed. Nevertheless, all assistants reply with a full sentence, repeating the title, interpret, and service the music will be played on.
As a representative of **timers and alarms**, we requested the assistants to set a timer to a specific number of minutes. In all three cases, responses were full sentences repeating the number of minutes and confirming that a timer was started. Very similar, the request of setting **reminders** was confirmed by sentences repeating the day and time of the reminder. Noticeably, the title of the event was not repeated.
For **home automation**, we choose the typical request of switching lights on or off. The responses were a simple 'Okay.' by Alexa and two variations additionally indicating which or how many lights have been switched on/off for Google Home and Apple Siri.
When analyzing the VAs' responses, the first thing to notice is that all tested VAs used a female voice as the default. In combination with the frequent use of singular first-person pronouns, female voices could reinforce gender stereotypes [14]. However, this is not the case here as the assistants rarely referred to phrasings such as 'I'll do that for you'. The more important finding for this work is that all assistants usually utilized full sentences to respond. A single exception is Amazon's Alexa in the home automation request, where it only confirms the request with "Okay" but is not giving any further information. Alexa is generally the assistant with the shortest responses, especially for commands such as **music**, **timer**, **reminder**, and **home automation**. Commands and queries are important distinctions that can be made regarding the requests to the VA. On the one hand, there are *commands* that only need to be executed and confirmed, and on the other hand, *queries* where information is needed that the assistant should communicate to the user. As a general observation, much of the information, such as the intent (e.g., setting a timer) and variables (e.g., 5 minutes) given in the request, will be repeated by the voice assistant in its response. One reason why voice assistants currently repeat that information is to make possible errors transparent to the user so they can recognize and correct them. It is important to mention that both Amazon's Alexa and Google's Assistant provide a "brief mode" that users can enable in their settings menu [10, 26]. Despite the naming of these modes, they do not actually shorten the voice output, but the "brief mode" lets the assistants emit a soft confirmation tone after smart home commands such as switching on lights or playing music instead of the spoken "Okay" confirmation.

To summarize, all three VAs are designed for natural language interaction and, therefore, expect full sentences as input and respond in full sentences per default. They mimic human language and try to create the impression of being a social companion instead of providing utilitarian responses that only include the bare information necessary to answer a given request successfully. This intention is also evident through many of the added features such as when asking the assistants questions about themselves, the possibility of letting them tell jokes or wishing a good night/morning. While social qualities can make the assistant seem likeable, long sentences as responses can also feel annoying to users as in other works "Alexa was commonly criticized for providing too much information to queries [...]" [9]. The dialog between user and voice assistant is not social conversation but transactional conversation, which is characterized by serving to convey factual or propositional information [15]. Therefore, it is unnecessary to have a chatty design of VAs with regards to the transactional nature of the dialog between human and machines, especially when considering frequently used commands. As the result of an interview with 20 VUI designers, Kim et al. [16] summarized that the designers "described natural human speech as often being indirect and inefficient, so these aspects of human conversation should be left out when designing for a natural VUI". These observations inspired us to design a different, more efficient response behavior and investigate its usefulness and acceptance.

4 USER STUDY

We conducted an online study to investigate the perceived usefulness, efficiency, and user acceptance of different levels of VA responses. Based on the analysis of requests to and responses by popular voice assistants, we designed and implemented a browser-based VA that can respond to the most frequently used requests in three different ways.

4.1 Requests and Response Styles

As in the analysis of the state of the art of VA responses, we choose to use the eight requests discussed earlier, answered in three response styles described below.

4.1.1 Requests. Only a limited set of features of the VA were required for the study, which is described in this section. Regarding **weather**, the assistant only answered requests about the weather the next day. When asked about **calendar** entries or the schedule for the next day, the assistant reported two pre-defined events at 12 PM and 7 PM. **News** could be requested by users, which led to the assistant reading out an exemplary headline from a news source. As a **general knowledge** question, we included the population size of the United States of America. For **music**, the VA could play only one song (Ukulele song by Rafael Krux from https://freepd.com/) back that faded out after 5 seconds of playback. **Timers** could be set to any specific amount of minutes, and **reminders** could be set for any time. Instead of using the popular "lights on" request as a representative of **home automation**, we let users change the background color of the website as it produces a similar effect in the sense that users are immediately able to perceive the result of their request.

4.1.2 Response styles. The requests were answered in three response styles. The first response style contained only the bare information; the second additionally contained feedback to the given request. The third was in full sentences and very similar to the state-of-the-art responses (see Table 2 and Table 3 for all responses). To determine the essential information for the first response style, it was important to distinguish between *commands* and *queries*. For *queries*, the assistant always needs to give verbal feedback. So the *minimal* response style provided this information for **weather**, **calendar**, **news**, and **general knowledge** (for all responses to queries, see Table 2). In our case **music**, **timer**, **reminder**, and **home automation** are *commands* as they do not

Topic	*minimal*	*keyword*	*full sentence*
Weather	"sunny, 5 to -3 degrees"	"Weather tomorrow: sunny, 5 to -3 degrees"	"The weather tomorrow is gonna be sunny with a high of 5 degrees and a low of -3 degrees."
Calendar	"Lunch with Steven, 12PM. Dinner with Ann, 7PM."	"Meetings tomorrow: Lunch with Steven, 12PM. Dinner with Ann, 7PM."	"You have two appointments. At 12 PM there is Lunch with Steven and at 7 PM dinner with Ann."
News	"BBC - Australia not intimidated by Facebook news ban."	"News today: BBC - Australia not intimidated by Facebook news ban."	"Here is what I found. BBCs latest headline is: Australia not intimidated by Facebook news ban."
General knowledge	"328 million."	"US population: 328 million."	"The total size of the US population amounts to 328 million."

