Expanding Speech Interaction for Domestic Activities

Nima Zargham

First Supervisor: Prof. Dr. Rainer Malaka
Second Supervisor: Prof. Dr. Benjamin Cowan

Digital Media Lab
Faculty 3: Mathematics / Computer Science
University of Bremen

Submitted on April 29th, 2024
Defended on June 25th, 2024

A dissertation submitted in partial fulfilment of the requirements for the degree of Doctor of Engineering (Dr.-Ing.).

WWW.UNI-BREMEN.DE

Wednesday 26th June, 2024

University of Bremen

Declaration by Postgraduate Students

Authenticity of Dissertation

I hereby declare that I am the legitimate author of this Dissertation and that it is my original work. No portion of this work has been submitted in support of an application for another degree or qualification of this or any other university or institution of higher education. All direct or indirect sources used are acknowledged as references. I hold the University of Bremen harmless against any third party claims with regard to copyright violation, breach of confidentiality, defamation and any other third party right infringement.

.

Faculty	Faculty 3: Mathematics / Computer Science
Degree	Doctor of Engineering (Dr.-Ing.)
Title	Expanding Speech Interaction for Domestic Activities
Candidate (Id.)	Nima Zargham (3014908)

Signature of Student _____

Date Wednesday 26th June, 2024

To Shahed and Pari

Thanks for all the love.

Acknowledgements

I would like to express my heartfelt gratitude to Prof. Dr. Rainer Malaka for his unwavering support and invaluable guidance throughout my PhD journey. His optimism and wisdom have significantly boosted my motivation and self-confidence. I would also like to express my profound appreciation to Prof. Dr. Benjamin Cowan for his expert insights and tremendous support throughout this dissertation. His constructive feedback and encouragement have been a constant source of inspiration. Special thanks to my advisor, Dr.-Ing. Nina Wenig, for her insightful feedback and guidance on the earlier versions of this manuscript. Additionally, I am deeply grateful to Prof. Dr. Yvonne Rogers from University College London and Prof. Dr. Lennart E. Nacke from the University of Waterloo for graciously hosting my research exchanges and fostering collaborative efforts. Their mentorship has greatly enhanced my understanding of human-computer interaction and provided invaluable insights.

I am extremely grateful to my colleagues at the Digital Media Lab and those who have collaborated with me on my research endeavors. A special thanks to Mehrdad Bahrini, Michael Bonfert, Leon Reicherts, and Dmitry Alexandrovsky for the close collaborations and valuable exchanges. I would also like to thank Thomas Mildner, Vino Avanesi, Bastian Dänekas, Thomas Münder, Evropi Stefanidi, Carolin Stellmacher, Nadine Wagener, Daniel Diethei, Ava Elizabeth Scott, Maximilian A. Friehs, Kamyar Javanmardi, Ioannis Bikas, Sebastian Höffner, Anke Reinschlüssel, Laura Spillner, Georg Volkmar, Leon Dratzidis, Ameneh Safari, Katharina Hasenlust, Robin Nolte, Rachel Ringe, Johanna Rockstroh, Fenja Schweder, Johannes Pfau, Susanne Putze, Robert Porzel, Karsten Sohr, Tanja Döring, Gerald Volkmann, Dirk Wenig, Irmgard Laumann, Philipp Harms, Svenja Voß and Evgenia Sazonkina for their collaboration, mutual inspiration, exchange of ideas, and camaraderie. Working with such a supportive and talented group of colleagues has been a privilege.

Last but certainly not least, I want to express my deepest gratitude to my family and friends for their love and support. Thank you, Mehrdad, Afi, Navid, and Niusha. Your presence in my life is my biggest source of happiness and strength. Without you, this dissertation would not have been possible.

Abstract

Due to technological advancements, communicating with computer systems using natural language has become a common phenomenon. Speech-based systems have become widely popular among people in everyday activities due to the intuitive nature of their interaction. Speech interaction inherently encompasses a social component as it reflects the fundamental human capacity for communication and enables interpersonal engagement through verbal exchange. This makes speech interaction with computers an essential topic of research in the field of human-computer interaction. Along with the development of speech-based systems, users' demands and expectations from such systems grow. Despite the popularity of speech interaction, designing a gratifying experience for users interacting with such systems remains challenging. This can be attributed partly to technical constraints, such as challenges with speech recognition, and partly to experiential limitations where these systems fail to meet users' needs and expectations as communication partners.

This dissertation explores human-agent speech interaction in domestic activities through a series of user studies and interviews. The primary focus lies in the practical application of speech technology for everyday activities, particularly within two application domains of homes and video games, to identify factors contributing to successful speech interaction. Drawing inspiration from communication and human-computer interaction models, a novel interaction model is introduced to provide a more nuanced understanding of the dynamics in human-agent speech interaction. Different dimensions of speech systems are analyzed, including the utility and efficacy of speech systems, diverse representations of speech agents, and the style in which these systems interact with users. By examining users' needs and addressing current issues, designing new features to enhance existing systems, and proactively anticipating potential future challenges, this work takes a comprehensive approach, encompassing a retrospective, current, and future outlook. The aim is to provide a broad perspective on designing speech systems, offering relevant design factors and recommendations to achieve a higher user experience with such systems. This thesis contributes to the fields of human-computer interaction, voice user interfaces, game user research, and user experience design within both academia and industry.

Zusammenfassung

Aufgrund des technologischen Fortschritts ist die Kommunikation mit Computersystemen unter Verwendung natürlicher Sprache zu einem alltäglichen Phänomen geworden. Sprachbasierte Systeme sind aufgrund der intuitiven Art ihrer Interaktion bei den Menschen in ihrem Alltag sehr beliebt geworden. Sprachliche Interaktion beinhaltet von Natur aus eine soziale Komponente, da sie die grundlegende menschliche Fähigkeit zur Kommunikation widerspiegelt und zwischenmenschliches Engagement durch verbalen Austausch ermöglicht. Dies macht die Sprachinteraktion mit Computern zu einem wesentlichen Forschungsthema im Bereich der Mensch-Computer-Interaktion. Mit der Entwicklung sprachbasierter Systeme wachsen auch die Anforderungen und Erwartungen der Nutzer an solche Systeme. Trotz der Beliebtheit der Sprachinteraktion ist es nach wie vor eine Herausforderung, die Interaktion mit solchen Systemen für den Benutzer angenehm zu gestalten. Dies kann zum Teil auf technische Einschränkungen zurückgeführt werden, wie z. B. Herausforderungen bei der Spracherkennung, und zum Teil auf erfahrungsbedingte Einschränkungen, wenn diese Systeme die Bedürfnisse und Erwartungen der Benutzer als Kommunikationspartner nicht erfüllen.

In dieser Dissertation wird die Mensch-Agent-Sprachinteraktion bei häuslichen Aktivitäten anhand einer Reihe von Nutzerstudien und Interviews untersucht. Das Hauptaugenmerk liegt auf der praktischen Anwendung von Sprachtechnologie für alltägliche Aktivitäten, insbesondere in den beiden Anwendungsbereichen Haushalt und Videospiele, um Faktoren zu identifizieren, die zu einer erfolgreichen Sprachinteraktion bei häuslichen Aktivitäten beitragen. In Anlehnung an Modelle der Kommunikation und der Mensch-Computer-Interaktion wird ein neuartiges Interaktionsmodell vorgestellt, das ein differenzierteres Verständnis der Dynamik in der Mensch-Agent-Sprachinteraktion ermöglicht. Es werden verschiedene Dimensionen von Sprachsystemen analysiert, die Aspekte wie den Nutzen und die Wirksamkeit von Sprachsystemen, verschiedene Darstellungen von Sprachagenten und den Stil, in dem diese Systeme mit den Benutzern interagieren, umfassen. Durch die Untersuchung der Bedürfnisse der Benutzer und die Behandlung aktueller Probleme, die Entwicklung neuer Funktionen zur Verbesserung bestehender Systeme und die proaktive Vorwegnahme potenzieller zukünftiger Herausforderungen verfolgt diese Arbeit einen umfassenden Ansatz, der einen Rückblick, eine aktuelle und eine zukünftige Perspektive umfasst. Das Ziel ist es, eine breite Perspektive auf die Entwicklung von Sprachsystemen zu bieten und relevante Designfaktoren und Empfehlungen zu geben, um eine höhere Benutzererfahrung mit solchen Systemen zu erreichen. Diese Arbeit leistet einen Beitrag zu den Bereichen Mensch-Computer-Interaktion, Sprachsteurung, Spieleforschung und User Experience Design in Wissenschaft und Industrie.

Contents

List of Figures

List of Abbreviations

1

Introduction

Speaking is a primary mode of human communication, serving as a fundamental method for conveying thoughts, ideas, and emotions since ancient times. Voice emerges as a fundamental and intrinsic aspect of social interaction within human beings and extends to interactions beyond our species, such as between humans and animals. With its diverse vocalizations and communicative abilities, voice has played a crucial role in developing and maintaining social bonds throughout the evolutionary process (Seaborn et al., 2021). Furthermore, voice has been instrumental in facilitating cooperation, coordination, and collective activities among individuals, ultimately contributing to the survival and success of social groups.

Technological advancements have now enabled us humans to communicate with computer systems using our voice. Speaking is a natural way of communication among humans, and people find it easier to interact with technology that resembles their own characteristics (Breazeal, 2003). Voice recognition systems have been arguably developed since the late 1950s when Bell Laboratories designed "Audrey," which could recognize spoken digits (Li and Mills, 2019; Meng et al., 2012). These early systems were relatively basic and required a controlled environment with limited background noise to function effectively. Thanks to advances in research and technology, such systems have grown significantly in capability and sophistication, now capable of handling complex tasks. Over the past few years, various technologies integrating voice-based capabilities have emerged. Interactive systems that enable users to communicate with computers, devices, or applications using their voice are referred to as Voice User Interfaces (VUIs) (Pearl, 2016). These systems often consist of an Artificial Intelligence (AI) Agent, which can perceive the environment, process information, and act autonomously to achieve specific goals (Seaborn et al., 2021). Voice interaction is now incorporated as a feature in a variety of devices, including but not limited to smartphones, personal computers, cars, and smart speakers (Knote et al., 2019). Social robots, Voice Assistants (VAs), and other voice-based embodiments of AI have become prevalent in contemporary society (Seaborn et al., 2021). The domestic setting, the focus of this dissertation, is one of the main applications of voice interaction through devices such as smart speakers or smart displays. Such devices are

used for various purposes, such as smart home control, scheduling, online shopping, and music playback, among others.

Yet, despite the technological strides and the advantages these systems offer, designing a truly satisfying experience with VUIs remains a formidable challenge. Many people still express dissatisfaction, frustration, or discomfort when interacting with voice systems (Carter et al., 2015; Luger and Sellen, 2016; Porcheron et al., 2018; Bonfert et al., 2018), primarily due to issues with speech recognition and constrained functionality (Jentsch et al., 2019). This is exacerbated by the fact that these systems often fall short of meeting users' expectations as conversation partners (Jentsch et al., 2019; Luger and Sellen, 2016; Murad and Munteanu, 2019; Doyle et al., 2019).

Verbal communication is a complex and multifaceted phenomenon, requiring the harmonious alignment of various elements. These elements include not just the words spoken but also nonverbal cues, tone, context, and the receptiveness of both the speaker and the listener. Humans excel at communication and hold high expectations for effective communication partners. This poses a significant challenge when AI agents attempt to assume the role of a communication partner.

To design speech-based systems that are more desirable and appealing, both the hedonic and pragmatic qualities of Human-Agent Speech Interaction (HASI) need to be considered. The term 'hedonic qualities' here refers to the subjective and emotional aspects of the User Experience (UX), encompassing elements of enjoyment, pleasure, and overall affective response derived from the interaction (Laugwitz et al., 2008a). On the other hand, 'pragmatic qualities' pertain to the utilitarian aspects, focusing on practical and functional attributes that contribute to fulfilling users' goals and satisfying their specific requirements. Hedonic qualities are associated with factors such as aesthetics, novelty, fun, and entertainment value. On the other hand, pragmatic qualities encompass factors such as efficiency, usability, task completion, and reliability. Therefore, the design goals for these systems generally fall into two broad categories: *functional goals* and *experiential goals* (Zamfirescu-Pereira et al., 2023). *Functional goals* focus on the agent's ability to complete tasks, answer questions, and provide useful information, while *experiential goals* emphasize creating a positive and engaging user experience that fosters satisfaction, trust, and a sense of natural conversation. To accomplish these goals, designers work at distinct levels of abstraction. This involves defining the structure of the dialogue flow, ensuring that each stage of the human-agent conversation serves the user's goals effectively. Additionally, it encompasses fine-tuning the agent's utterances to enhance the overall quality of interaction, ultimately seeking to achieve a more pleasant user experience (Zamfirescu-Pereira et al., 2023).

In the initial stages, speech-based systems were highly limited, operating on a command and response basis where users initiated commands, and the system executed corresponding actions if recognized (Bolt, 1980). These early systems prioritized utilitarian functionality, employing digital feedback forms to inform users about command processing. However, there was minimal emphasis on these systems' representation

and interaction style. As technology progressed, speech systems evolved to handle more complex tasks (Hannun et al., 2014). Along the way, with more research on this topic, researchers and developers began to explore the experiential dimensions of these systems. Synthetic voices improved, and exploration expanded to include diverse visualization forms and agent personalities. Most of these design aspects pursued human-likeness, entailing cognitive capabilities, personality, and physical appearance (Waytz et al., 2010; Zlotowski and Bartneck, 2013). This evolution transformed user interactions with speech systems, shifting from simple, single-turn systems with limited vocabulary to more sophisticated systems capable of multi-turn conversations that imitate human qualities. These systems have assumed roles as advanced social actors (Nass et al., 1994), resembling companions rather than mere utilities (Pradhan et al., 2019; Lee et al., 2017). Furthermore, there is a growing body of research on proactive speech systems (Edwards et al., 2021; Kraus et al., 2020; Miksik et al., 2020), which are systems that initiate conversations with users. This challenges several traditional interaction models, depicting the process as something that starts with a user input and ends with a system output. In light of these transformations, researchers argue for the term "partner-based" interaction (Peña et al., 2023; Doyle et al., 2023). This term aims to encapsulate the essence of modern speech systems, which act as digital partners or companions, possessing diverse human-like qualities and robust functional capabilities to assist users in their tasks. The interaction paradigm has evolved from a transactional process to a dynamic and engaging partnership.

To gain a comprehensive understanding of the intricate nature of user experience with speech-based systems, it is important to understand the underlying components that mold this interaction. Shifting our focus from the users' holistic encounter with speech-based systems, we turn our attention to the factors that shape this experience.

1.1 | Speech Interaction Components

In communication science, several communication frameworks have traditionally explained the intricate information exchange process between two actors, a sender and a receiver, through a specific channel (Fujishin, 2008). These models provide valuable overviews of the complexities inherent in communication dynamics. Additionally, multiple interaction frameworks have been developed for human-computer interaction, ranging from Norman's interaction cycle (Norman, 1986) and the general interaction framework by Abowd and Beale (Abowd and Beale, 1991) to Nigay's Human-Computer Interaction (HCI) model (Nigay, 1994). However, a noticeable gap exists regarding interaction frameworks tailored for speech, especially considering the evolution of technology and the subsequent changes in interaction patterns. One of the goals of this dissertation is to bridge this gap by introducing a novel interaction model specifically designed for speech with agents, acknowledging the unique challenges and opportunities

presented by this evolving landscape.

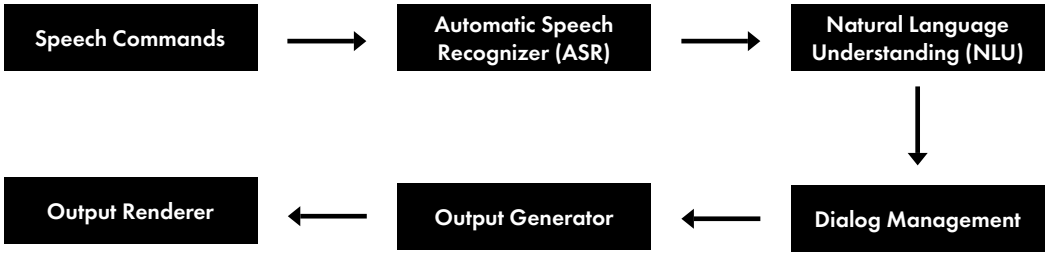


Figure 1.1: The dissected process of human-agent speech interaction.

From a technical point of view, the process of speech interaction initiates with the user speaking to a computer system via a microphone (Sridhar and Tolentino, 2017). The input the user gives to the system is called the "speech command". The speech command is then processed to filter out ambient noise. The filtered audio is then digitized or decoded using Automated Speech Recognition (ASR), which breaks down the speaker's unique voice pattern into discrete segments of several tones. Next, the input text undergoes analysis using Natural Language Understanding (NLU), where meaning is derived from the speech data. The system then employs dialog management to determine how to respond to the input, resulting in a computer-readable representation of the response. An output generator is then used to convert the machine-readable response into a human-readable format, typically in the form of text. Subsequently, an output renderer converts the text into speech, a process known as Text-to-Speech (TTS) (see Figure 1.1). Most of these processes occur "behind the scenes," allowing users to interact seamlessly with the voice-based system (Harris, 2004).

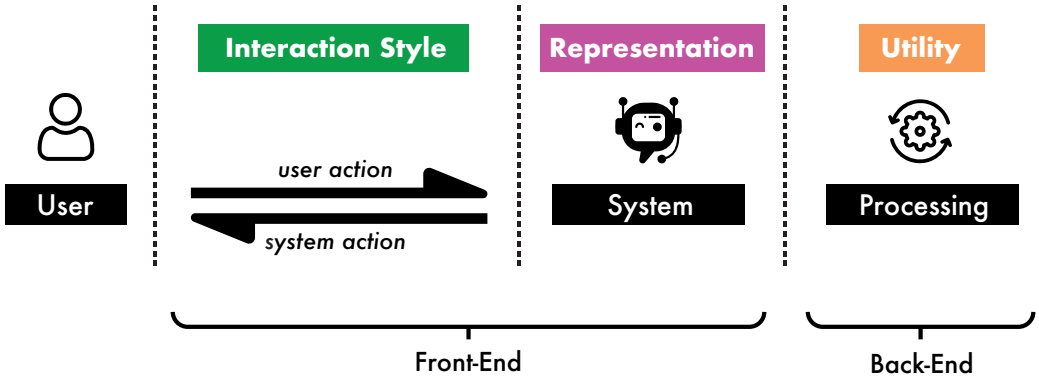


Figure 1.2: The components of human-agent speech interaction, depicting the front-end with the system's representation and interaction characteristics, and the back-end consisting of utility components.