Table 2: Table with responses for *queries* in the three response styles minimal, keyword, and full sentence.

necessarily require verbal feedback. The assistant does not need to provide any information back to the user to execute them successfully but simply performs the given task. For instance, switching on a light already provides sufficient feedback for the user. Therefore, at least in principle, commands do not require any additional verbal feedback, and the first response style is only *confirmation* (for all responses to commands, see Table 3). The second response style is in a *keyword* format for both - *commands* and *queries*. This response style provides additional feedback about the recognized request and input variables, and when applicable, the answer is added (e.g., for **weather** "Weather tomorrow: sunny, 5 to -3 degrees"). As a result, this response style provides exactly the same information as the full sentence baseline but without the additional words that would make it a grammatically correct, complete sentence. It thus also allows the user to identify potential recognition errors of the assistant but requires significantly fewer words and therefore less time than the full sentence responses by omitting fill and linking words.

The third response style is for *commands* and *queries* in *full sentences*. The sentences were designed to be similar to the sentences that we found in the analysis of available VAs. Although the current state of the art of full sentences in voice assistants aims to be as human as possible, it can also be argued that it is the very short responses of the *minimal* response style that genuinely are human. In conversations between humans, written grammar rules are often ignored, and the average number of words in a phrase is significantly lower than in written language [24]. From this perspective, these response styles are not a gradation from human to utilitarian, but extreme points of short and long responses as they could occur in human conversation. However, one response that a human person would never give is the *keyword* response style. This is a constructed response that assumes that misunderstandings will occur more often in the communication between humans and machines instead of between humans. Therefore, the shortest possible answer is accompanied by additional feedback. It helps the user to verify if the request was understood by the VA correctly by repeating the topic of the request, e.g., "Weather tomorrow: -3 to 2 degrees, sunny". This added feedback allows the user to determine if a response matches their request or if a recognition error occurred. In order to provide an interactive exploration of these response styles, we developed a web application for the Chrome browser. This is publicly available to try out at https://keepitshortdemo.github.io/Demo/.

To show the effects of the three response styles on the duration of the speech output, we timed the speech outputs of the VA for each response style. Whole sentences took a median of 4.03 seconds for the speech output, keyword responses took a median of 2.38 seconds, and minimal/confirmation responses took a median of 1.13 seconds. In other words, keyword responses require only 67% of the time of the full sentence responses, and minimal/confirmation responses require even less at only 40% of the time of the full sentence baseline.

Topic	*confirmation*	*keyword*	*full sentence*
Timer / Alarm	"Okay."	"Timer, 10 minutes"	"Okay, I set a timer to 10 minutes. Starting now."
Music	*music playback starts*	"Rafael - Ukulele song"	"Okay, I'm playing Ukulele song by Rafael for you."
Reminder	"Okay."	"Get Cake, 5 pm."	"Okay, I'll remind you to Get Cake at 5 pm."
Home automation	*background color changes*	"Background color: blue"	"Okay, your background color is set to blue."

Table 3: Table with responses for *commands* in the three response styles confirmation, keyword, and full sentence.

4.2 Apparatus

In the user study, participants had the task of posing various requests to a voice assistant. Therefore, a voice assistant was created using JavaScript and the WebSpeech API [32]. We used the "Google UK English Female" voice for speech synthesis, which can be selected and listened to in any WebSpeech implementation when using Google's Chrome browser. It was not necessary for the intended study setting to reproduce the full functionality of a modern voice assistant but individual functions only. Instead of using a wake word as standalone VAs in smart speakers do, we used an activation button that users had to click before issuing their request. To actively involve the participants, they were given instructions in a way that they needed to find their own words for a request rather than just reading it out (e.g., "Please ask the assistant about tomorrows weather."). After clicking the button, a user could issue requests as described in subsection 4.1. The user's request was transcribed, stored, and processed to issue the correct response. The request type was identified by searching for keywords such as 'weather', 'timer', 'reminder' in the transcribed request string. Then, depending on the identified request, the string was searched for variables such as the day of the week and time in the reminder request or the number of minutes for the timer request. When our algorithm identified all variables, the VA gave the corresponding response via speech synthesis. If a variable was found missing, a voice response would inform the participant that this specific information was necessary for the request and encouraged them to repeat the request. If no keyword could be detected to identify the request type, a voice response informed the user that the assistant did not understand the request.

4.3 Procedure

The user study consisted of three parts: First, there was a verification of whether the voice input and output were operating properly. Checking if everything is operating properly was done by letting users type a sentence that was synthesized as voice output after pressing a button and letting them dictate a given sentence to the system. Subsequently, the main part of the study took place, which was the interaction with the VA. After a user completed the interaction with the VA, a final questionnaire was used to collect overall feedback and demographics. For the main part, we used a repeated-measures design, in which each participant had to perform the eight different requests in three response styles, resulting in 24 request/response interactions for each participant. We decided to keep the request topics as blocks to allow for better comparability between response styles. However, the three response styles within these blocks were randomized so that each participant experienced a different order of responses. In this way, learning effects are counterbalanced and compensated. After each successful request, the VA spoke the appropriate answer, and the participant was asked to rate items described below on 7-point semantic differential scales. The corresponding questionnaire appeared only when the correct request was detected, assuring that participants actually issued the correct requests. We selected the semantic differentials from the extension of the user experience questionnaire (UEQ+) [19, 38], a modular framework that provides scales for many user experience-related questions. To keep response times between requests short, we only picked the two semantic differentials with the highest loading from three scales *Response behavior* (artificial - natural, unlikeable - likable), *Response quality* (not helpful - helpful, useless - useful), and *Comprehensibility* (complicated - simple, unambiguous - ambiguous). Next, we added the two items with the highest loading from the *Efficiency* scale (slow - fast, inefficient - efficient). Since the questionnaire creators do not validate this use of the items, we calculated the Cronbach alpha values as a measure of internal consistency for *Response behavior* (0.79), *Response quality* (0.92), *Comprehensibility* (0.15), and *Efficiency* (0.65). While we found high or sufficient alpha values for the other three, *Comprehensibility* has a low alpha value. For the individual items (complicated - simple and unambiguous - ambiguous) it is constructed of, opposing answers do not necessarily contradict each other (e.g., a response can be simple, which is usually good, but also ambiguous, which is bad) but together describe and contribute to the understanding of comprehensibility. Therefore, we argue that the scale is meaningful and can be used despite the low agreement between the two single items.