To describe the functioning of speech interaction systems, we can categorize them into

two components: the front-end and the back-end (see Figure 1.2). The front-end entails the user-facing aspects, which can be referred to as the presentation layer, dedicated to optimizing *experiential goals*. This includes the system's appearance, the visual or auditory representation of the speech agent, and the interaction characteristics. On the other hand, the back-end constitutes the processing and utility components. This layer mainly focuses on achieving *functional goals*. The successful integration and functioning of these two components are critical in delivering a satisfying and desirable speech interaction experience that meets the user's expectations.

This dissertation conducts a comprehensive analysis of both the back-end and front-end aspects of speech systems. On the back-end, the dissertation delves into strategies to enhance the utility aspect of speech systems, introducing approaches to improve their overall efficacy. Simultaneously, on the front-end, the work investigates diverse representations of speech agents, exploring their influence on users' overall experiences with these systems. Additionally, it examines novel interaction styles, seeking to understand their impact on user engagement and satisfaction.

1.2 | Terminologies

The existing literature on VUIs presents a diverse and sometimes confusing array of terms to refer to various elements of voice interaction. Voice is often associated with speech, which refers to the linguistic expression of words through the combination of vowel and consonant sounds, known as phonetics (Fitch, 2017). However, when referring to speech, the focus is on the linguistic content and the semantic meaning conveyed through spoken words. While speech represents a significant voice component, it is not the sole aspect. Voice also encompasses non-linguistic forms of communication, including intonation, pitch, timbre, and other non-verbal cues that contribute to communication and expression. It can convey various social attributes, including gender and personality traits (Seaborn et al., 2021). Therefore, understanding voice requires acknowledging its multifaceted nature and recognizing that speech is just one facet within the broader concept of voice.

The following aims to clarify some of the elements of voice interaction:

- **Voice User Interface**: The system elements that the users interact with, which may involve both audio and visual components (Porcheron et al., 2018).
- **Conversational Agent**: A virtual character capable of engaging in conversation with the user, equipped with the ability to perceive its environment, process information, and autonomously execute actions to achieve specific goals (Allouch et al., 2021). This can be a computer program or artificial entity designed to facilitate communication with users. This is also referred to as a *virtual assistant*, or *dialogue system*.
- **Speech Command**: The user's input provided to the system, serving as the directive for the intended action (Sridhar and Tolentino, 2017). It's the way users communicate

their desires to the system, initiating the interactive process.

- **Wake Word**: A specific word or phrase that activates the system, prompting it to listen for a user's command (e.g., "Hey Siri" or "Alexa").
- **Speech-to-Text (STT)**: The technology that translates spoken language into text, enabling computers and applications to understand and process human speech.
- **Text-to-Speech (TTS)**: The technology that translates text into spoken language, enabling VUIs to respond audibly to user inputs.

1.3 | Aims and Objectives

This dissertation focuses on the practical implementation of speech technology for domestic activities, primarily focusing on two specific domains: home assistants and speech-based video games. These applications are a common part of everyday activities in contemporary daily life. Home assistants are voice-controlled virtual assistants that aim to support people in various aspects of daily life within a household (e.g., Amazon Echo with Alexa [1], Google Home with Google Assistant [2], or Apple HomePod with Siri [3]). On the other hand, speech-based video games are interactive digital games that incorporate speech recognition as an integral component of the gameplay to control in-game events and engage in dialogues with Non-Player Characters (NPCs). This thesis delves into both these applications, exploring their unique features and the role of speech interaction within them.

The focus on the domestic setting emphasizes the accessibility to the general population rather than being confined to expert users. At the same time, it acknowledges the increasing prevalence and adoption of VAs in everyday household settings, a trend that is anticipated to escalate further in the future (Malodia et al., 2021). Speech interaction in homes accommodates both single-user and multi-user scenarios, facilitating a wide array of tasks, from essential activities like banking to casual and leisure pursuits. Homes are a unique setting where people feel most safe and are at maximum comfort. It is where technology steadily integrates into our daily lives. Furthermore, it explores the potential for interaction with other smart devices like cameras, TVs, and locks, enabling seamless integration with voice assistants to enhance their capabilities. Furthermore, the exploration of speech interaction within the video game domain serves as a deliberate choice due to its controlled environment offering a high degree of experimental control. Games provide an environment that is relatively easy to replicate, and developing research prototypes is often more financially feasible than in other settings. Additionally, there is a reduced concern for privacy compared to conducting studies in actual homes,

[1] https://alexa.amazon.com/
[2] https://assistant.google.com/
[3] https://www.apple.com/homepod/

where individuals may be uncomfortable with always-listening devices. Games, being immersive and entertaining, have the ability to foster intrinsic motivation for users to engage with such systems. This choice of domain thus combines experimental advantages with user engagement.

While these two domains could be rather distinct in their use cases and challenges, they share inherent similarities, particularly concerning their interaction process. Examining these diverse applications will illustrate that the underlying research questions can be applied to a variety of domains.

The following research questions are approached in this thesis:

- **TRQ1:** What factors contribute to successful partner-based speech interaction in domestic activities?
- **TRQ2:** How can we enhance the *efficacy* of speech systems for domestic activities?
- **TRQ3:** How does the *representation* of virtual speech agents contribute to a better interaction experience?
- **TRQ4:** What is the appropriate *interaction style* in speech systems for domestic activities?

The first research question (**TRQ1**) sets the overarching theme for this dissertation, serving as a foundational inquiry that intersects with the other three questions. **TRQ2** focuses on the utility dimension of speech systems. This dissertation tackles this question by analyzing existing challenges and shortcomings and proposing innovative solutions to enhance system efficacy through two dedicated studies. **TRQ3** examines the representation and appearance of speech systems. Through an exploration encompassing five separate studies, this dissertation investigates various elements of this dimension, including embodiment, personality, and multiplicity, shedding light on individual user perceptions and preferences. Lastly, **TRQ4** investigates speech systems' interaction styles to identify and tailor approaches that align with individual user preferences. Through three dedicated studies, this dissertation provides a deeper understanding of designing interaction styles better suited to individual user preferences, laying the groundwork for developing more personalized and user-centric speech systems.

These research questions serve as the foundation for guiding the research in this thesis, encompassing the development of prototypes and evaluations to uncover fresh insights. Throughout this dissertation, the specific research questions driving the thesis are referred to as **TRQs**, while the research questions guiding individual papers are denoted as **RQs** to ensure clarity in distinguishing between the thesis's overarching research objectives and the questions addressed within each paper.

Contribution

This doctoral dissertation by publication aims to broaden the horizons of speech interaction by exploring its applications for everyday activities. This work contributes to the body of the literature in Conversational User Interfaces (CUIs) and the broader field of HCI by identifying prevalent issues and challenges associated with existing speech systems for everyday activities and identifying gaps in the literature on this topic, determine opportunities for successful speech interaction for everyday activities, and propose insightful recommendations to refine the design of such systems. Moreover, the dissertation introduces a novel interaction model tailored explicitly for HASI. The model aims to offer an overview of the intricacies involved in the speech interaction process. The contributions stem from empirical evaluations involving real users in the application domains. These evaluations aim to provide applicable and valid insights into the underlying mechanisms of how people perceive and use the technology. The thesis relies on several prototypes developed for two application domains, employing various evaluation methods such as lab studies, online surveys, and interviews. These approaches extend the theoretical foundation of speech interaction for domestic activities, offering valuable information to guide the design and evaluation of such systems. The contributions are produced throughout various publications, forming the foundation of this dissertation. Each publication and its respective contribution are thoroughly outlined in the subsequent chapters of this dissertation.

1.4 | Document Structure

This thesis is structured into nine sections, each serving a distinct purpose in presenting and contextualizing the work. The 'Introduction' provides an overview of the dissertation's overarching topic, emphasizing its significance. The subsequent section, 'Background and Literature Overview,' delves into the development of speech-based systems, their use for domestic activities, and the challenges associated with speech, drawing insights from existing literature. Chapter three introduces the human-agent speech interaction model, inspired by models of communication and human-computer interaction frameworks, providing an overview of the impacting factors that shape the interaction between users and speech agents. Chapters four to six form the core of this thesis, encompassing the research underpinning this work. Each of these chapters consolidates two or more publications, situating them within the broader context of the dissertation. The individual publications are summarized, emphasizing findings relevant to this thesis. In some instances, additional information not included in the original publications is incorporated to ensure a comprehensive grasp of the research. Chapter four explores the utility aspect of speech interaction, focusing on methods to enhance system efficacy, particularly in recognition and error handling (**TRQ2**). Chapter five shifts the spotlight to the experiential dimensions of HASI, presenting works investigating

approaches to represent the speech agent for a better interaction experience (**TRQ3**). The sixth chapter investigates appropriate interaction styles best suited for speech interaction (**TRQ4**). In chapter seven, the results from all the research presented are combined and discussed to draw overarching conclusions applicable beyond the scope of this thesis. The first research question (**TRQ1**) is revisited and answered based on the conducted work, while the limitations of the approaches and avenues for future research are discussed. A conclusion for the presented work is given in chapter eight, summarizing the work. Lastly, the 'Publications' section contains ten original publications, upon which this dissertation is founded in their original format as they were published.

Throughout the dissertation, when discussing individual papers, details such as implementation specifics, results, and statistical analyses may be omitted for conciseness. For a more thorough exploration of this information, please refer to the corresponding individual papers. This thesis is a dissertation by publication, with many of the studies conducted collaboratively alongside colleagues who are co-authors of the included publications. In instances where the research was conducted collectively, the editorial "we" is used throughout this thesis. However, to emphasize personal contributions, the pronoun "I" is used when discussing individual work specific to this thesis.

2

Background & Literature Overview

2.1 | Speech-Based Systems

Voice interaction with computer systems has been subject of exploration and development since the early 1950s. In 1952, one of the earliest known speech recognizers, "Audrey", was developed by Bell Laboratories (Li and Mills, 2019; Meng et al., 2012). Audrey was capable of recognizing spoken digits through the use of an acoustic analyzer that converted sound waves into electrical signals (Donepudi, 2014). Later in the 1970s, researchers at the Massachusetts Institute of Technology (MIT) developed the "Put That There" system which enabled users to interact with a computer through both voice commands and gestures (Bolt, 1980). Users could issue voice commands to direct the system to perform tasks, such as moving objects on a computer screen. Even though these early systems were relatively basic and with very limited functionality, their success further showcased the potential of using voice as an interaction modality. Over the years, significant progress has been made in this field, driven by the remarkable strides in AI, machine learning, and Natural Language Processing (NLP). These technological breakthroughs have given rise to highly precise speech recognition systems (Hannun et al., 2014), empowering individuals to engage with computer systems across a wide array of settings and scenarios (Cambria and White, 2014).

There has been extensive research on speech-based systems and how people utilize them (Bentley et al., 2018a; Lovato and Piper, 2015; Clark et al., 2019; Allison, 2020). These systems have been evaluated in various domains, encompassing medicine (Austerjost et al., 2018; Miehle et al., 2017), education (Jung et al., 2019; Winkler et al., 2019), smarthomes (Bonfert et al., 2021; Lopatovska et al., 2019), automobiles (Berton et al., 2006; Schmidt and Braunger, 2018; Schmidt et al., 2019, 2020), entertainment (Allison, 2020; Allison et al., 2017), and mental well-being (Kocielnik et al., 2018; Wagener et al., 2023).

Similar to many innovative technologies, VUIs have the capacity to facilitate social interactions and exhibit forms of human-like behavior (Aeschlimann et al., 2020). Previous research has shown that when people engage in speech interactions with computers, they often perceive a form of social connection with the technology, prompting responses

similar to those directed toward humans (Nass and Brave, 2005). This phenomenon, coined as the *Computers are Social Actors* paradigm by Nass, Steuer, and Tauber, describes that users apply social conventions when interacting with a computer (Nass et al., 1994). These include courtesy (Nass et al., 1999; Bonfert et al., 2018) or attributing personalities to the speech agents (Reeves and Nass, 1996).

Generally, people have a tendency to ascribe human characteristics to non-human objects to better understand the entity's actions, a phenomenon referred to as anthropomorphism (Duffy, 2002). These attributes encompass cognitive capabilities, personality traits, and physical appearance (Zlotowski and Bartneck, 2013; Hart et al., 2013; Waytz et al., 2010). One of the primary drivers behind the growing research on voice interaction is that computers that closely resemble human characteristics are better received by users (Hart et al., 2013; Seaborn et al., 2021). Studies have demonstrated that integrating human-like qualities into computer systems significantly influences users' perceptions and attitudes toward these devices (Bonfert et al., 2021; Doyle et al., 2019; Cowan et al., 2017b). Incorporating such qualities for VUIs, which encompass realistic voices, embodiment, and agent personalities, have been shown to enhance Human-Agent Interaction (HAI) and improve user experiences (Völkel et al., 2020). A study by Seymour and Van Kleek (Seymour and Van Kleek, 2021) identified a strong correlation between the level of trust and the extent of anthropomorphism shown by users toward their voice assistants.

Human-likeness is a desirable quality when interacting with speech agents. However, if the human-like qualities closely resemble real humans but fall slightly short, it can lead to an adverse reaction known as the *"uncanny valley"* (Mori et al., 1970; Seyama and Nagayama, 2007). While realistic agents can be appealing to users (McDonnell et al., 2012; MacDorman et al., 2009), achieving this requires sufficient social responsiveness and aesthetic refinement (Hanson et al., 2005). To avoid the *uncanny valley* effect, researchers recommend maintaining consistency in realism while being deliberate in stylization (Schwind et al., 2018).

In HASI, the functionality and ease of use of the VUI, contributing to reliability, are undeniably pivotal and well-studied aspects. However, existing literature hints that the attractiveness of these systems can sometimes outweigh their reliability (Yuksel et al., 2017; Lopatovska et al., 2019). The attractiveness of virtual agents has been recognized as a significant factor in human-agent communication. Khan et al. (Khan and De Angeli, 2009) claim that users maintain a better evaluation of attractive agents regardless of their interaction. Similarly, Banakou et al. argue that agents with higher levels of attractiveness and sophistication tend to engage in more successful social interactions (Banakou et al., 2009). Another study by Khan et al. (Khan and Sutcliffe, 2014) emphasizes that attractive agents are more persuasive in influencing users' decision-making compared to unattractive agents.

To contribute to the evolving body of literature in this field, this dissertation employs human-centered design methodologies to investigate users' preferences and expectations

concerning the experiential aspects of HASI for domestic activities, with the primary objective of uncovering design considerations and requirements for developing desirable speech systems.

2.2 | Speech Interaction in Homes

The integration of speech technology into people's homes has become increasingly prevalent. Speech interaction aligns well with the growing trend of smart homes, where technology seamlessly integrates into the fabric of daily life, simplifying tasks and enriching the home environment. Given that homes function as the central hub for daily activities, they offer a distinctive setting for the implementation of new technologies. Contemporary households commonly feature home assistants like Google Home, Amazon Echo, or Apple HomePod. With the rising prevalence of these systems in homes, VAs play a more significant role as everyday digital assistants (Roslan and Ahmad, 2023; Barzilai and Rampino, 2020), assisting users in various common tasks, including smart home control, weather forecasts, music playback, and appointment scheduling, among others (Pyae and Joelsson, 2018).

Home Assistants (HAs) exhibit significant functional similarities to phone assistants, capable of addressing various use cases that individuals have traditionally relied on their phones for over the past decade. Despite these resemblances, notable differences exist in usage patterns and their integration into a person's daily routine between these two distinct voice assistant types (Bonfert et al., 2021). Speech interaction in homes possesses unique qualities. Beyond facilitating the control of smart home appliances, these systems are primarily stationary, allowing users to interact with them from a distance (Paay et al., 2022; Pradhan et al., 2018). Additionally, they are commonly shared devices within a household, serving multiple individuals, including family members or roommates.

As these devices become more ubiquitous in domestic settings, research in this area has been expanding significantly. Sciuto et al. investigated the integration of VAs into households by analyzing the logs of 75 Alexa users, totaling 278,654 voice commands (Sciuto et al., 2018). Their findings revealed variations in the usage patterns of these systems throughout the day. Notably, usage peaked in the morning before 9 am and again in the late evening, with late-night interactions being the least prevalent. The authors identified challenges such as the lack of feature discoverability and environmental awareness as significant issues with home assistants. Similarly, Bentley et al. investigated the long-term use of home assistants, seeking a deeper understanding of how users engage with these devices over an extended period (Bentley et al., 2018b). Their findings revealed that music playback queries were the most frequently used, closely followed by information retrieval and home automation. The authors also highlighted that specific types of commands exhibited variations at distinct times of the day, such as peaks in entertainment and home automation commands during the evening and requests

about weather and time in the early morning. Additionally, the study noted increased weekend usage compared to weekdays. Lopatovska et al. examined user interactions with Amazon Alexa, categorizing them as casual or leisurely, extending beyond information retrieval (Lopatovska et al., 2019). Authors witnessed an overall decrease in usage over time. Moreover, they observed that users expressed satisfaction with Alexa, even when it did not produce desirable outcomes, prioritizing the interaction experience over the quality of the outcome.

Home assistants are commonly shared devices in households, where multiple persons can interact with the same system. The shared use of technology is a common occurrence among friends and family, often indicative of the nature and strength of interpersonal relationships (Gruning and Lindley, 2016). Shared device usage commonly stems from the convenience of utilizing the same technology or may be influenced by economic considerations (Matthews et al., 2016). However, people are more likely to share devices with those they trust (Brush and Inkpen, 2007), as privacy is a significant consideration when sharing devices. Research by Hang et al. indicates that individuals carefully assess the trade-off between the potential loss of privacy and the practical advantages associated with sharing a device (Hang et al., 2012). Home assistants are challenged to cater to the diverse needs and preferences of individuals with unique characteristics simultaneously. This user base spans across various ages, cultural backgrounds, and levels of technological familiarity. Moreover, special considerations must be given to specific user groups, such as children, necessitating distinct design considerations and addressing unique risks that differ from those posed by adults (Luria et al., 2020). Additionally, the usage of these systems is intricately influenced by contextual factors (Reicherts et al., 2021). The device's location within the home, the individuals present, and the social dynamics among them all contribute to shaping the interaction with home assistants.

The more recent product category of home assistants has evolved to incorporate a screen that provides visual output. These devices, commonly referred to as *smart displays*, integrate a visual interface that could enrich the user experience by providing a multi-modal interaction, combining the benefits of voice communication with the additional context and information conveyed through visual elements (Shalini et al., 2019a; Bonfert et al., 2021). However, with the introduction of the new visual interface, new design considerations and questions emerge about how to best utilize this added modality to enhance the user experience. Research by Oh et al. highlights that people experience higher levels of social presence when a visual representation is available (Oh et al., 2018). Hernández-Trapote et al. found that users interacting with an embodied agent perceive interactions as more pleasant compared to using voice-only interfaces; however, they express greater privacy concerns (Hernández-Trapote et al., 2008a). Despite extensive research emphasizing the potential advantages of multimodal systems combining speech and visual modalities, most commercial home assistants remain voice-only devices.