After successful interaction with the assistant in all response styles, we collected users' preferences of response styles per request. To capture our participants' general attitudes toward voice assistants, we included four statements on which the participants had to indicate their agreement on a 7-point Likert scale (1 = Disagree strongly, 7 = Agree strongly). "I see a voice assistant as a technical system." and "I see a voice assistant as a social companion." were aimed at the relationship between user and VA, "Efficiency is most important when using a voice assistant." and "Human-likeness is most important when using a voice assistant." was supposed to show what is important to them when interacting with the assistant. Finally, we used the Affinity for Technology Interaction (ATI) questionnaire [12] to assess participants' tendency to engage in interaction with technologies. The Ten Item Personality Inventory (TIPI) [13] was used to measure the big five personality traits briefly. Thereby, we intended to test whether these characteristics of the participants have an influence on the preferred response length of a VA. At the very end of the survey, we included an open question to gather general feedback. This open question allowed the participants to address topics that have not been explicitly asked so far.

4.4 Participants

We recruited 72 participants via Prolific [1]. Prolific is a UK-based platform, and participants are primarily White/Caucasian from the UK (31.3 %), US (26.8 %), and Europe (17.4 %). As others have shown, the Prolific participant pool is equally reliable but more diverse than MTurk [33], which is often used in scientific work and has been shown to generalize well to a broad population [18]. One participant had to be removed from the results because they provided inconsistent answers. The age of participants ranged from 18 to 67 years, with a median of 23 and an interquartile range (IQR) of 7 years. Since we screened our participants in advance for those who use voice assistants, it is not surprising that many young people were recruited as participants. Young people are the largest group of voice assistant users [17]; however, there are also older users

[1] https://www.prolific.co/

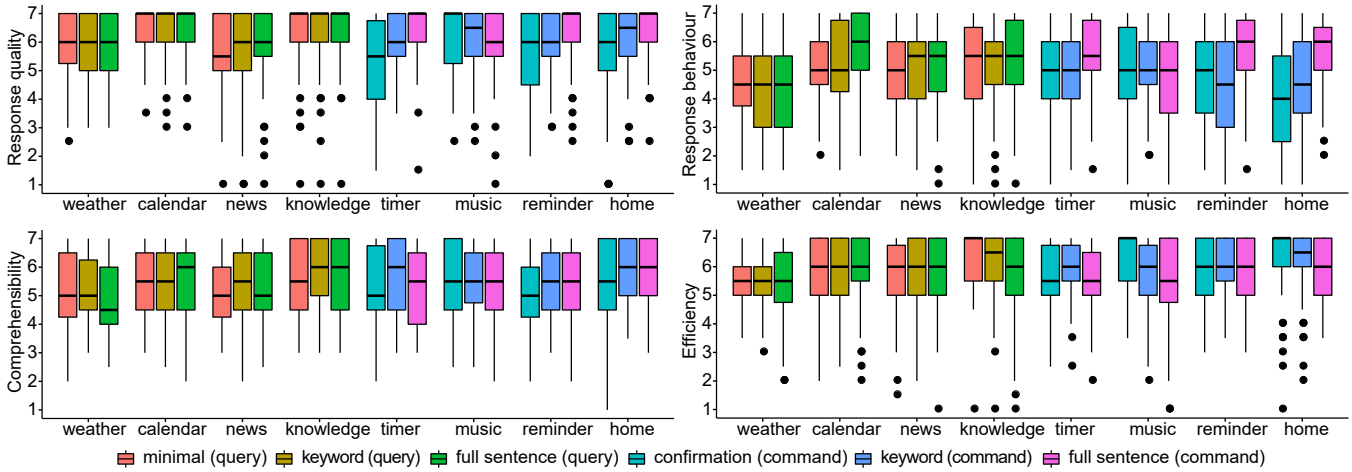

Figure 1: Ratings of response behavior, response quality, comprehensibility, and efficiency of the eight requests in three response styles.

about which we cannot draw strong conclusions. The gender distribution was shifted towards male, with 45 participants reported identifying as male (63%), 24 identified as female (34%), a single participant identified as diverse, and another one preferred not to tell. As pre-screening criteria, we only recruited participants with an acceptance rate above 95%, meaning they have successfully carried out at least 95% of previous surveys to ensure the high reliability of participants. We also decided only to recruit participants who indicated in Prolific to possess and have used a VA before. Our participants indicated they use Google's Assistant (42.3 %, n = 30), Amazon's Alexa (32.4 %, n = 23), Apple's Siri (18.3 %, n = 13), Microsoft's Cortana (5.6 %, n = 4), and Samsung's Bixby (1.4 %, n = 1). Having used a specific VA could also introduce some bias towards a response style. However, as we showed earlier, the current commercial VAs are very similar in their responses, and it is unlikely that there is a significant difference or bias between users of different VAs. Whether there is a bias between users of VAs and non-users cannot be answered by our experiment and should be investigated in future work. We considered it more important for our user study that our participants were familiar with the general interaction with voice assistants so that they are, in contrast to non-users, able to evaluate the different response styles. These non-users cannot assess how often requests are misunderstood in practice or how important the implicit feedback of the different response styles is in everyday use. Already possessing and using a VA makes them expert users as they already know the context of use of such a system and can generalize the different response styles more accurately to everyday situations.

5 RESULTS

The results are structured in three parts: first, the ratings per request and response pair are described, followed by the preferences and attitudes towards VAs. Finally, the open feedback is presented.

5.1 Ratings per Request and Response Style

The ratings of response quality, response behavior, comprehensibility, and efficiency are displayed in Figure 1. We used the non-parametric Friedman's test to identify statistically significant differences between the response styles for each request. In those cases, pairwise comparisons (pwc) were carried out using Wilcoxon signed-rank test compensated with Bonferroni correction.

The *response quality* scale is composed of the two semantic differentials, *not helpful (1) - helpful (7)* and *useless (1) - useful (7)*. The ratings of response styles for **news** showed significant differences ($\chi^2(2) = 6.99$, $p < 0.01$, $W = 0.049$). A pairwise comparison showed significant differences between minimal and full sentences ($p < 0.05$) response styles as full sentences were rated higher. In the **timer** request, significant differences with small effect size ($\chi^2(2) = 32.24$, $p < 0.0001$, $W = 0.227$) were found between all response styles (pwc: confirmation - keyword $p < 0.0001$, confirmation - full sentences $p < 0.0001$, keyword - full sentences $p < 0.05$) as confirmation was rated lowest, followed by keyword and full sentence responses. When setting a **reminder**, ratings of the response styles also showed significant differences ($\chi^2(2) = 24.27$, $p < 0.0001$, $W = 0.171$) for all response styles (pwc: confirmation - keyword $p < 0.01$, confirmation - full sentences $p < 0.0001$, keyword - full sentences $p < 0.01$). Ratings of responses for the **home automation** request also showed significant differences ($\chi^2(2) = 11.2$, $p < 0.005$, $W = 0.079$) for confirmation - keyword ($p < 0.05$) and confirmation - full sentences ($p < 0.005$). In the ratings of **weather**, **calendar**, **knowledge**, and **music** no significant differences were found.

The *response behaviour* scale is composed of the two semantic differentials *artificial (1) - natural (7)* and *unlikeable (1) - likeable (7)*. We found significant differences with a small effect size in the **calendar** request ($\chi^2(2) = 18.73$, $p < 0.0001$, $W = 0.132$). Pairwise comparison showed that the minimal responses ($p < 0.001$) and

the keyword responses ($p < 0.01$) were significantly less natural and likeable as the full sentence response style was rated higher than the other two. We also found significant differences for the **timer** request ($\chi^2(2) = 19.65$, $p < 0.0001$, $W = 0.138$). Again, the full sentence response style scored higher than the other two (confirmation - full sentence $p < 0.0001$, keyword - full sentence $p < 0.001$). The same pattern also shows in the **reminder** request ($\chi^2(2) = 20.69$, $p < 0.0001$, $W = 0.146$, pwc: confirmation - full sentences $p < 0.01$, keywords - full sentences $p < 0.0001$). In the **home automation** request, we also found significant differences ($\chi^2(2) = 36.89$, $p < 0.0001$, $W = 0.260$) and pairwise comparisons showed significant differences for all combination (confirmation - keywords $p < 0.01$, confirmation - full sentences $p < 0.0001$, keywords - full sentences $p < 0.0001$). The ratings of requests for **weather**, **news**, **knowledge**, and **music** showed no significant differences.

Comprehensibility is constructed from *complicated (1) - simple (7)* and *ambiguous (1) - unambiguous (7)*. We found significant differences with a small effect size ($\chi^2(2) = 17.00, p < 0.0005, W = 0.120$) for the **weather** request. The pairwise comparison revealed that the full sentence was rated significantly lower than both shorter response styles (minimal - full sentences $p < 0.05$, keyword - full sentences $p < 0.005$). The ratings for **news** also showed significant differences($\chi^2(2) = 6.71$, $p < 0.05$, $W = 0.047$). Interestingly, the keyword condition is rated highest and significantly more comprehensible than the minimal response style ($p < 0.05$). The same pattern applies to the response styles for **general knowledge** with significant differences ($\chi^2(2) = 11.38$, $p < 0.005$, $W = 0.080$) between keywords and full sentences ($p < 0.05$). Ratings for **calendar**, **timer**, **music**, **reminder**, and **home automation** showed no significant differences.

The last scale to be described is *efficiency*. It is constructed from the two semantic differentials *inefficient (1) - efficient (7)* and *slow (1) - fast (7)*. The ratings of responses to the **general knowledge** request showed significant differences with small effect size ($\chi^2(2) = 7.95$, $p < 0.05$, $W = 0.056$). While medians of 7, 6.5, and 6 for the response styles minimal, keywords, and full sentences hint a clear link between response length and rated efficiency, pairwise comparisons only showed significant differences between minimal and full sentences ($p < 0.05$) responses. The same pattern applies to **music** ($\chi^2(2) = 21.36$, $p < 0.0001$, $W = 0.150$) although pairwise comparisons show significant differences between confirmation and keyword ($p < 0.005$) as well as confirmation and full sentences ($p < 0.001$). Likewise, the same pattern applies to **home** ($\chi^2(2) = 10.15$, $p < 0.01$, $W = 0.072$) with significantly more efficient confirmation ratings compared to full sentences ($p < 0.05$). **Weather**, **calendar**, **news**, **timer**, and **reminder** showed no significant differences in their ratings.

5.2 Preferences and Attitude towards VAs

After interacting with the VA in all response styles, participants were asked to select their favorite response style for each request. This information is displayed in Figure 2. In general, no clear trend was apparent as to which response style was most preferred. Participants only preferred the minimal and confirmation styles most often for **home automation** requests. The keyword response style

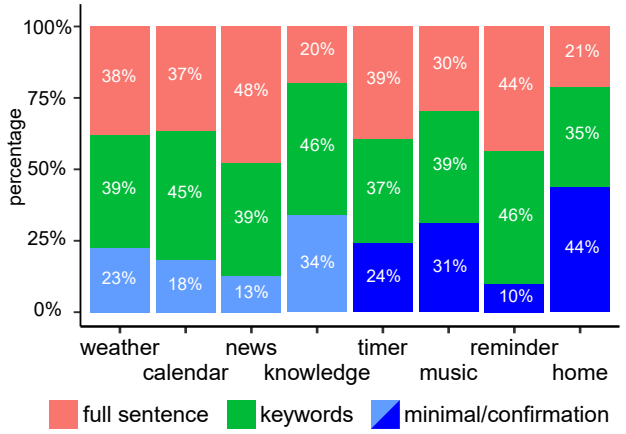

Figure 2: The participants' preferences regarding the three response styles per request.

is the most often preferred response style, namely for **weather**, **calendar**, **general knowledge**, **music**, and **reminder**. Full sentences are only preferred most often for **news** and **timer**.