Through a comprehensive investigation, this dissertation seeks to explore design possibilities for speech agents in domestic settings in order to guide the design of more

desirable systems and foster the widespread adoption of VUIs.

2.3 | Speech in Video Games

Engaging individuals of diverse ages in the comfort of their homes, playing video games stands out as a prevalent domestic activity. The immersive nature of gaming has drawn in many individuals as they seek interactive and engaging experiences. Recently, the video game industry has shown a notable interest in voice interaction technology (Carter et al., 2015; Allison et al., 2019). The intriguing and intuitive nature of voice has incentivized developers to incorporate it into video games as an input method. Thanks to advancements in speech recognition technology and the increased availability of microphones in consumer gaming devices, several game companies have been integrating voice-based services into their games (Allison et al., 2017). Moreover, the video game industry has been making significant strides towards creating more immersive gaming experiences (Cairns et al., 2014). One prominent avenue of advancement is the integration of cutting-edge technologies like Virtual Reality (VR), which has notably enriched the immersive aspects of video games (Winkler et al., 2020; Yao and Kim, 2019). Voice interaction in games has also been shown to increase immersion (Zhao et al., 2018; Lee et al., 2006) and social presence (Hicks et al., 2018), ultimately enhancing the Player Experience (PX) (Allison et al., 2019).

Voice-controlled video games originated in the 1970s when a handful of experimental games incorporated voice interaction as a novel feature (Reddy et al., 1973). One of the earliest examples was *VoiceChess*, which used a speech recognition system to support standard chess instructions (Allison et al., 2017). Over time, fueled by advancements in hardware and software, its utilization gained prominence. A notable turning point occurred with the release of more modern gaming consoles in the 2000s, where game developers began to embrace voice interaction more frequently. In 2002, Xbox introduced voice interaction through a microphone peripheral in the Xbox Live Headset, enabling players to use their voice for certain in-game actions like menu navigation and option selection. In 2006, the release of the Nintendo Wii equipped its motion-sensing controller with a built-in microphone, which players could use for voice-based input. These early systems laid the groundwork for more sophisticated voice-controlled features in games. In the years that followed, the use of voice controls in video games continued to evolve. Recently, the emergence of Large Language Models (LLMs) and generative AI technologies has significantly expanded the potential for leveraging this modality in the realm of gaming.

Along with technological developments, research on voice interaction in video games has also been growing over the past few years (Allison, 2020; Anzai et al., 2021; Allison et al., 2019; Hedeshy et al., 2022; Hong et al., 2021). In the context of video games, voice interaction can be classified into two categories: verbal and non-verbal

interactions (Allison et al., 2018). Verbal interactions, also referred to as speech interaction, involve spoken words or sentences as input for game interaction, requiring a speech recognition system to understand and respond to player commands accurately. In contrast, non-verbal voice interactions use other voice characteristics, such as volume or pitch, without relying on explicit speech recognition to facilitate player input and interaction. The advantage of non-verbal interactions lies in their ability to circumvent potential challenges associated with speech recognition, ensuring a more reliable gaming experience (Allison et al., 2019).

One of the important aspect of video games, highly valued by many players, is the opportunity for social interaction (Klimmt et al., 2010). This aspect is particularly prominent in multiplayer games, where natural language communication among players, whether they are collaborating or competing, creates a lively social atmosphere. However, this social dimension is often missing in single-player games. In such games, the social interaction is facilitated only through communication between the player and NPCs, and these interactions are commonly supported through dialog boxes controlled by the player's input. However, this method lacks the dynamic and real-time nature of interactions with other human.

In addition to the potential for introducing intuitive and innovative gameplay mechanics, voice interaction for games holds a great promise for users with disabilities (Wilcox et al., 2008). Individuals with motor control or vision impairments often face limitations in playing video games using traditional controls such as a mouse and keyboard, excluding them from this form of entertainment and social interaction (Mustaquim, 2013; Harada et al., 2011a). Moreover, speech-based games have demonstrated practical applications for speech therapy, offering opportunities for remote treatment (Ahmed et al., 2018; Lopes et al., 2016; Navarro-Newball et al., 2014).

Despite several advantages that voice interaction in video games could offer, it is mainly regarded as an optional game feature. One of the main reasons for this is problems with recognition (Petta and Woloshyn, 2001). While the technology has been improving extensively in recent years, speech-based systems are still susceptible to recognition failures, and creating a seamless experience for users is still highly challenging (Kinoshita et al., 2020). Given the capacity to bypass recognition challenges, researchers have previously encouraged designers to explore non-verbal forms of voice interaction, as they have shown to provide enjoyable game experiences (Sporka et al., 2006; Vieira et al., 2014; Parker and Heerema, 2008; Harada et al., 2011b; Allison et al., 2018, 2019). However, these games typically offer players limited control and often feature relatively simple mechanics. Consequently, many players tend to discontinue playing such games after only a few uses, seeking more engaging and complex gaming experiences. Additionally, it is important to acknowledge that video games are mainly goal-oriented activities with varying challenges, and players derive enjoyment when they actively work towards these goals (Reid, 2012; Juul, 2007). If the challenge is right, the players are in a state of flow (Csikszentmihalyi, 1990). When problems related to speech recognition arise, they introduce an additional

layer of challenge alongside the existing game dynamics, hindering players from achieving their goals and maintaining a state of flow (Csikszentmihalyi, 1990). Consequently, players often become frustrated, leading to their eventual abandonment of the game (Dow et al., 2007).

Part of the objectives of this dissertation is to tackle some of the challenges mentioned above to facilitate the smooth integration of speech interaction in video games with an aim to provide players with a more efficient and immersive experience. Using speech in single-player games to interact with NPCs can offer a unique potential to enhance player experience (Allison et al., 2019; Allison, 2020). This feature can create a more dynamic and lifelike virtual world where players feel actively engaged with the game environment. By enabling players to interact with NPCs vocally, the gameplay could become more personalized, allowing for natural and fluid conversations that mirror real-life interactions. Speech interaction has the potential to deepen emotional engagement (Bonfert et al., 2021; Chen et al., 2022; McLean et al., 2021). Hearing NPCs respond directly to your speech input can evoke a stronger emotional connection and a sense of empathy with the virtual characters (Nass and Brave, 2005). This could lead to more meaningful and memorable in-game experiences as players form stronger bonds with the characters and their stories.

2.4 | Challenges and Complications with Speech

While the literature on speech interaction emphasizes its substantial potential in HCI, it is essential to acknowledge the challenges associated with adopting speech as an interaction modality. Understanding and addressing these challenges is crucial in the pursuit of designing desirable speech-based systems.

Developing speech-based systems requires techniques, methodologies, and development tools that enable flexible and dynamic interactions to accommodate the diverse needs of various user groups and contextual settings (Turunen et al., 2005). This complexity is further magnified when targeting a global market, necessitating the consideration of various languages, accents, and dialects to ensure effective and reliable recognition systems (Pyae and Scifleet, 2018). Consequently, creating a satisfying experience with speech-based systems becomes difficult with inherent complexities and challenges. Technological advancements and the accessibility of open-source speech libraries have somewhat streamlined the development process, making it more feasible and efficient. Moreover, advances in NLP have enhanced the sophistication and reliability of speech systems, thereby enabling more robust language understanding and generation capabilities (Cambria and White, 2014). Nevertheless, challenges with speech interactions persist. Researchers and developers continue to grapple with these issues in their pursuit of delivering desirable user experiences.

From a technical stance, there are still several challenges concerning the smooth functionality of speech. Researchers believe that technical and functional limitations are

still one of the main reasons for user frustration and their skepticism towards VUIs (Pearl, 2016; Suhm et al., 2001). Speech systems are still limited with regard to recognition accuracy, specifically for non-native speakers and people with unique accents and dialects (Pyae and Scifleet, 2018). When the system does not correctly recognize users' speech input, user frustration, disappointment, and dissatisfaction arise, often leading to a lack of progress or the inability to complete tasks (Rotaru et al., 2005; Bohus and Rudnicky, 2005; Swerts et al., 2000; Bentley et al., 2018b; Cowan et al., 2017c; Luger and Sellen, 2016). When systems fail to recognize the user's input, error handling methods are often used as fallback strategies to redirect users (Li et al., 2018). These include asking the user to repeat the command, redirecting the user to the tasks the system can support, or presenting user options to correct their commands (Pappu and Rudnicky, 2014; Bohus and Rudnicky, 2005; Li et al., 2018). If the error handling is done well, it will not derail users, and the system can get them back on track (Pearl, 2016). However, such strategies often fail to provide users with a desirable outcome (Wang et al., 2020). Furthermore, most VUIs can only support a limited set of tasks, and this constrained functionality is another reason for the users' resistance to using such systems (Jentsch et al., 2019). Additionally, even though VUIs can enhance the accessibility of a system, they can also raise certain accessibility concerns for people with speech impairments or language difficulties. Due to such limitations, research indicates that users lack trust in speech systems for complex tasks and generally consider interactions with speech agents as secondary tasks (Luger and Sellen, 2016). Nonetheless, the emergence of NLPs and generative AI technologies such as ChatGPT has ushered in notable improvements (Bubeck et al., 2023).

One of the most significant barriers to users' acceptance of VUIs is the issue of privacy (Cha et al., 2020; Malkin et al., 2019; Miksik et al., 2020; Tabassum et al., 2019a). Previous research has shown that many individuals hesitate to embrace speech-based systems, especially home assistants, due to concerns about privacy and a lack of trust in the companies behind these devices (Lau et al., 2018). Users express discomfort with permanently preserving user recordings and strongly oppose using their speech data by third parties, particularly when it involves children and guests (Malkin et al., 2019).

Another prominent challenge with speech-based systems is their restricted use case. These systems face environmental limitations, particularly in shared or public spaces, where users may be reluctant to utilize speech interaction due to concerns about disturbing others or audibly revealing personal information (Pearl, 2016). People might also experience self-consciousness when using speech commands to communicate with technology, especially in public or shared spaces (Pearl, 2016). Engaging in a conversation with a computer may induce a sense of social awkwardness, as users find themselves essentially conversing with an inanimate object.

Addressing issues with speech systems goes beyond relying solely on software or hardware advancements. In many instances, the user's commands may be ambiguous, personalized, or complicated, making the system unable to understand them. Overcoming challenges in speech interaction demands a collaborative effort, with users playing

a crucial role in articulating commands the system can accurately recognize. When interacting with a VUI, users tend to adjust their speech patterns, anticipating that these systems might not fully grasp natural language. Observable adaptations include slowing down speech pace, rephrasing command sentences, and adjusting physical positions relative to the system (Jentsch et al., 2019).

On the experiential side, one of the main reasons that people find interaction with speech-based systems unsatisfactory or disappointing is that these systems do not fulfill the user's expectations as an interlocutor (Jentsch et al., 2019; Luger and Sellen, 2016; Murad and Munteanu, 2019; Doyle et al., 2019). Although anthropomorphization of speech agents has shown to enhance the UX, often by adding human-like qualities, these systems also face the challenge of giving rise to unrealistically high expectations regarding the system's intelligence, capabilities, and conversational fluidity (Luria et al., 2019; Foner, 1993; Murad and Munteanu, 2019; Doyle et al., 2019). They often fall short of meeting people's expectations (Sheehan et al., 2020), frequently being perceived as robotic, cold, socially awkward, untrustworthy, and incompetent (Feine et al., 2019; Go and Sundar, 2019; Shin et al., 2023; Zargham et al., 2023c). Studies indicate that users push the limits of VUIs by posing diverse questions, often surpassing the agent's abilities (Luger and Sellen, 2016; Lovato and Piper, 2015). This pattern also extends to children, as observed in a study by Lovato et al., where children predominantly asked Siri personal questions to test the agent's potential (Lovato and Piper, 2015). When initial user expectations are not met, it results in disappointment and an overall negative user experience (Porcheron et al., 2018).

The majority of VUIs operate as voice-only systems, limiting communication to a single channel. However, effective communication between people encompasses more than verbal exchanges alone. Non-verbal elements, such as facial expressions and body movements, play a crucial role in conveying information. Visual cues in non-verbal communication allow individuals to express more than just the explicit meaning of their messages; emotions, current mood, and aspects of one's personality can be communicated through these cues (Castillo et al., 2018). By acknowledging these non-verbal factors, a more comprehensive understanding of communication emerges, capturing both the semantic content of the message and the rich layers of information conveyed through non-verbal means. Introducing other communication channels, such as a visual dimension, also holds value in enhancing accessibility for individuals with hearing impairments (Massaro et al., 1999; Virkkunen et al., 2018).

Murad et al. highlight that a common source of user frustration in HASI is the perceived lack of agency and control, underscoring the importance of incorporating user control and freedom into speech interfaces (Murad et al., 2018). Existing research emphasizes that the lack of flexibility hinders productivity and satisfaction (Molnar and Kletke, 1996). Most VUIs tend to take a one-size-fits-all approach, neglecting the potential benefits of adapting to user preferences. Due to the individual differences in preferences for a desired system, systematic adaptation of these systems to the user is challenging (Völkel et al., 2020). Despite these challenges, various studies have shown

that users prefer VUIs that can adapt to their preferences and background (Cowan et al., 2015, 2016; Dahlbäck et al., 2007; Lee and Nass, 2003; Braun et al., 2019). The inclusion of customization and personalization features in VUIs offer users greater control over their interactions, potentially enhancing user satisfaction and improving performance (Molnar and Kletke, 1996; Murad et al., 2018; Choi et al., 2020; Wolters et al., 2009). *Customization* describes the extent to which technology or service can be modified to comply with user preferences (Hsieh and Chen, 2016; Teng, 2010). On the other hand, *personalization* refers to automatic adaptation to users' needs based on observed behaviors (Hsieh and Chen, 2016). These customization and personalization features can be particularly beneficial for individuals with special needs (Abdolrahmani et al., 2020, 2018).

Indeed, the one-size-fits-all approach adopted by VUIs has not only technical implications but also carries social concerns, particularly concerning stereotyping and the inadvertent imposition of social and political agendas. Gender stereotypes, in particular, have been a recurring issue with VUIs, drawing attention from researchers concerned about the potential reinforcement of stereotypical gender scripts (Zdenek, 2007; Sutton, 2020; Danielescu, 2020). Notably, a significant number of voice assistants still default to a female voice in many countries, a practice that may contribute to the amplification of gender stereotypes (Hwang et al., 2019). The concern extends beyond gender, as other forms of bias, such as ethnic stereotypes, can also be inadvertently perpetuated. Designers must conscientiously navigate the ethical implications associated with agent characteristics, ensuring they do not reinforce harmful agendas.

Ultimately, technical limitations such as restricted functionality, accuracy issues in recognition, inadequate fallback strategies, limited task support, concerns about privacy and security, and absence of communication channels beyond voice, coupled with experiential challenges like the inability to act as a proficient communication partner, restricted user agency and control, users' heightened expectations for an effective interlocutor, and skepticism towards the system's technological capabilities form the primary reasons behind users' reluctance to adopt VUIs. Acknowledging these challenges, this dissertation explores innovative solutions and strategies to address some of the functional and experiential concerns associated with VUIs.

3

Speech Interaction Framework

Research on VUIs has been growing extensively in recent years, and VAs are becoming increasingly prominent in people's daily lives (Bonfert et al., 2021). Despite the notable progress in this domain, a critical gap persists — the absence of a dedicated interaction model designed to systematically account for the various components involved in human-agent communication. The current trajectory of VUI research has primarily focused on the broader aspects of speech recognition and natural language understanding (Zargham et al., 2022b). While these are foundational elements, they represent just one layer of the complex interaction between users and speech systems. The absence of a specific interaction model tailored to speech agents hinders our ability to optimize user experience. Meanwhile, the community offers valuable domain-specific heuristics (Langevin et al., 2021) and frameworks (e.g. (Yeh et al., 2022; Mildner et al., 2024)).

Traditionally, researchers and practitioners have sought to explain the dynamics of speech interaction by borrowing models from communication or human-computer interaction (Langevin et al., 2021; Murad et al., 2021). However, these models often fail to capture the nuances inherent in spoken interactions, leading to an incomplete understanding of the underlying processes. Further complicating this issue is the constant evolution of technology, which continuously shapes how people interact with digital technologies. As a result, contemporary models may be constructed around technologies that change interaction principles or modalities, thus losing effectiveness in assisting practitioners in their work.

In response to these limitations, we propose a new model of interaction specifically tailored to speech. This model is conceived to align more closely with cutting-edge research in VUIs, providing a more accurate and comprehensive representation of the multifaceted nature of contemporary speech interactions. It is important to note that the model proposed in this work is explicitly designed for VUIs, catering to the specific elements and dynamics of speech-based interaction. Therefore, the model might not fully apply to the overarching category of CUIs. CUIs encompass a broader range of conversational systems, including VUIs as well as non-voice-based systems like chatbots.

Therefore, it might lack the versatility to accommodate other conversational systems' diverse modalities and interaction patterns.

Our proposed speech interaction model contains not only the traditional elements of sender, message, and receiver but also incorporates the dynamic interplay of context, user preferences, and the evolving capabilities of speech agents. The model draws inspiration from traditional frameworks found in communication theories and HCI. By recognizing the comprehensive nature of communication in VUIs, we aim to provide a more detailed understanding of how users engage with speech agents in real-world situations. To the best of our knowledge, the existing research landscape has not yet provided a structured framework that accounts for the constitutive components of speech interaction with agents.

3.1 | Models of communication

Communication is typically understood as the process of transmitting information, a concept present in diverse disciplines, including psychology, sociology, engineering, technology, and artificial intelligence (Fedaghi et al., 2009). In the field of communication studies, various models have been proposed to explain the process of communication, with the aim of providing a concise overview of its components. These models have been instrumental in advancing our understanding of communication while guiding research and development. Nevertheless, they are often criticized as they can be perceived as oversimplified and might leave out critical components (Kimmel, 2020). Typically depicted diagrammatically, these models share similar fundamental components, involving a sender encoding a message transmitted through a channel to a receiver, which decodes the message and provides a form of feedback.

Hartley's pioneering work introduced the quantification of "signals as means to convey information," laying the groundwork for information theory (Hartley, 1928). Hartley's rule quantifies the maximum rate of information transmission over a communication channel. In the broader context of communication models and theories, Hartley's work expands our understanding of the fundamental constraints and parameters influencing information flow through communication channels.