We were interested as to whether the preferred response length correlated with any other traits of the participants. Therefore, we calculated the non-parametric Spearman's correlation between participants' preferred response length and their personality traits of the big five measured with the TIPI [13], affinity for technology interaction [12], age, and gender. We found a weak positive statistically significant correlation between participants preferences for response length and age ($R = 0.25$, $p < 0.05$), implying that younger participants prefer shorter answers and older participants prefer longer answers. Other correlations were found to be not significant.

The agreement (see Figure 3, 1 = Disagree strongly, 7 = Agree

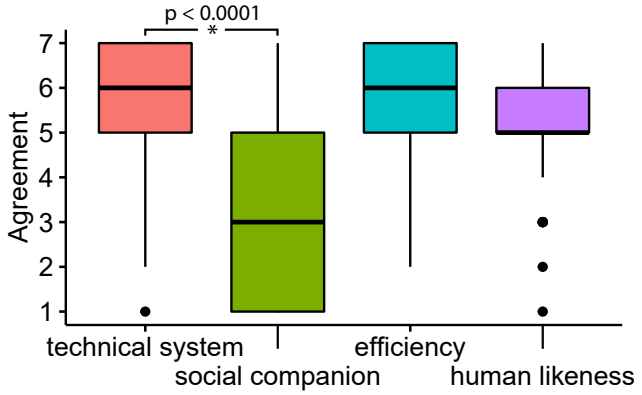

Figure 3: Agreement (1 = Disagree strongly, 7 = Agree strongly) of the participants with statements of VAs being a technical system / a social companion and statements of efficiency is most important / human-likeness is most important in a VA.

strongly) with the statement "I see a voice assistant as a technical system." was high with a median of 6 and an IQR of 2. The statement "I see a voice assistant as a social companion." received a lower agreement ($M = 3$, $IQR = 4$). A Wilcoxon signed-rank test showed that the difference between the agreement to those statements was highly significant ($V = 1797$, $p < 0.0001$). Similar, but less clear, was the agreement with "Efficiency is most important when using a voice assistant." ($M = 6$, $IQR = 2$) and "Human-likeness is most important when using a voice assistant." ($M = 5$, $IQR = 1$). The difference between the agreement to both statements is just below being statistically significant ($V = 999.5$, $p = 0.05157$).

5.3 Open Feedback

We used thematic analysis to structure and interpret the responses to the open feedback included at the end of the survey. The thematic analysis was performed in a lightweight process of inductive category development as described by Kuckartz [20] in QCAmap [2]. The analysis of this feedback has brought up several interesting aspects. Out of the 72 participants, 42 used this opportunity to provide additional feedback. Ten, again, expressed their preference for one of the response styles. Five participants mentioned that they want the assistant to repeat the requests' input parameters (e.g., weather tomorrow) because it provides important feedback if the assistant understood the user's request correctly. Three participants explicitly shared their desire for an option to adjust the assistants 'verbosity' or response style, and two mentioned that full sentences made the assistant sound more human. Twelve participants provided feedback on the implementation of our VA. Regarding the synthesized voice, two mentioned that the speech rate was sometimes too fast. Six participants mentioned that the synthesized voice could or should be more natural and human-like. Finally, four participants commented that they liked the assistant and were impressed by its accuracy. Furthermore, four participants reported usage behavior with their own VA, and one mentioned the confirmation tone of their Google Home (when "brief mode" is enabled) as a welcome alternative to the "Okay." response of our assistant.
Participants also provided general feedback regarding their participation in the online study. Six of them explicitly mentioned that they liked participating in the study, and three reported voice recognition errors that occurred during the study.

6 DISCUSSION

For most participants, VAs are definitely more technical systems than social companions (see Figure 3). Accordingly, the efficiency of the answers was rated more important than human-likeness. Interestingly, however, a large proportion of the participants stated that human-likeness was nevertheless rather important to them. These results are also reflected in the preferences for certain response styles. Especially in requests where the shorter response styles were perceived as more effective, an increased preference towards shorter answers was observed. Some exceptions, though, can probably be explained by the habit of the everyday life of the participants. The individual requests are discussed separately below.

[2]https://www.qcamap.org/

6.1 Commands

Commands in which the result was directly observable by the participants were rated as more efficient if only the action was executed (as in the confirmation response style), for example, when setting the background color **(home automation)** or playing **music**. The preference for the confirmation response style (e.g., 44% for **home automation**) was correspondingly high for these questions. The proportion of participants who wanted to hear complete sentences in response to these questions was likewise low (e.g., 21% for **home automation**). It is also noticeable that human response behavior seems to play a rather subordinate role for the users' preferences for these requests. In **music** and **home automation**, the response styles of full sentences were rated highest in *response behaviour*, but were least often preferred. Those ratings indicate that when only executing commands, the voice assistant is perceived more like a machine than as a social companion with whom one converses.
The situation is slightly different for the other two commands **timer** and **reminder**, especially for the **reminder**. For **reminder**, only 10% of the participants preferred the response style in which the VA only confirmed the execution of the command. The situation is similar for the **timer**, where 24% of the participants preferred the confirmation response style. We suspect that the rather low preference for the confirmation response style is linked with comprehensibility, which was significantly rated lowest. We presume that this is due to the lack of feedback via the response, leading to a lack of trust in the assistant. This lack of trust was also explicitly mentioned by participants in the open feedback: "It is good to acknowledge instruction so you know that it has gotten it right." While getting direct feedback of whether the assistant understood a command correctly when playing the correct music, for example, this feedback was lacking in the responses to the **timer** and **reminder** requests.