Correspondingly, communication models have been classified as either linear or non-linear models (Narula, 2006). Linear models focus on the one-way flow of information from a sender to a receiver. In contrast, non-linear models encompass a broader perspective, acknowledging multi-directional interactions in the communication process. In the following, we explore different classifications of communication models while discussing some of the better-known models.

3.1.1 | Transmission Models

Linear communication models, also referred to as *transmission models* or *action models* of communication, are unidirectional models that describe the process from a sender

to a receiver (Ellis and McClintock, 1990). An example of such a model is Shannon and Weaver's model of communication (Shannon, 1948), one of the key models in communication studies (Fiske, 2010).

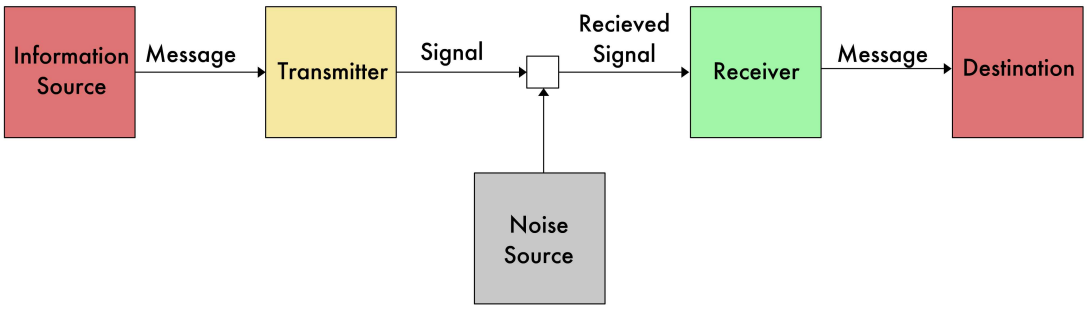


Figure 3.1: The Shannon and Weaver's model of communication.

This model systematically breaks the message flow into five fundamental components: a source, a transmitter, a channel, a receiver, and a destination (see Figure 3.1). The source produces the original message, and the transmitter encodes it into a signal, which traverses through a channel. At the receiving end, the message is decoded back from the signal and made available to the destination. A sixth component, noise, represents any disruptive factor that may interfere with the message's journey along the channel, potentially causing the received signal to deviate from the one sent. The Shannon-Weaver model is recognized as a foundational concept in Communication Studies (Fiske, 2010). This model originated from the Bell Telephone Laboratories in the US. Developed to enhance communication channel efficiency, the model was primarily concerned with maximizing information transfer through channels such as telephone cables and radio waves. The theory provided a systematic approach to optimizing information transmission and measuring channel capacity. While rooted in engineering and mathematics, the designers claim the theory's broad applicability to human communication. Subsequent models have rejected this linear approach for failing to account for the role of feedback in the communication process.

3.1.2 | Interaction Models

To address the limitations of linear models, experts in communication studies presented *Interaction Models*. *Interaction Models* introduce a feedback loop, allowing the listener to respond to the speaker by expressing their opinion or seeking clarification (Littlejohn and Foss, 2009). This two-way communication process involves a dynamic exchange of messages, making it more representative of conversations. It views communication as an action-reaction sequence, where the communicators take turns sending and receiving messages.

An example of the *interaction model* is the Schramm model (Schramm, 1997). According to this model, communication initiates when a source generates an idea and conveys it as a message. The message is then transmitted to a destination, where it undergoes decoding and interpretation for comprehension. In response, the destination formulates its own idea, encodes it into a message, and sends it back as feedback. Both the source and the destination engage in encoding, interpreting, and decoding (see Figure 3.2). Unlike linear models, Schramm's model does not view the audience as passive recipients but recognizes them as active participants, fostering a more interactive communication exchange.

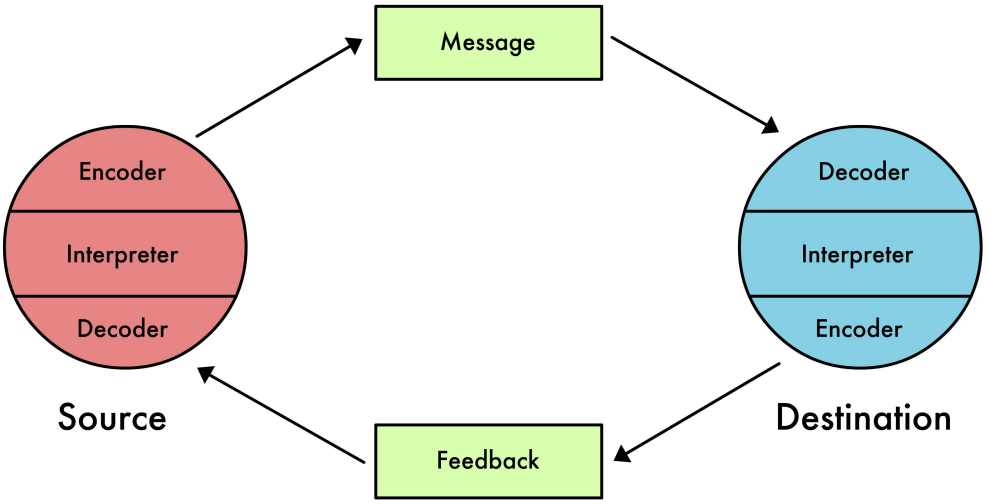


Figure 3.2: The Schramm interaction model of communication.

In contrast to the *transmission model*, which focuses on successful message transmission and reception, the *interaction model* is more concerned with the communication process itself. In this model, the effectiveness or ineffectiveness of communication is not solely determined by the successful transmission and reception of a single message. Schramm argues that the communication process is also influenced by participants' past experiences (Schramm, 1997). Communication failure may correspondingly occur if the message extends beyond the receiver's scope of experience, impeding their ability to decode and connect it to the intended message from the sender. Additionally, failures in communication may occur due to external noise and errors in decoding and encoding. Schramm later updated his model to highlight the importance of participant relationships in determining communication goals and roles. Unlike Shannon and Weaver's model, Schramm's communication model does not explicitly incorporate noise. Instead, it focuses primarily on the circular communication process and the behaviors of both senders and receivers.

3.1.3 | Transaction Models

Even though the added feedback help make the *interaction model* a more comprehensive representation of the communication process, *transaction models* take our understanding of the interaction a step further by proposing simultaneous sending and responding, emphasizing the interactive and concurrent sharing of ideas and feelings (Hamilton, 2016; Kastberg, 2019). Such models acknowledge that communication is not strictly a circular process, and the sending and receiving processes occur simultaneously, accounting for adaptability mid-communication, enabling adjustments based on real-time feedback from communication partners (see Figure 3.3). Additionally, these models consider the simultaneous exchange of non-verbal feedback, including body language, gestures, and facial expressions during the communication process (see Figure 3.3).

Dean Barnlund, one of the early proponents of a transactional communication model (Barnlund, 2017), argued that communication is essentially "the production of meaning rather than the production of messages." Barnlund's model emphasizes the shared responsibility of both parties in creating meaning, with each party influencing and being influenced by the other through a series of private, public, and behavioral cues (Hamilton, 2016). Public cues include factors in the physical or social environment that are available for meaningful interpretation by anyone in their presence. Private cues refer to interpretable factors that, similar to public cues, emerge and stay beyond the control of the communicators but are solely accessible to the individual. Examples include elements such as cognition, sensations, and emotions. Behavioral cues are those interpretable factors, both non-verbal (i.e., gestures, body language, and facial expressions) and verbal (i.e., written or spoken information), entirely controlled by the communicators (Watson and Hill, 2015).

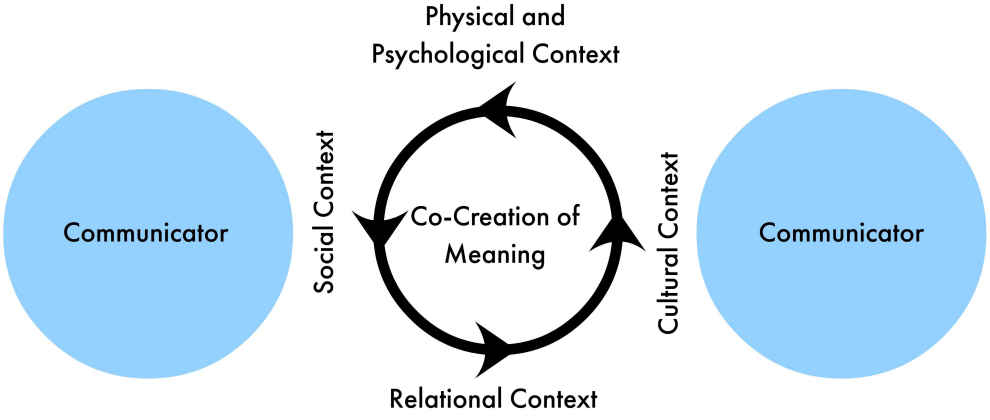


Figure 3.3: The transaction model of communication.

The primary objective of *transaction models* is to reduce uncertainty and achieve a shared understanding. This model provides a more nuanced understanding of context than the *interaction model*. The *interaction model* focuses on physical and psychological

influences impacting message transmission and reception. The *transaction model* sees communication as a force shaping realities beyond individual interactions and accounts for social, relational, and cultural contexts that affect communication encounters (Jones, 2016). Social context involves the rules and norms guiding communication, learned through community socialization. Examples of such social rules are not interrupting people, speaking politely, or not lying. Rules are often reiterated, and failure to adhere to them may result in consequences. On the other hand, norms are social conventions that we learn through observation, practice, and trial and error. Relational context considers the interpersonal history and the type of relationship between communicators, influencing communication dynamics. Initial interactions follow established norms, but bending or breaking social norms becomes easier with an established relational context. For instance, the level of formality when communicating with strangers differs from that established with long-term friends. Cultural context encompasses aspects such as nationality or ethnicity. Individuals possess multiple cultural identities influencing communication. People with historically marginalized identities are regularly aware of their cultural influence, affecting how others communicate with them. Conversely, those with dominant identities may rarely consider cultural identities' role in their communication. The importance of context stems from the dynamic and multifaceted nature of human interaction. Existing literature highlighted the significant role of contextual elements in human-agent speech interaction and their influence on the user and the system, as they can facilitate or disrupt communication (Zargham et al., 2022c, 2023b). The *transaction model* emphasizes that communication is the process of creating relationships, forming intercultural alliances, shaping self-concepts, and engaging in dialogue to build communities. In this approach, individuals are not labeled as senders or receivers but, to highlight their agency, are acknowledged as communicators.

Ultimately, the *transaction model* stands out as the most comprehensive communication model, encompassing elements from previous models while introducing new aspects such as the significance of context and the dynamic roles of sender and receiver (Hamilton, 2016; Jones, 2016). The discipline of communication studies extends beyond human interaction, encompassing communication with non-human entities, including computers. In the following section, we shift our focus to human-computer interaction frameworks to deepen our understanding of how humans interact with computers, ultimately working towards establishing a comprehensive model for human-agent speech interaction.

3.2 | Human-Computer Interaction Frameworks

Along with the advancements in technology, the field of HCI gained more prominence, necessitating a deeper understanding of how humans interact with computer systems. To support this process, several models and frameworks were introduced to shape the design and evaluation of interactive systems. These models played a crucial role in shaping HCI

into a structured discipline by offering conceptual frameworks encapsulating key elements of the interaction process to explore, analyze, and improve the dynamics between users and technology. Over time, these models have evolved to keep pace with technological advances and incorporate emerging concepts and principles.

The following discusses some of the more prominent HCI models and frameworks, and discuss their applicability to speech interaction with agents.

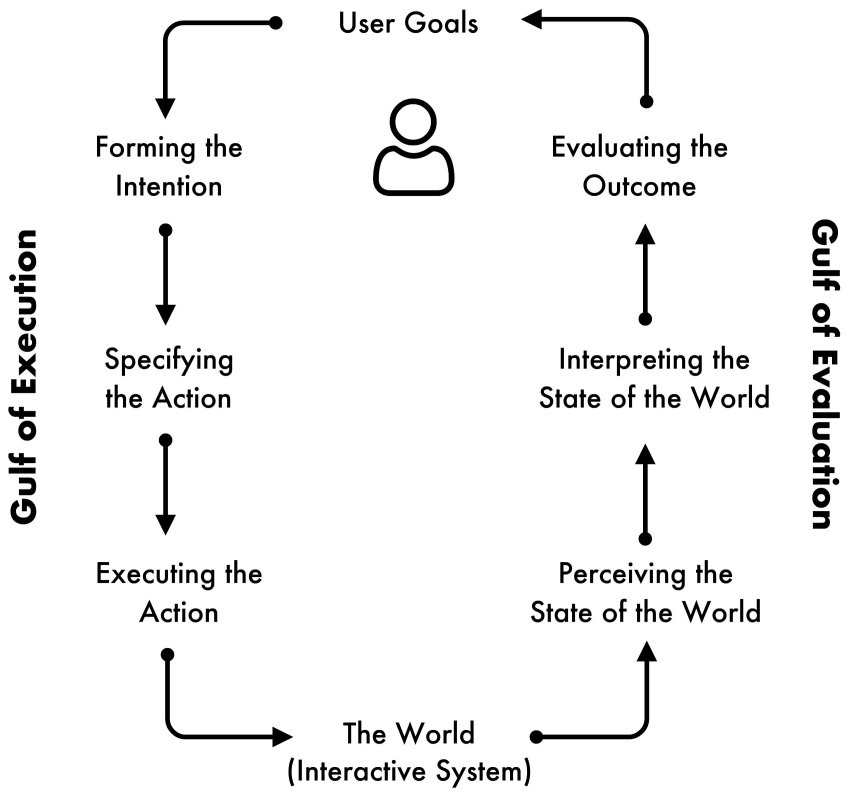


Figure 3.4: Norman's action cycle and the seven stages of action.

3.2.1 | Norman's Action Cycle

In his book "The Design of Everyday Things," Don Norman introduces "Norman's Action Cycle," an interaction model that revolves around two key components: execution and evaluation (Norman, 1988). This model portrays interaction as a cyclical process, emphasizing the iterative nature of human-computer interaction. The cycle begins with the user establishing a goal, forming the intention, specifying the action sequence, and executing the action at the interface (see Figure 3.4). Subsequently, the user perceives the system state, interprets the system state, and finally evaluates it with respect to the initial goals and intentions. In addition to the action cycle, Norman introduced the concepts of the "Gulf of Execution" and the "Gulf of Evaluation" to address potential challenges in user interactions. The Gulf of Execution refers to the disparity between a user's intentions

and the system's actions, emphasizing the ease or difficulty of translating goals into executable actions. On the other hand, the Gulf of Evaluation pertains to the cognitive gap between the system's output and the user's comprehension, addressing the ease or difficulty of understanding the system's feedback. These concepts underscore the importance of reducing the gaps for an optimal design. Norman acknowledges that his model is an approximate representation, not claiming to be a comprehensive psychological theory. Nevertheless, it provides valuable insights into the human-computer interaction process.

3.2.2 | Abowd and Beale framework

Building on Norman's model, Abowd and Beale proposed the "General Interaction Framework" (Abowd and Beale, 1991). This framework outlines the interaction process between system and user components through an interface's input and output components (see Figure 3.5). This framework mirrors a cyclic sequence wherein a user articulates a task, the system executes and presents the task, and the user observes the results, allowing them to formulate subsequent tasks (Mitchell et al., 1996). This framework provides an insightful extension to Norman's model, offering a more detailed exploration of the interactive process's bidirectional communication between users and systems.

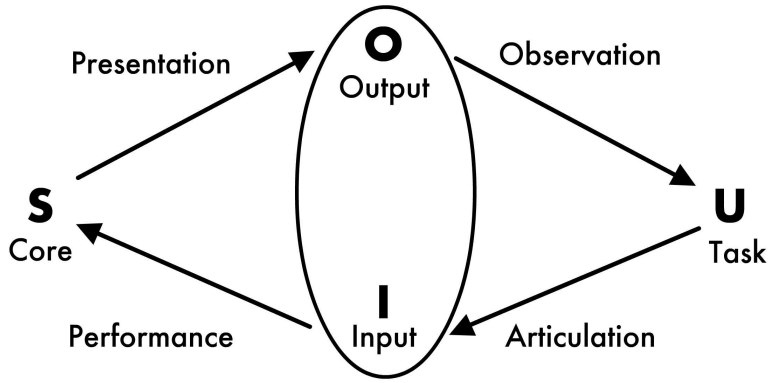


Figure 3.5: Abowd and Beale's general interaction framework.

The interactive cycle contains four distinct stages symbolized by arrows in the diagram. The cycle begins with the *articulation* stage, where the user formulates a goal and corresponding tasks to achieve that goal. This is followed by the *performance* stage, during which the identified tasks are translated into operations to be executed by the system. Subsequently, the system changes state and communicates this altered state through the *presentation* stage. Lastly, the *observation* stage involves the user evaluating the outcomes by observing the presented output. Each stage involves a translation process. Abowd and Beale's general interaction framework is argued to extend beyond computer systems, showcasing its applicability to a broader spectrum of interactive contexts.

3.2.3 | Nigay's Model

Nigay's fundamental HCI model breaks down the human-computer interaction process into two main entities: the user and the computer (Nigay, 1994). On the computer side, there are two components: the interface and the functional core. The interface serves the purpose of establishing a connection between the user and the computer's functional core by integrating software and hardware (see Figure 3.6). This model portrays the interface as a mediator facilitating the relationship between the two entities (Chignell and Hancock, 1988). However, the model is overly simplistic and fails to fully capture the complexity of the processes unfolding between the user and the computer.

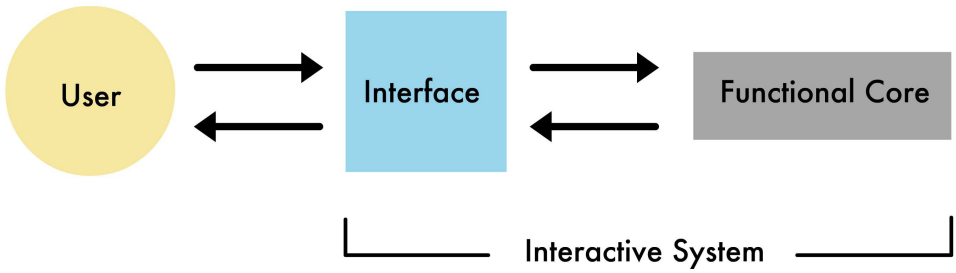


Figure 3.6: Nigay's fundamental HCI model.