6.2 Queries

The **general knowledge** request regarding the population of the USA had the highest preference for short answers. Here, only 20% of the participants preferred the full sentence response. But also for the other queries, the majority of participants tended to prefer responses in the keyword or minimal response style. The **news** request can be considered an outlier where a full sentence response was preferred by almost half of the respondents, and only 13% preferred the minimal response style.
The results of the ratings can explain these preferences. There were hardly any differences in the perceived response quality for the queries, so participants perceived all response lengths similarly helpful and useful. Only the **general knowledge** request regarding the population size of the USA was perceived as noticeably more efficient the shorter it was formulated. Again, efficiency seems to be the most important factor for the participants.
The ratings of **weather** and **calendar** were relatively similar. The majority of the participants preferred the shorter response styles over complete sentences, although they were not perceived as more efficient. Interestingly, for the **calendar**, unlike all other requests, response behavior and comprehensibility were rated significantly higher for full sentences than for the two shorter response styles. We suspect that the nevertheless high agreement for short responses

could again be due to the habit of the participants. If asking someone what they are doing on a certain day (or for the weather respectively), one does not expect them to repeat the date embedded in the question.

The comparatively high preference for complete sentences in the **news** request is difficult to explain with the results of the ratings. Only the minimal response style showed significantly lower scores and was rated as no more efficient than the other two answers. We suspect that the high tendency towards whole sentences comes from the participants' habit of having **news** read to them since newsreaders read in full sentences and do not just list keywords.

6.3 General Observations

As the three response styles do not represent an ordinal scale and due to the different forms of requests (commands and queries), the interpretation of the results is more complex. Although there is an overall trend towards a preference for short answers, minimal answers are mainly preferred for commands, and keyword responses are preferred for queries. We attribute this to the fact that the execution of a command in itself provides good feedback, e.g., by turning on the specified light or playing the requested song, the user already knows that the VA understood them correctly. For queries, it is apparently more important for the users to be able to check whether the response that the VA provided matches their request or whether there was a misidentification of the intent and the provided response does not match the users' request. Therefore, the additional feedback of the keyword response is preferred for general knowledge queries such as the population size of the US. Just the provided amount as a number could fit many requests and a misidentification of the query cannot be detected by the user in the minimal response style. Besides, we also found a small subset of participants who value social behavior and communication with VAs. One of our participants stated: "I like my vocal assistant to answer me in full sentences as if it was a real person talking to me. This is because it sounds less alienating to me when thinking about talking to a machine." This statement shows that there are also users that always prefer complete sentences, care less about increasing the efficiency of the interaction but enjoy conversing with the assistant.

6.4 Human Emulation Implies Human Capabilities

When considering our findings along with those of Clark et al. [7] and Luger et al. [27], it is very likely that our participants are biased by the capabilities and characteristics of the VAs they are currently using. Emulating human behavior in a VA raises the users expectations, as it implies human capabilities [27]. However, current VAs cannot live up to these expectations if a user expects and relies on human capabilities. To prevent breakdowns caused by such unrealistic expectations, Cowan et al. proposed that "using a less human-like voice that signals more basic conversational abilities [...] may facilitate a mental model that is closer to the true abilities" of a VA [8]. However, intentionally degrading voice reproduction is not the right solution from our perspective. Instead, we argue that the same effect could be achieved by using a less social dialog, such as in the keyword response style that is clearly not human-like in it's grammatical structure but is efficient and provides the necessary feedback for users.

6.5 Implications for the Design of VAs

In essence, our results show that the more straightforward and well-defined the task is, the simpler and shorter the response should be. Current VAs are not considered truly conversational partners but are only capable of performing rather simple tasks. Therefore, there is no need to establish a meaningful connection with them, and brief, basic responses were often considered appropriate. As VAs evolve, this preference will need to be re-evaluated according to the capabilities of the devices. However, with the current state-of-the-art, short, utilitarian responses can increase the users' overall experience compared to always resorting to full sentence responses.

An interesting extension to the concept of utilitarian VAs are more adaptive VAs. For example, based on the briefness and the way a users' request is spoken, responses of a VA should be very brief or more extensive. A first step towards this is to actually implement the partially existing "brief mode", moving beyond the substitution of the spoken "Okay" confirmation by a confirmation tone, utilizing the presented results. VAs should also respond more precisely to requests: does the user want to hear just a number or also background information? Currently, many requests lead to the same response. For example, asking Siri to "Tell me about the population of the United states." leads to the same response as "How many people live in the United states?".

One of the main takeaways of this study is that there is not a one-fits-all solution for the design of VAs. The use of VAs is very diverse and ranges from smart home commands to control the lights in a living room to isolated trivia questions on a smartphone. This context, of course, influences if a user wants to converse with a VA or wants to get a task done or some information delivered as quickly as possible. Moreover, VAs are also used as accessibility tools by people who cannot use any visual interfaces. Branham and Mukkath Roy [3] showed that the recommendation of development guidelines by VA suppliers does not match the needs of blind users. However, they provide implications for the design of inclusive VAs that match surprisingly well with what we found: the response style should be user-configurable, either on the fly (e.g., by letting the user ask for a brief weather update that leads to a shorter response than a regular weather request) or on a system level (e.g. by putting the VA in a brief mode for smart home users so voice commands are not answered by full sentence responses). Making the briefness and style of responses adaptive or adjustable makes sure to not deprive any user of the ability to engage in conversation with the VA but will allow those users who desire a higher level of efficiency to receive exactly that.

6.6 Limitations

The validity of this online study is, of course, inferior to that of a long-term field study. The best case would be implementing the response styles in the voice assistants actually used by the participants as this takes place in the real context of use. Such a study would require full implementation of the assistant with three response styles, not only eight specific requests. However, the resulting time

and effort required for such an implementation were not feasible for this research project. The study we conducted has the advantage that the participants could directly compare different response styles. In addition, the users' impressions can be directly captured, which compensates for at least some of the drawbacks. In future work, we would like to cover a more realistic pattern of use and also recruit participants who have no previous experience with voice assistants. Thereby we want to evaluate the potential influence of prior use of voice assistants on the preference and user experience of utilitarian voice assistants. A possible limitation is our participant pool that is not representative in terms of age and gender, resulting from an online study and a pre-screening for voice assistant users. Although VAs are mostly used by younger people [17] and the age of our participant pool resulted from them being our target group, it would still be desirable to cover a broader population. The majority of our study participants identified themselves as male, and about a third identified themselves as female. We tested the results for a relationship between gender and reported preferences and found no effect. Therefore, we are confident that there is no systematic gender effect that would question our findings. A more general limitation of online studies is the lack of control about how participants perform these studies. We were able to verify at the beginning of the study that voice input and output worked correctly. However, any other problems, for example, speech recognition errors, as three participants reported, may introduce some distortion in the results.