Addressing this limitation, this model was later extended to the Pipe-Lines Model (Nigay and Coutaz, 1997), which emphasizes the functional equivalence between user and computer system transformations, encompassing interpretation and rendering functions (see Figure 3.7). This model positions the user as the controller, initiating requests processed by the computer system, which then responds, depicting a seemingly one-way interaction. However, this neglects the collaborative nature of the communication.

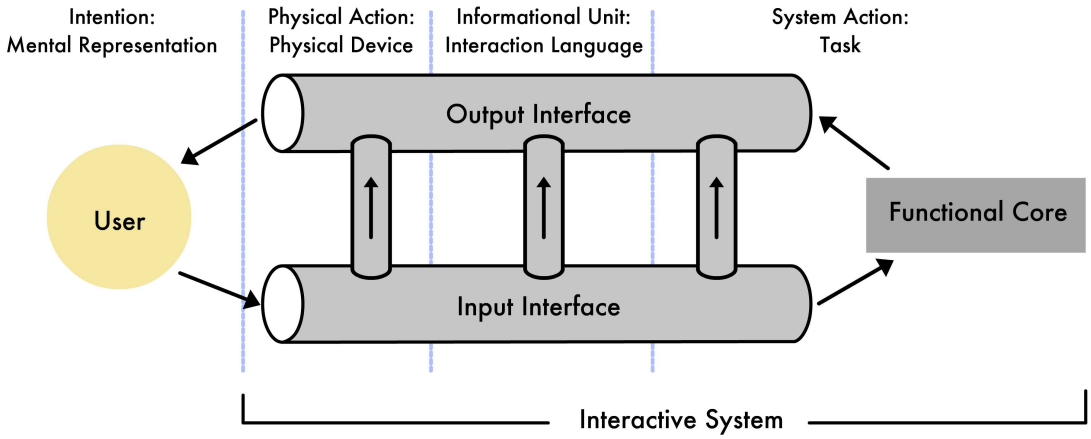


Figure 3.7: The Pipe-Lines model which extends Nigay's fundamental HCI model.

3.2.4 | Schomaker's Model

Schomaker proposed a model for identifying basic processes in multimodal human-computer interaction (Schomaker, 1995). This model outlines two separate cycles or loops: intrinsic feedback, resembling eye-hand coordination, and extrinsic feedback, imposed by the computer. The model assumes at least two agents—human and machine—which are physically separated but can exchange information through various channels (see Figure 3.8).

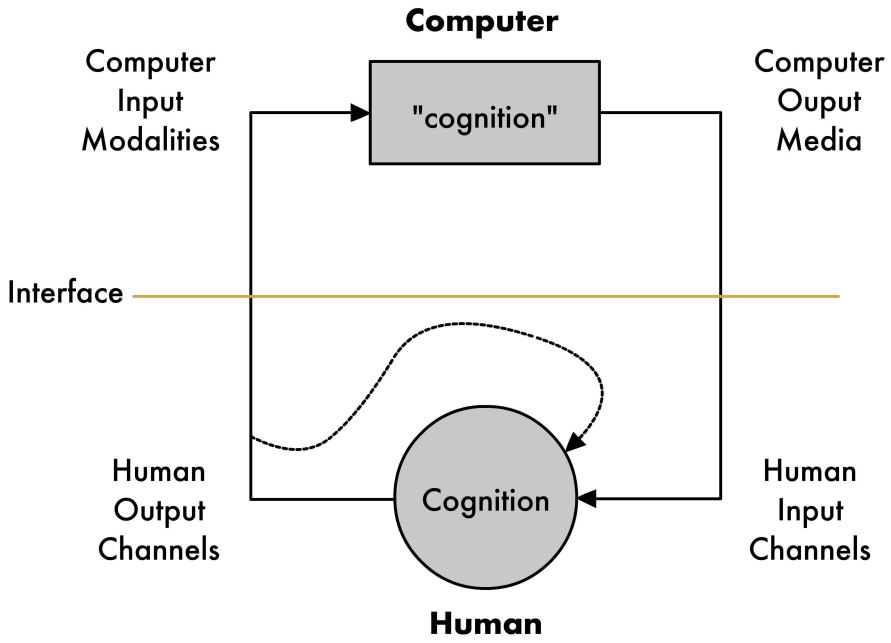


Figure 3.8: Schomaker's model for processes in multimodal human-computer interaction.

On the human side, there are two basic processes: perception and control. The perceptive process contains human input channels (HIC) and computer output media (COM), while the control process includes human output channels (HOC) and computer input modalities (CIM). Within both agents, a cognitive or computational component that processes incoming input information and prepares the output can be identified. The model emphasizes a functional and cognitive parallel between the computer and the human, portraying HCI as an event sequence where the human controls and manipulates the computer, and processes of perception and control occur at the human level. In this model, the interface is not seen as solely a computer element.

3.2.5 | Summary

Exploring previous HCI models has provided a foundational understanding of how humans interact with computer systems, shedding light on the dynamics between users and technology. However, while each model holds value, certain elements present in one

model may be absent in others, suggesting a gap for a comprehensive and conclusive model encompassing all discussed aspects. Furthermore, while the discussed models provide valuable insights, especially in conventional scenarios, they are designed for broader contexts and may fall short of fully grasping the intricacies of speech-based interactions. Creating a specialized model tailored to address the distinctive challenges and opportunities presented by speech interaction can be highly valuable.

3.3 | Partner-based Speech Interaction

The evolution of speech systems has been marked by a transformative journey from basic, task-oriented functionalities to highly intelligent and capable systems facilitated by advancements in machine learning, AI, and NLP (Zargham et al., 2023b). Initially, emphasis was placed on functionality and recognition, ensuring tasks were properly executed. However, recent developments have witnessed a notable shift towards refining these systems' representation and interaction styles as they become more prominent in people's daily lives (Zargham et al., 2022a). These systems can tackle more intricate tasks and engage in complex conversations. Modern speech systems incorporate anthropomorphic features, employing human-like voices and often featuring specific character traits (Bonfert et al., 2021).

The diverse domains and use cases of speech systems significantly influence their design. Depending on their application, while some systems mainly focus on functionality and task completion, concentrating on core efficacy, others also prioritize experiential aspects, such as systems designed to facilitate social interactions. For instance, a banking assistant prioritizes efficiency, focusing on quick and accurate financial transactions. On the other hand, consider a home assistant that supports a variety of tasks and can engage with users in casual conversations, share jokes, or perform smart home tasks. These systems should pay more attention to their hedonic dimensions as task criteria often could be less serious and time-critical.

Technological progress empowers designers and developers to create speech systems that authentically emulate human characteristics, enhancing natural behavior. In light of these developments, there is a perceptible shift in people's attitudes toward speech agents. Once perceived as limited computer programs for basic tasks, these systems can now be regarded as companions or assistants with unique features and traits (Pradhan et al., 2019). This transformation is particularly evident in speech systems with virtual assistants, like home or smartphone assistants, where a more intimate and interactive relationship is fostered. This form of speech interaction is referred to "partner-based speech interaction" (Peña et al., 2023; Doyle et al., 2023).

Earlier frameworks distinguished the process of human-agent communication from human-human communication, as computer systems had limited capabilities (Doran et al., 2003). With technological advancements, computer systems interacting with users

in a natural language become better at replicating human characteristics. This raises the argument that modeling human-agent communication based on human-human communication may be appropriate (Doran et al., 2003; McDaid, 2009). However, this notion is subject to debate among researchers and practitioners. While some advocate that this alignment could enhance authenticity and effectiveness, some argue that the distinct characteristics of human-agent interaction require tailored approaches. As discussed earlier, when examining existing communication models, the transaction model stands out as the most comprehensive communication model, as it acknowledges the significant impact of contextual elements and the communicators' dynamic roles and individual traits.

Other models tend to be predominantly turn-based, which limits the dynamics of the interaction. Furthermore, traditional frameworks typically depict a unidirectional interaction where users initiate and conclude the interaction upon receiving the system's output. In contemporary contexts, however, technological advancements have empowered computers to be more intelligent and context-aware, allowing them the possibility to be the one who initiates the interaction (Zargham et al., 2022c). Such systems are referred to as proactive systems (Reicherts et al., 2021). These advancements challenge the conventional depiction of user-driven interactions. Additionally, most existing frameworks serve as high-level or meta models, accommodating a broad spectrum of interaction types. In pursuit of universality, these frameworks often lack detailed consideration of specific nuances.

Given these considerations, a more specialized and comprehensive framework explicitly tailored for speech interaction with artificial agents can be beneficial.

3.3.1 | Human-Agent Speech Interaction Model

We propose a model of interaction, highlighting the influencing factors of HASI (see Figure 3.9). The user and the system (VUI) form the core components of the interaction process. The model is structured across three primary layers: the *interaction layer*, the *traits layer*, and the *contextual layer*. These layers are not isolated and influence one another. The user and system dynamics are linked with the broader contextual conditions, collectively shaping the interaction.

3.3.2 | Interaction Layer

The *interaction layer* is the central layer where the actual communication between the user and the system takes place. It involves the 'user action' and 'system action.' The interaction style, which contains elements such as the phrasing of the actions, the types of feedback provided, and the timing of the interaction, is a critical element within this layer, dictating how the user and the system communicate and guide them through the interaction process.

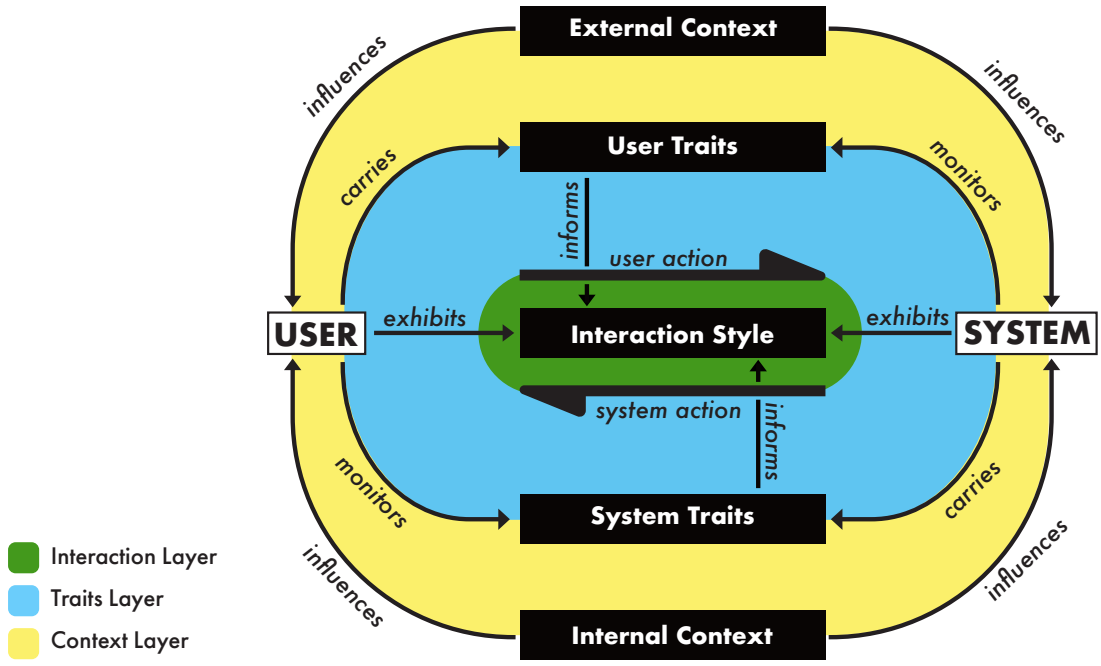


Figure 3.9: The human-agent speech interaction model highlighting factors influencing the interaction process.

3.3.3 | Traits Layer

Surrounding the *interaction layer* is the *traits layer*, which consists of 'User Traits' and 'System Traits.' The user traits encompass their personality and preferences, current emotional state, cognitive and physical abilities, as well as their cultural background and ethnicity (Zargham et al., 2022c). Furthermore, users' insights about the system, including their interaction knowledge and their mental model about the system, could also influence their interaction with the system (Schramm, 1997). On the other side is the system traits. The system traits include aspects such as the utility and the representation of the system. Utility contains functional elements such as speech recognition and error handling (efficacy), accessibility features, and privacy and security. These components form the backbone of the VUI, ensuring its efficiency and reliability. The representation includes the agent's characteristics, such as voice, embodiment, or personality, contributing to users' perceptions and interactions with the system. The agent's voice quality, visual appearance, and behavioral attributes all fall under this category, influencing how users engage with and trust the system. These traits are essential parts of the interaction process as they shape the nature and quality of the communication.

3.3.4 | Context Layer

As discussed earlier with the transaction model, one needs to account for contextual elements that might affect the communication encounters (Jones, 2016). The *contextual layer* is the external layer that could influence both the system and the user. It is further divided into 'Internal Context' and 'External Context.' The internal context encompasses elements shared by both the system and the user, including the topic of interaction and the history of past interactions. Both parties accumulate a history of interactions, shaping their current interaction patterns and expectations. For users, past experiences with the system influence their levels of trust, confidence, and satisfaction. Similarly, the system's memory of past interactions enables it to personalize and adapt responses based on user preferences and previous interactions, enhancing the user experience by anticipating their needs. The external context includes broader contextual factors beyond the immediate user and system dynamics. These contextual elements contain interpersonal, socio-cultural, and environmental aspects, which can all play an important role in how the interaction process unfolds (Zargham et al., 2022c). Interpersonal conditions, such as the presence of others during interactions and the type of relationship between those present (e.g., family members, friends, colleagues), can directly impact the interaction with the system. For instance, in a family setting, interactions with a speech-based system may involve collaborative decision-making or negotiation among family members, whereas interactions in a professional setting may prioritize efficiency and task completion. Moreover, socio-cultural factors, such as societal norms and standards, can influence interaction. Cultural expectations regarding politeness, formality, attractiveness, and communication styles can impact user behavior and preferences when engaging with the system. For instance, in cultures that value direct communication, users may prefer concise and straightforward interactions, whereas in cultures that prioritize indirect communication, users may expect more nuanced and contextualized responses from the system. Environmental factors include ambient noise levels, physical surroundings, ongoing activities, and the proximity of the user and the system. For instance, a user engaging with a voice assistant in a busy urban environment may face challenges due to high noise levels impacting speech recognition accuracy. Conversely, a user in a quiet home setting may experience a more seamless interaction.

3.3.5 | Relations Between the Components

The HASI model demonstrates a fluid and interactive relationship among its layers and their components. The actions undertaken by the user within the *interaction layer* are informed by their personal trait and are further shaped by the surrounding contexts. The system, characterized by its own specific traits, creates a space for interaction while actively monitoring and adapting to the user's behaviors. At the context layer, the internal and external contexts influence both the user and the system. For instance, regarding the

external context, the physical environment in which the interaction happens can directly affect the user and the system. In a loud environment with high background noise, both the user's cognitive ability as well as the system's utility might be impacted. The internal context also influences both the user and the system. For instance, the interaction history can shape the user's perception and expectations of the system, ultimately impacting the interactions with the system. Likewise, the system can use this historical data to refine future interactions, tailoring responses to align with the user's established preferences. The traits layer encapsulates the unique characteristics of both the user and the system. On one side, users carry individual traits that the system can monitor and learn from to provide more personalized and responsive services. This adaptation can enhance user experience as the system aligns its operations with the user's specific needs and preferences. On the other hand, the system itself is defined by its own traits, including its functional capabilities and representation. As users become familiar with the system's traits, they can adjust their expectations and interactions to better utilize its strengths and understand its constraints.

The primary purpose of the HASI model is to facilitate a shared understanding among researchers, designers, and developers involved in the design and implementation of speech systems. By outlining the key components and dynamics of human-agent speech interactions, considering users' individuality, systems' capabilities, and multifaceted contextual factors, this model could serve as a foundational tool for initiating discussions and sharing perspectives within the research community.

It is important to acknowledge that as our understanding of human-agent interactions evolves and technology advances, the HASI model may require updates and refinements. While the model could provide valuable insights into the interaction process, it is not a fully complete or universally applicable model. Instead, it can be a starting point for further exploration and refinement, accommodating technological changes and user preferences over time.

This dissertation delves deeper into this model, focusing on its individual elements, with a significant emphasis on the system traits. The analysis concentrates on dissecting and examining three key factors within the system side: utility, representation, and interaction style. The approach adopted in this thesis systematically investigates how utility, representation, and interaction style influence the interaction dynamics between users and speech-based agents. By examining these elements, this dissertation seeks to understand their interconnections and collective influence on the broader interaction model.

4

System Utility

The functionality and efficacy of speech systems are essential in shaping a satisfying interaction. Efficient speech systems streamline workflows and enable users to accomplish tasks quickly and precisely. Researchers attribute the disapproval or non-adoption of voice systems primarily to performance challenges, specifically in speech recognition and constrained functionality (Jentsch et al., 2019).

Over the past years, significant advancements have been made in enhancing the functionality of speech systems, particularly with regard to ASR, which can now achieve accuracy levels surpassing 90% (Radzikowski et al., 2019). However, despite these notable improvements, challenges in performance persist for speech systems as they continue to be vulnerable to recognition inaccuracies. When the system fails to interpret users' speech input accurately, it results in user dissatisfaction (Purington et al., 2017), often impeding progress or hindering the completion of tasks (Mavrina et al., 2022).

Recognition failures in human-agent speech interaction can be attributed to three primary sources (Li et al., 2018). First, the system may fail to understand the user's command. Second, the command itself could be misunderstood. Third, the provided command might be out of context, falling beyond the system's vocabulary. Generally speaking, speech recognition errors can be categorized into misrecognitions and non-recognitions (Bohus and Rudnicky, 2008). Misrecognitions refer to cases where the system misinterprets the user's input, whereas, in non-recognitions, the system fails to obtain any interpretation. Various factors contribute to these issues, including users providing complex or unclear input, background noise, limited vocabulary in the system, or faulty hardware (Anusuya and Katti, 2010). In response to recognition challenges, users tend to adjust their communication strategies when interacting with a VUI. Common strategies include hyperarticulation (Stent et al., 2008), which involves speaking more clearly and precisely by exaggerating articulatory movements. Individuals may adjust their speaking pace, reformulate commands, or increase their volume to enhance communication clarity. Another tactic involves physically relocating either themselves or the system to optimize the interaction environment (Jentsch et al., 2019). Language barriers also exacerbate issues related to inaccurate speech recognition. Previous research

highlights that VUIs are more user-friendly and easier for native English speakers to interact with compared to non-native speakers (Pyae and Scifleet, 2018).

On the software front, researchers and developers have explored various strategies to enhance ASR. One conventional approach involves training the system with extensive voice samples to improve recognition accuracy (Li et al., 2018). Additionally, research has shown that augmenting training data with synthesized material can effectively enhance speech recognition (Rosenberg et al., 2019). Machine learning techniques, particularly deep learning (Nassif et al., 2019), have also been proposed to discern underlying patterns in speech data, enhancing recognition accuracy (Haeb-Umbach et al., 2019). Other researchers recommended multimodal approaches where the system combines input from multiple sources to improve recognition accuracy (Mustaquim, 2013; Suhm et al., 2001). For instance, in cases of uncertain recognition results, the system can leverage keyboard or gesture input to confirm or correct the recognition outcome. These highlight some of the ongoing efforts to optimize ASR systems through diverse methods to design more efficient speech systems.

Despite the great progress in speech technology, technical limitations and recognition issues persist as primary factors contributing to user frustration and skepticism when utilizing VUIs (Wei et al., 2022; Ma et al., 2023). This chapter of the dissertation aims to address these efficacy concerns to enhance the quality of human-agent speech interaction. The objective is to identify avenues for improvement and enhancement and contribute nuanced insights and empirical findings supporting advancements in HASI.

This chapter investigates thesis research question 2 (**TRQ2**):

- How can we enhance the *efficacy* of speech systems for domestic activities?

4.1 | Speech Recognition

This section is based on **Publication 1**:

Nima Zargham, Mohamed Lamine Fetni, Laura Spillner, Thomas Muender, and Rainer Malaka. **"I Know What You Mean": Context-Aware Recognition to Enhance Speech-Based Games**. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24), Association for Computing Machinery.

My contribution to this work: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, part of software, supervision, validation, visualization, and contribution to all parts of the manuscript.

One of the most fundamental aspects of HASI is speech recognition, reflecting the system's ability to transcribe spoken words into recognizable commands accurately. The efficacy of this process directly influences the overall user experience, as prompt and accurate recognition is crucial. Recognition issues are a key source of user frustration and dissatisfaction when engaging with speech systems (Jentsch et al., 2019). Achieving a seamless user experience remains a daunting task. The reliance on ASR systems in VUIs contributes to persistent usability issues, including bypassing non-speech conversational cues (Murad et al., 2019). Additionally, factors such as background noises, hardware constraints, and language barriers add to the complexity of this process (Springer and Cramer, 2018).

In an attempt to enhance the recognition accuracy of speech systems, we developed context-aware speech recognition, which will be presented in detail in the following section.

Context-Aware Speech Recognition

In everyday human interactions, imagine someone entering a room and asking, "Where's the charger?" This question becomes clearer when we consider the context. Observing the room, noting the usual places for electronic devices, and recalling recent actions, such as someone using a laptop, helps us understand the request. This comprehensive understanding allows the person responding to accurately guide the inquirer to the specific location of the charger. Similarly, in human-agent interaction, incorporating contextual details like the environment's layout, object locations, user gaze direction, and a history of the user's actions could enhance the virtual agent's ability to understand commands within a specific context. Just as humans use contextual cues to navigate and understand their surroundings, integrating contextual information in human-agent interactions aligns with natural communication patterns, potentially optimizing the agent's responsiveness and overall user experience.

Inspired by this, we implemented a novel approach called context-aware speech recognition, incorporating information about the user's actions and environment to improve prediction accuracy. For many speech systems, during the development process, a speech recognizer's vocabulary is created where a specific set of commands is defined. When processing speech commands, a common approach involves comparing the recognized output text with all available commands in the vocabulary to enhance the prediction of the intended command and improve recognition. This comparison is based on their Levenshtein distance, representing the minimum number of single-character edits required to transform word A into word B (Levenshtein et al., 1966; Ziółko et al., 2010). The system then executes the command with the lowest distance. However, recognition accuracy is often decreased due to acoustic similarities between different commands (Zgank and Kacic, 2012). Our method aimed to aid this process by using additional contextual data for better prediction.

We designed a speech-based video game that uses context-aware speech recognition, which handles players' speech commands, considering the game environment and actions. In our method, when a command exactly matched one of the commands in the vocabulary, the system would execute that command. However, if a command did not precisely match any available commands, the recognition system would calculate a confidence score for each possible command. In the first step, the set of possible commands was limited to those commands in the vocabulary that were similar to what was recognized. We set a maximum Levenshtein distance threshold of 20 - all commands with a distance > 20 were not considered possible commands. This number was chosen empirically after our initial testing sessions of the game, as it showed to be an appropriate number to effectively detect phrases that are too long, too short, or too different from the list of accepted commands (vocabulary). A fallback interaction was triggered if there was no possible command with a Levenshtein distance below this threshold. In such cases, the game's main character would respond with a message indicating that they did not understand the player's instruction. Otherwise, the confidence score was calculated for all possible commands with a distance of ≤ 20, and the command with the highest total score was executed.

We designed two versions of the game. In the control group, when a command did not precisely match any available commands at that level, the recognition system would calculate similarity scores using the Levenshtein distance between the recognized input and the available commands and execute the one with the lowest distance. If the Levenshtein distance were higher than a set minimum, the system would consider that command unrelated and trigger the fallback interaction. We refer to this as the *scope filter* in this work. In the intervention group, we additionally implemented an *environment* filter and an *actions* filter. The *environment filter* takes into account the environment of the player inside the game at the given moment, while the *actions filter* is based on context information about the possible commands at a certain point in the gameplay. The final confidence score was calculated as a weighted sum of the three scores based on the *scope filter*, *environment filter*, and *actions filter* (see Figure 4.2). In a between-subjects user study with 40 participants, we compared these two conditions.

For this study, we aimed to answer the following research questions:

RQ1: Can data derived from the game environment and actions aid command prediction?
RQ2: Does using context-aware speech recognition based on game environment and actions enhance usability and player experience?
RQ3: Does using context-aware speech recognition enhance players' performance in a speech-based game?

Figure 4.1: When pointing at interactable objects, a list of actions appears in the top-right corner of the screen. The already performed actions are crossed out.

Game Design

To evaluate our method, we designed "Escape the Echo," a speech-based escape room game where players have to communicate with the main character "Sophie" using speech commands. The player's objective is to help Sophie escape various rooms by guiding her to inspect specific objects and use them to exit the room. The game unfolds across three levels – a jail cell, a bathroom, and a classroom. In each level, players can instruct Sophie through a series of actions to progress and successfully escape.

Players have to instruct Sophie to perform specific tasks using speech commands linked to in-game objects like mirrors, desks, or doors. When players target an interactable object, its name appears in the center of the screen, along with a list of available actions in the top-right corner (e.g., inspect, move, or break), as shown in Figure 4.1.

Players determine instructions based on these actions, such as interpreting the hint "Break" and targeting the game object "mirror" to command, "Break the mirror." The game has a total of 86 unique actions and 36 unique interactable objects distributed between the three levels. The speech system was programmed to handle various phrases for each action. If the command has been executed already, Sophie would reply, "I have already done that." If the command could not be performed on that game state, she would reply, "I cannot do that." Player control is limited to mouse movement for exploring the room through a handheld camera controlled by Sophie. Character movement is constrained to actions the player instructs, aligning with the established story and player identity in the game.

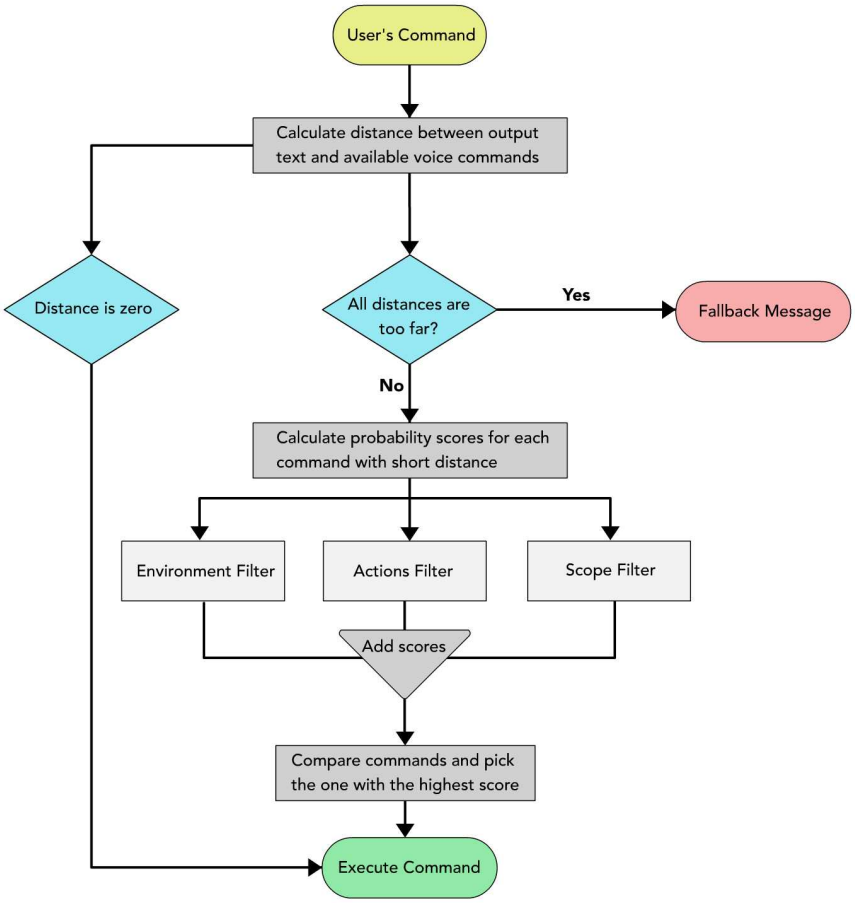


Figure 4.2: The general process of the command prediction in the intervention group using all three filters.

Filters

Scope Filter The *scope filter* assigns a score to each command based on its similarity to the recognized output text. The system compares the recognized output text with available commands in the level, executing the one with the lowest Levenshtein distance. For instance, suppose the player's recognized intent is "Bake the mirror," and the expected command in the system's vocabulary is "Break the mirror." In this case, the Levenshtein distance would be 3 (add 'r' after 'B,' add 'e' after 'r,' and remove 'e' after 'k'). If there were no other commands with a lower distance, then the "Break the mirror" action would be triggered. Due to the distance threshold, the Levenshtein distance of a possible command would be between 1 and 20. Let **N** be the Levenshtein distance of a given command to the output text, and **M** be the maximum possible distance based on the threshold. Then, the scope filter score was calculated as follows: $(M - N + 1) * 4$.

Environment Filter The *environment filter* assigns scores to each available voice command in the current level based on the game environment at the time the command was given.

The maximum possible score for this filter is 30. The environment filter score for a given command is calculated based on the interactable objects visible in the frame (camera view) and which, if any, interactable object the player is currently targeting (meaning that it is at the center of the field of view). If the object mentioned in a command is the one that is currently being targeted, this adds 15 points to the environment filter score of this command. Additionally, all those commands corresponding to objects visible in the frame but not necessarily being targeted also get a number of points calculated as $15/N$, with N being the number of objects in view. This is because players might not necessarily aim at an object as long as it is visible. If three objects are in the frame and one is being targeted, then commands referring to the target object will receive an environment score of $15 + 5 = 20$. The other two will receive an environment filter score of 5 each. This approach allows the system to prioritize commands related to more prominent objects in the frame.

Actions Filter The *actions filter* assigns scores to possible commands based on the actions that should be performed, that is, how a player wants to interact with an object instead of which object it is. Similar to the *environment filter*, the *actions filter* also has a maximum score of 30 and a minimum of zero. This filter takes into account four facts about the current context: whether or not the action has just been revealed as an option to the player (after the player inspected the same item in the previous step), whether or not the action is known to the player in general, whether or not it is possible in the current game state, and whether or not the action has already been tried in this state. If an action has just been revealed, it must be possible and is now known to the player. However, it can happen that the action has already been tried even before it was revealed. Thus, for a given possible command, one of the following will apply:

- All four facts are true, as the action has just been revealed and has not been tried before. In this case, the command receives 30 points (maximum).
- The action has just been revealed. However, the player has already tried it before in a previous step: 15 points.
- The action has not just been revealed, but it is possible, known to the player, and not tried yet: 15 points.
- The action has not just been revealed, and it is also not possible. However, it is otherwise known to the player and yet to be tried: 1 point.
- In all other situations, this command receives 0 points.

Measures

We employed standardized questionnaires to evaluate both player experience and the perceived usability of the speech system. These included the System Usability Scale (SUS) (Brooke et al., 1996) and the Player Experience of Need Satisfaction (PENS) (Ryan et al.,

2006). In addition, we asked a set of customized questions to gather insights into players' game experience. Lastly, each participant underwent a brief semi-structured interview to delve into qualitative aspects of their player experience, usability, and individual preferences (Wilson, 2013). Furthermore, after each gameplay session, a log file captured details, including the total number of commands, directly recognized commands (without using filters), predicted commands (using filters), playtime, average prediction scores from *scope*, *environment*, and *actions* filters, the overall confidence score for commands, and the number of predicted commands that would yield the same outcome using only the *scope* score. While the control group's gameplay was influenced solely by the *scope filter* during the experiment, data from the *environment* and *actions* filters were logged for analysis.

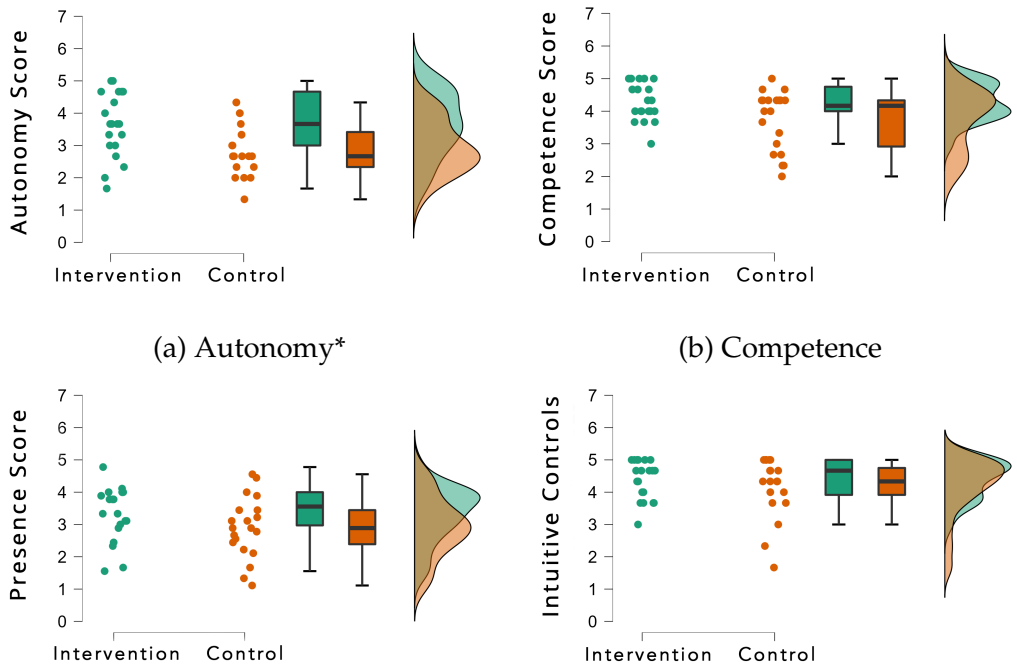
Findings

Players in both groups expressed overall enjoyment and provided positive feedback on the game. They found the experience of controlling the game with their voice to be exciting and novel. Post-experiment, participants inquired about the possibility of new playable rooms, with several expressing a desire to replay the game to uncover additional content. Interviews showed that players felt immersed when conversing with the main character, sensing a connection to the game's world. This aligns with existing literature, indicating that in-game voice commands contribute to a sense of embodying a character within the game's world (Allison et al., 2019) and is in line with previous research on voice-controlled games, highlighting the potential for heightened immersion through voice interaction (Zhao et al., 2018; Lee et al., 2006; Osking and Doucette, 2019).

Results obtained from the game logs indicate that our proposed method had a noticeable impact, influencing 37% of the predicted commands. Despite the *scope filter* carrying a higher weight than the supplemental *environment* and *actions* filters, they collectively affected around one-third of all given commands. These findings robustly support the substantial influence of our proposed method on command prediction. However, *RQ1* could not be fully answered as we lack ground truth and insights into players' intended actions for each command.

A significant difference was observed in favor of the intervention group regarding *Autonomy* (see Figure 4.3). This might be attributed to the greater flexibility in command formulation, which players also raised throughout the interviews. Players in the intervention group felt less restricted by command variability, suggesting that context-aware speech recognition can enhance perceived freedom of control and flexibility in the game. Additionally, players in the intervention group reported higher enjoyment and overall experience ratings.

Moreover, our results revealed significantly higher usability scores for the intervention group. This is further supported by our customized questionnaire, where players in the intervention group perceived significantly fewer errors despite the game logs showing a



(a) Autonomy*

(b) Competence

(c) Presence

(d) Intuitive Controls

Figure 4.3: The distribution of variables and the mean and confidence intervals of the PENS results between the control and intervention groups.

significantly higher total number of correctly recognized intents (no filters applied) in the control group. The high usability ratings could imply that the game was more convenient to play as the system could accurately interpret players' intended commands, minimizing the need for repeated instructions. Thus, we affirm that context-aware speech recognition can enhance usability and player experience in speech-based games (*RQ2*).

We found no significant differences in playtime, the number of voice commands invoked, or the number of times the filters were used. Additionally, there were no distinctions in prediction scores and filter scores. This indicates that both groups encountered similar playing conditions and faced comparable recognition errors and interactions with the environment and game state. The consistent number of commands and playtime implies that players performed at the same level regardless of the recognition method. Players also observed this as they rated their performance similarly in both conditions in the customized questions. Therefore, we conclude that the context-aware recognition method did not necessarily enhance players' performance (*RQ3*).

Overall, the findings of this study suggest that data from the game environment and actions can be leveraged in video games or virtual environments to improve speech recognition accuracy. This could enhance the usability of the speech system and improve the interaction process.

4.2 | Error Handling

This section is based on **Publication 2**:

Nima Zargham, Johannes Pfau, Tobias Schnackenberg, and Rainer Malaka. **"I Didn't Catch That, But I'll Try My Best": Anticipatory Error Handling in a Voice Controlled Game**. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22), Association for Computing Machinery.

My contribution to this work: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, part of software, supervision, validation, visualization, and contribution to all parts of the manuscript.

Speech systems are typically trained with extensive voice data and connected with ontologies and knowledge graphs to identify and understand user commands and reply with a satisfying answer (Li et al., 2018). Nevertheless, user commands can be fuzzy, personal, and complex, leading to the occurrence of recognition errors. When such errors occur, appropriate error handling is crucial. Effective error handling will not derail users. It can keep them on track and lead to successful task completion (Pearl, 2016). However, if error handling is done poorly, it may cause users to fail tasks and potentially refuse to use the system again. This makes error handling a critical part of designing VUIs (Li et al., 2018).