7 CONCLUSION

By analyzing and mapping out the response behavior of current voice assistants, we found that they almost always utilize full sentences as responses, even for commands where no verbal feedback would be necessary at all. We designed an interactive VA prototype that can respond to eight typical requests in three response styles. The implementation of this VA allowed us to compare a state-of-the-art, full-sentence baseline to short, utilitarian responses that require only 40% - 67% of the time for the speech output. We provide empirical findings from this comparison that highlight the conflict between the current implementation of VAs that only use whole sentences and their perceived efficiency. While two of eight requests were preferred to be answered in full sentences, the execution of simple commands does not require detailed verbal feedback since the execution of a command already provides feedback on its own. For instance, in home automation requests, the preference for the full-sentence responses was as low as 21%. We also found that younger participants more often preferred the shorter response styles, and the keyword response style were perceived as equally useful, likeable, and sometimes even more comprehensible. The perceived efficiency (based on users' ratings) and actual efficiency (measured) of keyword responses was shown to be higher than the baseline.

Designing the response of a VA with a human-sounding voice but without using whole sentences opens up the design space and contributes to a successful and human-centered interaction [39]. With such responses, users only receive the information from the device they have requested, and the VA is not imposed on them as an equal partner. We want to emphasize that the current trend towards humanizing and personifying assistants limits the design options and has clear negative implications for efficiency. Because of a narrow focus on the personification of the VAs, other properties currently receive little attention. As also evident in our analysis, people have diverse needs and preferences. We do not suggest that all voice assistants should only use keywords to answer requests. Still, the design of VAs should reflect these needs and preferences and give users the option to adapt their digital assistant to their very own personal needs.

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] Daniel Lee Ashbrook. 2010. *Enabling mobile microinteractions.* Dissertation. Georgia Institute of Technology. https://smartech.gatech.edu/handle/1853/33986 Accepted: 2010-06-10T17:03:08Z Publisher: Georgia Institute of Technology.

[3] Stacy M. Branham and Antony Rishin Mukkath Roy. 2019. Reading Between the Guidelines: How Commercial Voice Assistant Guidelines Hinder Accessibility for Blind Users. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 446–458. https://doi.org/10.1145/3308561.3353797

[4] Michael Braun and Florian Alt. 2019. Affective Assistants: A Matter of States and Traits. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. Association for Computing Machinery, Glasgow, Scotland Uk, 1–6. https://doi.org/10.1145/3290607.3313051

[5] Hyojin Chin, Lebogang Wame Molefi, and Mun Yong Yi. 2020. Empathy Is All You Need: How a Conversational Agent Should Respond to Verbal Abuse. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, Honolulu, HI, USA, 1–13. https://doi.org/10.1145/3313831.3376461

[6] Eugene Cho, S. Shyam Sundar, Saeed Abdullah, and Nasim Motalebi. 2020. Will Deleting History Make Alexa More Trustworthy? Effects of Privacy and Content Customization on User Experience of Smart Speakers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, Honolulu, HI, USA, 1–13. https://doi.org/10.1145/3313831.3376551

[7] 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)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300705

[8] 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)*. Association for Computing Machinery, New York, NY, USA, Article 43, 12 pages. https://doi.org/10.1145/3098279.3098539

[9] 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, 1–12. https://doi.org/10.1145/3338286.3340116

[10] Rita El Khoury. 2019. Google Assistant on the Pixel 4 can be 'brief' with its answers. https://www.androidpolice.com/2019/10/24/google-assistant-on-the-pixel-4-can-be-brief-with-its-answers/

[11] Simon Forrest. 2019. Virtual Assistants to Exceed 2.5 Billion Shipments in 2023. https://www.futuresource-consulting.com/insights/virtual-assistants-to-exceed-2-5-billion-shipments-in-2023/?locale=en

[12] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human–Computer Interaction* 35, 6 (April 2019), 456–467. https://doi.org/10.1080/10447318.2018.1456150 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2018.1456150.

[13] Samuel D Gosling, Peter J Rentfrow, and William B Swann. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in Personality* 37, 6 (Dec. 2003), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

[14] Florian Habler, Valentin Schwind, and Niels Henze. 2019. Effects of Smart Virtual Assistants' Gender and Language. In *Proceedings of Mensch und Computer 2019*

(MuC'19). Association for Computing Machinery, New York, NY, USA, 469–473. https://doi.org/10.1145/3340764.3344441

[15] C. J. Hookway. 1978. Semantics By John Lyons Cambridge University Press, 1977. Vol. 1, xiii 371 pp., £12.00, £3.95 paper; Vol. 2, xiv 526 pp., £15.00, £4.95 paper. *Philosophy* 53, 205 (1978), 421–423. https://doi.org/10.1017/S0031819100022579

[16] Yelim Kim, Mohi Reza, Joanna McGrenere, and Dongwook Yoon. 2021. *Designers Characterize Naturalness in Voice User Interfaces: Their Goals, Practices, and Challenges.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3411764.3445579

[17] Bret Kinsella. 2019. Voice Assistant Demographic Data – Young Consumers More Likely to Own Smart Speakers While Over 60 Bias Toward Alexa and Siri. http://voicebot.ai/2019/06/21/voice-assistant-demographic-data-young-consumers-more-likely-to-own-smart-speakers-while-over-60-bias-toward-alexa-and-siri/ Section: Amazon alexa.