Several guidelines for designing fallback strategies have been proposed, which include asking the user to repeat the command, redirecting the user to the tasks that the system can support, or presenting user options to correct their commands (Pappu and Rudnicky, 2014; Bohus and Rudnicky, 2005; Li et al., 2018). Occasionally, speech systems fall back on humor in response to complicated commands that the system cannot appropriately handle otherwise, which users might see as sarcastic or entertaining (Porcheron et al., 2017). A study by Bohus et al. suggests advancing the conversation by ignoring the non-recognition and trying an alternative dialog plan (Bohus and Rudnicky, 2008).

To contribute to the existing literature on error handling methods in speech systems, we have developed a novel approach called anticipatory error handling. This will be presented in detail in the following section.

Anticipatory Error Handling

Previous literature suggests that after a recognition error, the likelihood of errors in subsequent intents increases (Rotaru et al., 2005; Swerts et al., 2000; Bohus and Rudnicky, 2008). One contributing factor to this could be that, as errors accumulate, the user's patience diminishes, giving rise to heightened frustration. This, in turn, can result in

acoustic and language mismatches (Bohus and Rudnicky, 2008). Human operators often bypass signaling non-understandings and try to advance the task by asking different questions, generally leading to a quicker recovery (Skantze, 2003). In line with this, for speech-based systems, researchers recommend using alternative dialog plans to progress the task when non-understandings occur rather than focusing on repairing the current problem (Bohus and Rudnicky, 2008).

Expanding on the research about the error handling of speech systems, we developed anticipatory error handling, an approach to bypass unrecognized commands and avoid the need for command repetition for correction. In this method, when a command was not recognized, the system would continue by executing a locally optimized action, focusing on goal completion and obstacle avoidance, without alerting the user about the recognition failure.

To assess our approach, we developed a speech-based video game called "Listen, Sparky!" to investigate the user experience. In our game, players use speech commands to control the protagonist. We conducted a between-subjects user study with 34 participants, comparing traditional repetition-based error handling with the anticipatory error handling approach implemented in the game. In the control group, the game notified the player of recognition failure, prompting command repetition (see Figure 4.4). Conversely, with anticipatory error handling, if a command went unrecognized, the game would execute a locally optimized action considering goal completion and obstacle avoidance without notifying the player about the recognition failure.

Figure 4.4: In the control group, when a command is not recognized, the game displays question marks over Sparky's head.

In this work, we looked into the following research questions:

RQ1 Does performing a locally optimized game action in times of misrecognition lead to a measurably improved usability in a speech-based video game?
RQ2 What are the effects on player experience regarding competency, autonomy, presence, and intuitive control if error handling mechanisms decide for unintended actions?

Game Design

We designed and implemented "Listen, Sparky!", a speech-controlled arcade game. In this game, players control the sheepdog "Sparky," who has to guide a sheep through challenging courses while avoiding hazardous encounters. Players assume the role of a shepherd, issuing speech-controlled commands to direct their sheepdog. With the progression of the levels, the challenge of the game would similarly increase. After completing the initial two levels, a hostile wolf character was introduced, posing a threat to the survival of the escorted sheep. If the sheep approached the wolf too closely, the level would fail and have to be restarted.

Figure 4.5: Voice commands making up the core game controls. This menu was accessible anytime during gameplay.

Players could choose from five possible actions to command Sparky at each game state, as shown in Figure 4.5. The system accommodated multiple phrases per action. For example, players could instruct Sparky to 'flank right' using phrases such as 'go right,' 'right side,' or 'move right.' If the voice recognition system recognized a command, Sparky would execute the corresponding action. If no matching command was found, the system

treated it as a failed attempt, triggering the error handling system based on the respective experimental group.

We recorded error rates for each game session, representing the number of commands that went unrecognized by the system throughout the session. To evaluate the error handling methods, it was essential to have noticeable instances of recognition failure. To ensure this, both game versions were programmed to provide a minimum overall error occurrence of 15% after the initial ten commands. This meant that if a player achieved an error rate below the target, the system would intentionally misrecognize the next request.

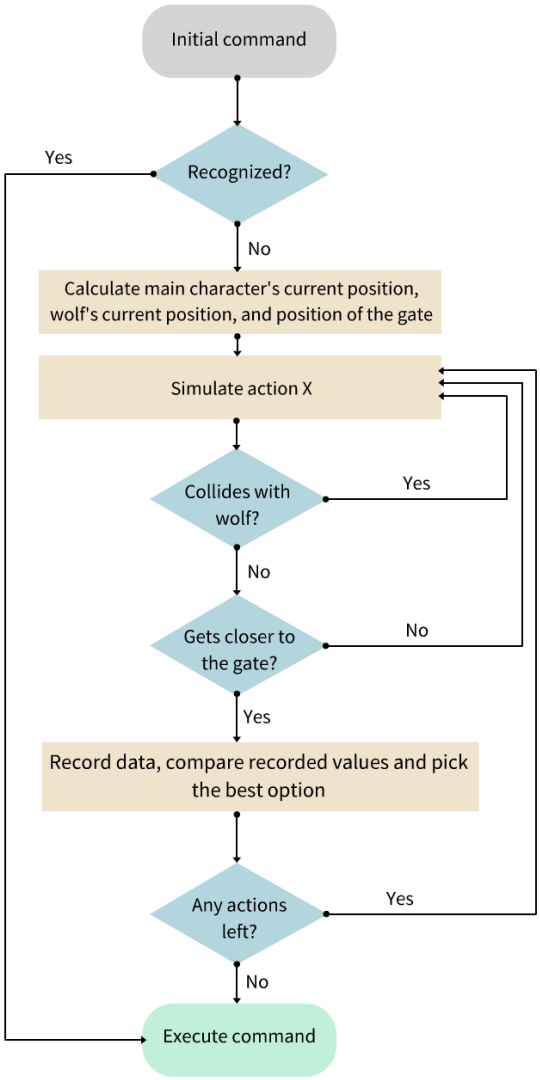


Figure 4.6: General process of the anticipatory error handling.

In the control group, in the case of non-recognition, the character would not react but only indicate that the command was not recognized by displaying some question marks above its head (see Figure 4.4). In the intervention group, players played a version that implemented anticipatory error handling based on the underlying game state. If a

command went unrecognized, the game would execute a locally optimized action focusing on obstacle avoidance and goal completion without notifying the player of the recognition failure. The primary priority was to prevent the sheep from being caught by the wolf (obstacle), followed by considering actions that would position the sheep closest to the gate (see Figure 4.6).

As an example, in the depicted game situation illustrated in Figure 4.7, if the player commands Sparky to "bark" but the intent is not recognized, the game would consult the error handling system. The error handling system would then determine the most optimal action at that moment, aiming to ensure the sheep avoids the wolf and/or moves closer to the gate. In this scenario, the system selects "flank left" as the anticipated solution, as it offers the best possible outcome by keeping the sheep away from the wolf while moving it closer to the gate.

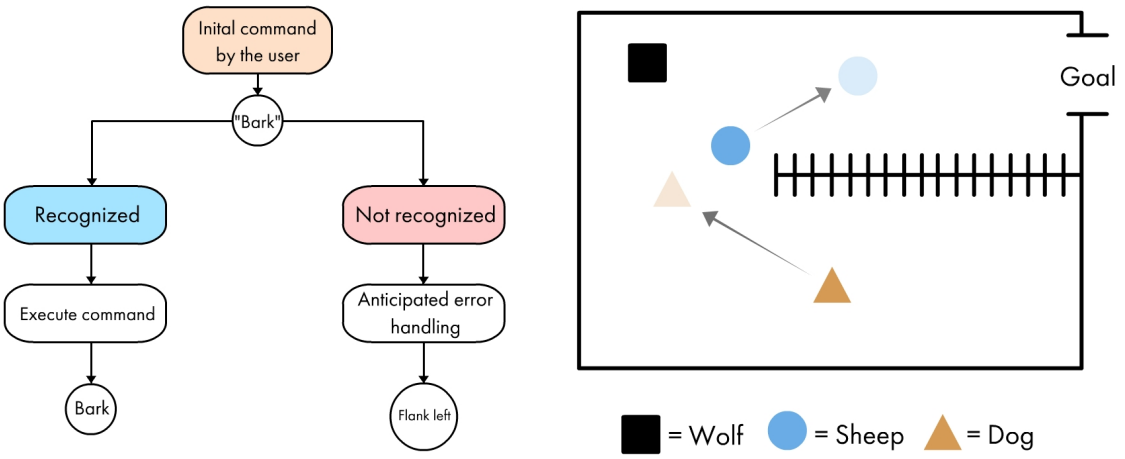


Figure 4.7: Displaying a specific game situation where the recognition fails, and the system chooses to flank left as it would have the best possible outcome (right). The flowchart shows the process of anticipatory error handling in the intervention group (left).

Measures

We used standardized questionnaires to evaluate the player experience and the perceived usability of the system. Our post-exposure questionnaires included the SUS (Brooke et al., 1996) and the PENS (Ryan et al., 2006). A series of customized questions were also recorded on 5-point Likert scales, focusing on the extent to which Sparky behaved as the participant expected him to do so, Sparky's perceived intelligence, and overall game experience. Furthermore, participants were prompted to estimate the number of unrecognized commands and describe Sparky's response when commands were not recognized. We concluded each session with a brief, semi-structured interview to gather additional insights, allowing for a qualitative exploration of player experience, usability, and individual preferences in both conditions (Wilson, 2013).

Findings

In general, users provided positive and supportive feedback about "Listen, Sparky!". Players from both conditions expressed enjoyment in playing our speech-based game. Throughout the experiment, participants willingly repeated levels even after successfully completing them, and many expressed a desire to continue playing beyond the experimental session. Participants highlighted that controlling Sparky with their voice enhanced their sense of immersion, making them feel "like an actual shepherd."

Several players struggled with command recognition, particularly at the beginning of the game. However, over time, participants improved their understanding of the recognition system, refining their command formulation and enunciation for better recognition. As they progressed through various levels, players developed proficiency in utilizing game mechanics. Furthermore, we noted that when players felt time pressure, it contributed to more complications in command recognition. This was primarily attributed to variations in talking pace and quick decision-making, resulting in unclear and incorrect commands. Additionally, non-native speakers exhibited a higher error rate, increasing frustration for these players during the game.

The study revealed significantly enhanced usability (*RQ1*) for the version incorporating anticipatory error handling. However, our qualitative findings indicate that the usability increase is primarily attributed to cases where error handling aligned with the user's intention, which was not always the case despite selecting the technically optimized solution. Mismatches between participants' intended commands and the system-selected command were perceived as a distinct error type, negatively impacting user learning curves. We also did not observe any differences in terms of player experience other than the subscale of intuitive control. Thus, we argue that anticipatory error handling may not inherently enhance the user experience of speech-based games (*RQ2*). The key focus of this handling technique should be on tailoring predictions to individual users rather than approximating technically optimized decisions. Furthermore, quantitative analysis of recorded errors showed that, despite the intervention group participants committing more errors on average, they reported a significantly lower perceived error count than the control group. This was partly due to unrecognized commands in the intervention group being the optimized action, eliminating the perception of recognition failures. However, users were still less likely to notice this intervention, even when misrecognition was addressed by an optimized action diverging from the intended command.

Upon revealing both conditions to participants, diverse opinions emerged about the appropriate error handling method. Some preferred anticipatory error handling for its contribution to maintaining the game's flow, while others disliked it, perceiving it as masking the problem rather than solving it. A participant even suggested introducing a random action for added challenge and surprise. Given these varied perspectives, the optimal solution may vary among players. Therefore, we recommend developers to consider incorporating multiple error handling methods as optional features, allowing

players to choose based on their preferences.

Overall, we observed that players did not necessarily prefer anticipatory error handling if the executed action did not align with their initial intent. This suggests that maintaining a sense of full control and agency, even with suboptimal actions for level completion, might be preferable. However, the frustration of repeating actions when they are not recognized was even more pronounced. Hence, emphasis should be placed on understanding the user's initial actions rather than prioritizing optimal actions. Otherwise, incorrect handling can negatively impact the experience, hinder learning progress, and raise doubts about error handling in general. Nonetheless, this work introduced an initial approach to anticipatory error handling, showcasing "optimal decisions" based on heuristics. For a more comprehensive understanding, further exploration is needed to consider factors such as player types, mood, and game genres.

4.3 | Conclusion

This chapter discussed two distinct research works dedicated to improving the functionality of speech systems with regard to the processing of intents. The presented works employ software-based solutions to tackle efficacy concerns related to speech interaction and aim to enhance the ability of speech systems to resolve uncertainties. The primary focus of both studies is on developing innovative methods to refine the processing of speech inputs and effectively manage ambiguous or unclear commands. The first work introduces an approach leveraging data from user actions and environmental cues to enhance the prediction of the users' intended speech commands. This method explores the integration of contextual information to provide more accurate and context-aware speech recognition. The second work investigates a novel error handling approach, where the system determines and selects the optimal course of action for the user in instances of recognition failures. These research works target different stages in the processing of speech intents. Collectively, they contribute to the holistic improvement of users' speech interaction experiences.

For each study discussed in this chapter, a fully functional game featuring a voice user interface was developed. These works specifically focused on speech-based video games to explore their proposed methods. This choice is justified by the prevalence of playing video games as a common everyday activity and the rising popularity of speech-based video games, an area with limited research. Additionally, video games provide controlled environments, facilitating high experimental control. Nevertheless, the methodologies employed in these studies hold relevance beyond their specific gaming contexts. In the case of context-aware recognition, while tailoring the *environment* and *action filters* to each game is essential, the fundamental principles introduced serve as a foundation for designing similar systems across diverse gaming contexts. Moreover, these principles extend to other realms of HCI, particularly in virtual environments like virtual reality. In

non-virtual settings, integrating multimodal systems and combining data from various sources like gesture or gaze can enhance command prediction accuracy. Similarly, the findings related to anticipatory error handling can be applied to a broader spectrum of technologies and HCI applications. In goal-oriented activities involving collaboration between users and systems, anticipatory error handling could prove beneficial for successful interaction. However, it might not be the optimal approach for tasks that involve creativity or exploration, where greater emphasis on identifying the user's initial intention is crucial for an enhanced user experience. Nevertheless, even though the broader insights of these works can apply to the use of VUIs in general, in future work, these methods could be transferred and evaluated in other domains, such as navigation, medicine, education, and smart homes, to explore different settings.

In our investigations, we sought to address both non-recognitions and misrecognitions to resolve uncertainties. Anticipatory error handling emerges as a potential fallback solution for both non-recognitions and misrecognitions. On the other hand, context-aware recognition holds promise in mitigating misrecognitions by incorporating additional environment and action filters. Evaluating misrecognized intents posed a significant challenge, as identifying the player's intended action was not always feasible. Therefore, this method may not impact non-recognized intents. Nevertheless, these approaches, especially context-aware recognition, can significantly benefit non-native English speakers or those with distinct accents and dialects. The additional information from these filters can enhance the system's ability to predict users' intended actions, reducing the likelihood of misrecognition.

A logical next step in refining the methods introduced in this chapter would be to incorporate deep learning techniques and user models to enhance the prediction of intended commands. This could lead to more sophisticated and personalized systems that adapt to individual users' speech patterns and preferences, thereby improving overall accuracy and user satisfaction. Additionally, this also supports user agency, providing individuals with greater empowerment by aligning system outputs more accurately with their intentions and promoting an adaptable and user-centric approach.

The positive feedback and enthusiasm for both games developed for evaluation can be partly attributed to the unconventional nature of speech-based video games. The demographic data and participants' perception of the games' novelty indicate that voice-controlled games remain an unfamiliar category. Our research emphasizes that integrating speech-based interaction in games enhances inclusion and immersion by actively engaging with in-game characters. We encourage further exploration in this field. The studies discussed in this chapter offer valuable insights for researchers and developers on addressing and managing speech recognition in video games and broader applications of voice user interfaces.

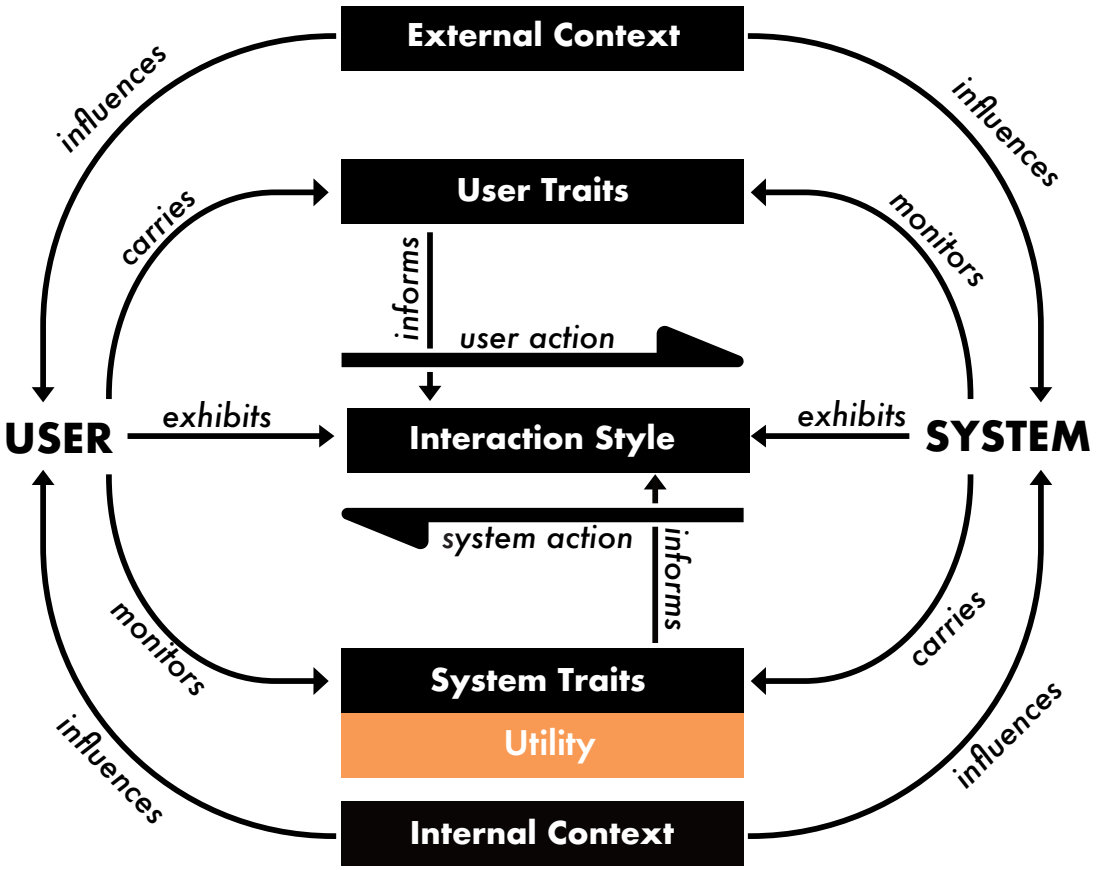


Figure 4.8: The role of system utility in the HASI model.