[18] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 453–456. https://doi.org/10.1145/1357054.1357127

[19] Andreas M. Klein, Andreas Hinderks, Martin Schrepp, and Jörg Thomaschewski. 2020. Construction of UEQ+ scales for voice quality: measuring user experience quality of voice interaction. In *Proceedings of the Conference on Mensch und Computer (MuC '20)*. Association for Computing Machinery, New York, NY, USA, 1–5. https://doi.org/10.1145/3404983.3410003

[20] Udo Kuckartz. 2019. Qualitative Text Analysis: A Systematic Approach. In *Compendium for Early Career Researchers in Mathematics Education*, Gabriele Kaiser and Norma Presmeg (Eds.). Springer International Publishing, Cham, 181–197. https://doi.org/10.1007/978-3-030-15636-7_8

[21] Anastasia Kuzminykh, Jenny Sun, Nivetha Govindaraju, Jeff Avery, and Edward Lank. 2020. Genie in the Bottle: Anthropomorphized Perceptions of Conversational Agents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, Honolulu, HI, USA, 1–13. https://doi.org/10.1145/3313831.3376665

[22] Katharina Kühne, Martin H. Fischer, and Yuefang Zhou. 2020. The Human Takes It All: Humanlike Synthesized Voices Are Perceived as Less Eerie and More Likable. Evidence From a Subjective Ratings Study. *Frontiers in Neurorobotics* 14 (2020), 105. https://doi.org/10.3389/fnbot.2020.593732 Publisher: Frontiers.

[23] Eun-Ju Lee. 2010. The more humanlike, the better? How speech type and users' cognitive style affect social responses to computers. *Computers in Human Behavior* 26, 4 (July 2010), 665–672. https://doi.org/10.1016/j.chb.2010.01.003

[24] Geoffrey Leech and Jan Svartvik. 2013. *A Communicative Grammar of English* (3 ed.). Routledge, London. https://doi.org/10.4324/9781315836041

[25] 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 (May 2018), 173–186. https://doi.org/10.1093/iwc/iwy005

[26] Craig Lloyd. 2019. What Is Alexa's Brief Mode and How Do I Turn It On (or Off)? https://www.howtogeek.com/346176/what-is-alexa%e2%80%99s-brief-mode-and-how-do-i-turn-it-on-or-off/

[27] 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 - CHI '16*. ACM Press, Santa Clara, California, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[28] Michal Luria, Rebecca Zheng, Bennett Huffman, Shuangni Huang, John Zimmerman, and Jodi Forlizzi. 2020. Social Boundaries for Personal Agents in the Interpersonal Space of the Home. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, Honolulu, HI, USA, 1–12. https://doi.org/10.1145/3313831.3376311

[29] Clifford Nass and Youngme Moon. 2000. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues* 56, 1 (2000), 81–103. https://doi.org/10.1111/0022-4537.00153 _eprint: https://spssi.onlinelibrary.wiley.com/doi/pdf/10.1111/0022-4537.00153.

[30] Clifford Nass, Youngme Moon, B. J. Fogg, Byron Reeves, and D. Christopher Dryer. 1995. Can computer personalities be human personalities? *International Journal of Human-Computer Studies* 43, 2 (Aug. 1995), 223–239. https://doi.org/10.1006/ijhc.1995.1042

[31] 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 (CHI '94)*. Association for Computing Machinery, New York, NY, USA, 72–78. https://doi.org/10.1145/191666.191703

[32] André Natal, Glen Shires, Philip Jägenstedt, and Hans Wennborg. 2020. Web Speech API. https://wicg.github.io/speech-api/

[33] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70 (May 2017), 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

[34] 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)*. Association for Computing Machinery, Montreal QC, Canada, 1–12. https://doi.org/10.1145/3173574.3174214

[35] Amanda Purington, Jessie G. Taft, Shruti Sannon, Natalya N. Bazarova, and Samuel Hardman Taylor. 2017. "Alexa is my new BFF": Social Roles, User Satisfaction, and Personification of the Amazon Echo. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '17*. ACM Press, Denver, Colorado, USA, 2853–2859. https://doi.org/10.1145/3027063.3053246

[36] Felix Richter. 2017. Users Learn to Appreciate Smart Speakers' Many Talents. https://www.statista.com/chart/9579/smart-speaker-use-cases/

[37] Lionel Robert. 2017. *The Growing Problem of Humanizing Robots.* SSRN Scholarly Paper ID 3027628. Social Science Research Network, Rochester, NY. https://papers.ssrn.com/abstract=3027628

[38] Martin Schrepp and Jörg Thomaschewski. 2019. Design and Validation of a Framework for the Creation of User Experience Questionnaires. *International Journal of Interactive Multimedia and Artificial Intelligence* 5, Regular Issue (2019), 88–95. https://www.ijimai.org/journal/bibcite/reference/2730

[39] Ben Shneiderman. 2020. Design Lessons From AI's Two Grand Goals: Human Emulation and Useful Applications. https://ieeexplore.ieee.org/document/9088114

[40] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Three Fresh Ideas. *AIS Transactions on Human-Computer Interaction* 12, 3 (2020), 109–124. https://doi.org/10.17705/1thci.00131

[41] Sarah Theres Völkel, Daniel Buschek, Malin Eiband, Benjamin R. Cowan, and Heinrich Hussmann. 2021. Eliciting and Analysing Users' Envisioned Dialogues with Perfect Voice Assistants. *arXiv:2102.13508 [cs]* na (Feb. 2021), na. https://doi.org/10.1145/3411764.3445536 arXiv: 2102.13508.

[42] 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 (CHI '20)*. Association for Computing Machinery, Honolulu, HI, USA, 1–14. https://doi.org/10.1145/3313831.3376210

[43] Astrid Weiss, Anna Pillinger, Katta Spiel, and Sabine Zauchner-Studnicka. 2020. Inconsequential Appearances: An Analysis of Anthropomorphic Language in Voice Assistant Forums. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, Honolulu, HI, USA, 1–7. https://doi.org/10.1145/3334480.3382793