Reflection on the HASI Model

Based on the proposed HASI model, the utility dimension investigated in this chapter is part of the system traits within the traits layer (see Figure 4.8). Our research demonstrated the significance of external context, particularly the environment, in the speech interaction process. We discussed that certain environmental factors, such as distance or background noise, can influence the efficacy of speech systems. Understanding and mitigating these environmental challenges is crucial for ensuring robust speech interaction systems that can perform reliably across various real-world settings. On the other hand, our research showcased that leveraging contextual information could also enhance speech interaction accuracy. Using contextual data can support the system to better predict users' intended commands, offering a pathway to designing more precise speech systems. Furthermore, user traits, such as their familiarity with the system, could enhance the effectiveness of user interaction with VUIs. Users tend to adapt their interaction patterns based on past experiences, influencing the system's performance. Additionally, we observed that deviations from user preferences or expectations can lead to user dissatisfaction and

raise concerns regarding user agency. Our study highlighted that understanding users' intentions is more important than selecting an optimal option for them. Prioritizing user preferences can enhance users' sense of agency when engaging with VUIs. Therefore, systems need to actively monitor and understand user traits, particularly their preferences, and adjust interactions accordingly. Incorporating these insights can help optimize speech recognition technologies, resulting in improved user experiences, streamlined interactions, and increased overall effectiveness.

In response to the dissertation's **TRQ2**, we posit that utilizing contextual information to enhance the prediction of users' intended commands has the potential to strengthen the efficacy of speech recognition. Moreover, prioritizing the prediction of user-intended actions over technically optimized actions may improve overall experience and heighten user satisfaction. We witnessed that in both studies, the usability of the speech system for the proposed methods was rated significantly higher compared to the conventional approaches, highlighting the potential of using such methods to enhance HASI.

5

Agent Representation

Much of the research on speech interaction focuses on enhancing the efficiency and accuracy of these systems by investigating methods to improve recognition and its accuracy rates (Allison et al., 2018). However, it is essential to recognize that beyond the functionality of speech systems, their appearance and representation are of significant importance. Yuksel et al. argue that an agent's appearance and attractiveness might be even more important than its reliability (Yuksel et al., 2017). Similarly, Lopatovska et al. (Lopatovska et al., 2019) posit that, depending on the task the user would like to accomplish with the speech system, prioritizing a pleasant UX may supersede the quality of the outcome. This perspective underscores the idea that the end result does not solely determine users' engagement and satisfaction. Other elements in the interaction process profoundly influence the users' experience with such systems.

An important dimension within the speech interaction process involves the representation of speech systems, where the visual and auditory elements play an essential role in shaping user perceptions and, consequently, influencing the overall user experience. Often, a prevalent misconception persists that views the representation of an AI agent as a superficial layer, seemingly detached from being an integral component that can significantly contribute to the system's effectiveness and user acceptance (Khan and De Angeli, 2009; Yuksel et al., 2017). However, this mindset has been changing in recent years. With the proliferation of speech systems, specifically in homes and smartphones, research on the representation dimension of speech interaction has been expanding.

As highlighted in previous chapters, existing literature indicates that conversational agents often fall short of meeting user expectations as effective interlocutors (Jentsch et al., 2019; Luger and Sellen, 2016; Murad and Munteanu, 2019; Doyle et al., 2019). A potential contributing factor to this shortfall is the insufficient emphasis placed on the representation of these entities as engaging conversation partners. Steering the complex design process of a speech agent entails a critical examination of its representation. Designing the qualities and characteristics the agent should embody is vital to elevate user experience and satisfaction (Doyle et al., 2019). One of the most common design strategies for conversational agents is anthropomorphism. Incorporating

anthropomorphism in product design has demonstrated notable benefits (Hart et al., 2013). Currently, due to commercial speech systems being mainly voice-only systems, the representation of speech agents is predominantly conveyed through their voice, linguistic characteristics in responses, designated personifications (e.g., Alexa instead of the product name Amazon Echo), and the physical design of the device (Reeves and Nass, 1996; Bickmore and Picard, 2005; Beneteau et al., 2019). Research has consistently emphasized the profound influence of such design decisions on user experience, spanning considerations such as the agent's gender (Brahnam and De Angeli, 2012; Hwang et al., 2019), voice characteristics (Chidambaram et al., 2012), and visual attributes (Wang et al., 2019; Andrist et al., 2017). However, a prevalent issue arises from the standard implementation of one-size-fits-all systems, which may not align with the diverse needs and preferences of the user population. This lack of adaptability in design decisions can lead to inherent problems, perpetuating societal stereotypes (Hwang et al., 2019), lacking inclusivity, overlooking diverse populations, and neglecting accessibility considerations (Abdolrahmani et al., 2020). Considering these dimensions is imperative for the development of more user-centric and universally accessible conversational agents.

This dissertation chapter aims to delve into the representation aspects of speech agents to examine how various attributes and characteristics of these agents influence user interaction and experience. Additionally, the chapter seeks to uncover insights that contribute to the effective and user-centric design of speech systems, aligning them more closely with users' preferences and promoting inclusivity, accessibility, and adaptability. It is important to note that the studies discussed below concentrate on virtual agents rather than physical manifestations such as robots.

This chapter explores **TRQ3**:

- How does the *representation* of virtual speech agents contribute to a better interaction experience?

5.1 | Number of Agents

With the wide design space of VAs, their development involves intricate and crucial design decisions. One such parameter is the number of interlocutors in a conversation. Typically, most voice assistants employ a single human-like voice to respond to user queries, creating the impression of a single agent assisting with tasks. Notably, in commercial products, the default setting often utilizes a female voice, potentially reinforcing gender stereotypes. Hwang et al. delved into the reflection of gender stereotypes in female-voiced assistants, identifying characteristics like bodily display, subordinate attitude, and sexualization, which could establish a power dynamic between users and female agents (Hwang et al., 2019). While conversation scenarios involving more than two interlocutors have been extensively explored in human-human interactions (Branigan, 2006) and interactions with multiple persons conversing with a single artificial agent (Johansson et al., 2014;

Pappu et al., 2013), limited research has investigated scenarios with multiple agents conversing with one user. A study by Abdolrahmani et al. proposes that offering simultaneous access to multiple VA personas can effectively support blind users in varied contexts (Abdolrahmani et al., 2020). As the appropriateness of output heavily relies on interaction context and content, users might benefit from access to customizable personas tailored for specific tasks, such as cooking or scheduling.

Building upon prior research on incorporating multiple agent personas into a single VA system, we introduce the concept of multi-agent voice assistants. In this framework, we co-embody multiple agents into one system, each specialized in a distinct task domain, aiming to explore its impact on user experience. To assess our concept, we conducted two separate studies—one in a virtual smart home setting (Publication 3) and another in a VR video game (Publication 4). This section will delve into the details of these two studies. We conducted both of these studies in an immersive VR setting. State-of-the-art VR technology offers interactive, high-fidelity simulations, providing experimental control, cost-effectiveness, and replicability (Kinateder et al., 2014). With regards to Publication 3, conducting the user study in VR provided a home-like setting without intruding into participants' actual homes, ensuring a wiretap-free environment. Concerning Publication 4, the VR environment offered a more immersive experience, allowing users to better engage with the game environment while eliminating outside distractions. It also offered technical ease of implementation while maintaining authentic interactions. Prior research has highlighted the effectiveness of field studies simulated in VR as a robust research tool, demonstrating largely similar behavioral patterns between virtual and real settings (Mäkelä et al., 2020; Paneva et al., 2020; Agethen et al., 2018; Moussaïd et al., 2016; Deb et al., 2017). Insights from these studies emphasized the importance of designing scenarios that encourage natural behavior and allow users to explore the technology freely, a principle we applied in our work to yield ecologically valid results.

5.1.1 | Multi-Agent Home Assistants

This section is based on **Publication 3**:

Nima Zargham, Michael Bonfert, Robert Porzel, Tanja Döring, and Rainer Malaka. **Multi-agent Voice Assistants: An Investigation of User Experience**. In Proceedings of the 20th International Conference on Mobile and Ubiquitous Multimedia (MUM 2021), Association for Computing Machinery.

My contribution to this work: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, validation, visualization, and contribution to all parts of the manuscript.

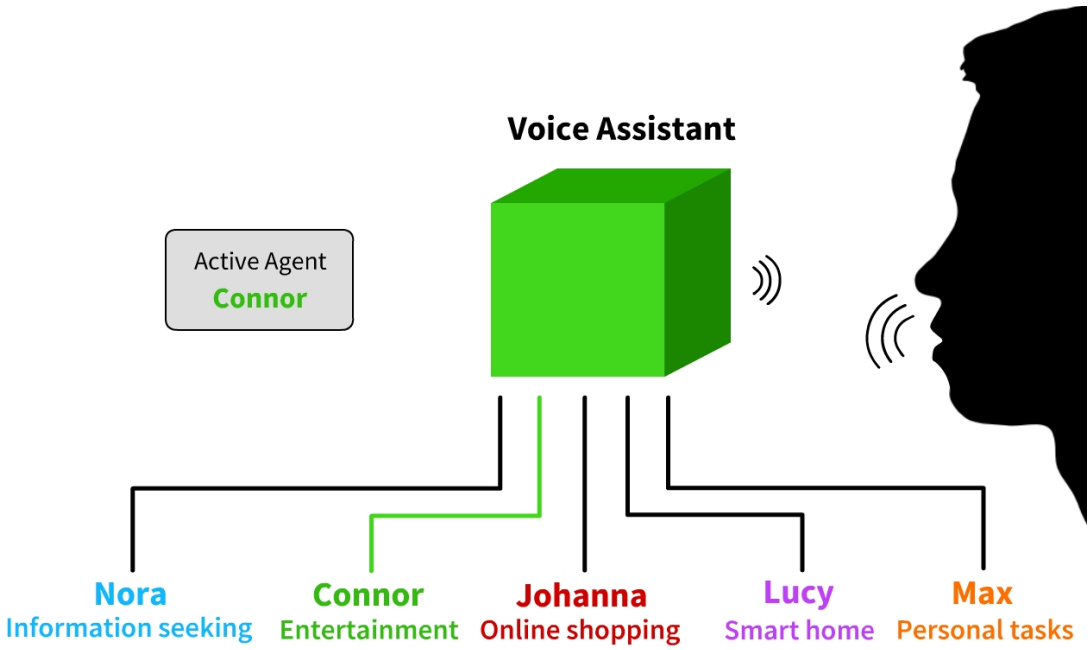


Figure 5.1: Schematic illustration of a multi-agent voice assistant with five available agents, their specialized task domain, and representing color. Here, the user talks to the currently active agent, Conner, as indicated by the cube-shaped device lighting up green.

We sought to conduct a user study to unveil insights into the user experience of a multi-agent system within a single home assistant device, exploring both its potential benefits and challenges. To evaluate the impact of our concept within a home setting, we developed a voice assistant in virtual reality that appears as a unified device housing multiple agents, each equipped with a distinct voice and assigned to a specific task domain (See Figure 5.1). The system automatically selects the most suitable agent for a given task when the user initiates a dialogue. With this, the characters can be perceived as individuals and competent experts in their responsibilities.

Prototype Design

We designed an apartment in VR where users encountered a VA and several smart appliances, emulating a smart home environment (see Figure 5.2). In our prototype, the VA device exhibited a grey hue during idle periods and would change to yellow while actively listening. The device changed colors based on the responding agent, with each agent being represented by a unique color. This visual cue aimed to convey the presence of the respective agent occupying the device temporarily. Similar to conventional home assistants, users initiated commands by uttering a wake word. They could always interrupt it by saying, “Stop.” When activated, the VA’s visual representation turned yellow and began rotating, signaling readiness to receive commands while remaining grey and static otherwise. The system only recognized commands relevant to the experiment, rejecting all other inquiries with a standard response indicating unsupported features.

Figure 5.2: The virtual smart home environment. The voice assistant device on the right is embodied as a hovering cube. Its orange color represents the currently active agent, Max, responsible for "personal tasks".

In the control condition, mirroring industry standards, a single agent with a female voice was implemented. Conversely, in the intervention condition, we introduced the multi-agent VA comprising five distinct characters, each tailored to assist users in specific task areas. To visually distinguish the agents, each was assigned a unique color. Drawing from prior research on the primary purposes of using VAs (Pyae and Joelsson, 2018), we identified five prevalent task domains: *information seeking*, *entertainment*, *online shopping*, *smart home control*, and *personal tasks*. Nora (blue, female voice) specialized in responding to information-seeking tasks, including news and weather-related queries. Connor (green, male) took charge of entertainment-related tasks, such as music and video preferences. Johanna (red, female) handled online shopping inquiries. Lucy (purple, female) was responsible for smart home-related tasks, and Max (orange, male) focused on personal tasks, including reminders, alarms, and shopping lists. Agent assignments, including gender and task domains, were arbitrary. While the environment and mechanics were consistent across both control and intervention conditions, the tasks varied to prevent redundancy. For instance, if users were required to set an alarm in one condition, they were tasked with setting a reminder in the other condition.

We conducted a within-subject study with 20 participants, comparing the two conditions. Participants were tasked with completing 12 assignments, with each agent responsible for at least two tasks in the multi-agent condition. Participants were assigned a series of typical tasks involving a home assistant, triggering expected system reactions

such as switching lights, playing music or videos, locking doors, inquiring about the weather, and making online purchases. At the beginning of each session, the VR provided a concise introduction, outlining its capabilities to assist the user. In the multi-agent condition, individual agents introduced themselves, specifying their designated task domains and emphasizing their collective affiliation within the same system.

Measures

Following each round, participants completed post-exposure questionnaires, including the SUS (Brooke et al., 1996) and the User Experience Questionnaire (UEQ) (Laugwitz et al., 2008b). The experiment concluded with brief semi-structured interviews, where participants shared their overall opinions and provided insights into the potentials and challenges of the multi-agent voice assistant system.

Findings

Regarding user experience, for the subscales of perspicuity, efficiency, and dependability, all pragmatic qualities, no statistically significant differences emerged between the two conditions. This result was expected as pragmatic qualities relate to the perceived usability of a system, and there were no differences in system performance or functionality between conditions. SUS results affirmed the comparable usability of both systems, indicating no significant differences. This implies that the multi-agent approach is not more complex to learn or understand, offering ease, speed, practicality, and predictability similar to the single-agent VA.

On the other hand, participants gave significantly higher ratings to the user experience in the multi-agent condition for hedonic qualities, specifically in the subscales of *novelty* and *stimulation*. As anticipated, the innovative concept of receiving assistance from a team of agents was perceived as more novel. Moreover, the multi-agent approach received a significantly better rating in terms of *stimulation*, which could be attributed to the diversified interactions, reducing monotony during task performance, or the perceived support from a team of agents working collectively to assist the user. There were also indications of higher ratings for *attractiveness* in the multi-agent system. Furthermore, 70% of users preferred the multi-agent system overall, and an equal percentage found it more entertaining.

Regarding qualitative feedback, participants perceived different voices as distinct characters with personalities. They believed certain agents could be more suitable for specific task domains, conveying different character traits through voice factors like tone, gender, or accent. Expectations toward agent characters varied depending on their domain; for example, users preferred a trustworthy character for online shopping tasks involving sensitive data. The availability of diverse agents for distinct task domains emerged as a significant advantage, accommodating the need for individual character traits aligned with varied responsibilities. While we initially designed all agents to be as

neutral as possible for comparability, the study results highlighted the potential benefits of incorporating a variety of personalities. For instance, building a close relationship with an agent for entertainment or music could enhance user experience, while agents responsible for calendars or banking may prioritize dependability and trustworthiness. Consequently, we recommend tailoring the personality of a voice assistant agent to align with its assigned expertise. Participants encountered difficulty keeping track of different characters. Users needed more time to recognize agents and their domains, suggesting multi-agent systems should not exceed a certain complexity to prevent overwhelming users and facilitate establishing connections with individual characters. The complexity of interactions, influenced by the number of agents and human interlocutors, should align reasonably with the application's intensity and duration, considering contexts like business (daily and short), customer service (once and short), tutoring (short-term and intense), or at home (long-term and intense).

We crafted our multi-agent system to automatically select agents based on the task domain, aiming to prevent confusion and avoid overwhelming participants with the need to memorize agent names and domains. Despite this design choice, some participants wanted to manually assign task domains to individual agents, seeking a form of customization. Generally, addressing all user preferences equally with a single implementation is impossible. User responses in our study indicated the desire to customize different design factors of the voice assistant, such as the number of agents, their voices, and their roles.

In summary, our findings reveal promising potential for adopting multi-agent VAs and high user approval. Specifically, we observed significant differences in the hedonic aspects of the user experience, indicating the perceived value and enjoyment associated with multi-agent VAs.

5.1.2 | Multi-Agent Game Companions

This section is based on **Publication 4**:

Nima Zargham, Michael Bonfert, Georg Volkmar, Robert Porzel, and Rainer Malaka. **Smells Like Team Spirit: Investigating the Player Experience with Multiple Interlocutors in a VR Game**. In Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '20), Association for Computing Machinery.

My contribution to this work: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, validation, visualization, and contribution to all parts of the manuscript.

Next, we explored our multi-agent approach in a speech-based VR game to assess its impact on player experience. Collaborative multiplayer games emphasize teamwork to achieve shared goals, with massively multiplayer online role-playing games (MMORPG) being a notable example of this cooperative gameplay (Coulombe and Lynch, 2020). Previous research indicates that cooperative games can mitigate the adverse effects of violent video game play on cooperative behavior (Greitemeyer et al., 2012). For many players, the social aspects of online gaming are a crucial factor (Griffiths et al., 2004). In single-player games, this cooperative component can exist between the player and in-game characters. Some single-player games enable players to control multiple characters, each with a specific purpose or ability, placing the player in the role of a team leader, as seen in games like *Commandos* (Pyro Studios, 1998) or *Desperados: Wanted Dead or Alive* citedesperados. However, in this genre, the player is not a member of the team but rather in control of it.

In our study, we extended the concept of being assisted by multiple experts to a collaborative VR game where the player actively participates as a team member, engaging in natural language conversations with other in-game agents. Our game required bilateral voice interaction with in-game agents for successful gameplay. We developed two versions of the game: one where the player communicates with a single universally assisting character and another where the player interacts with a team of specialists (see Figure 5.3). The agents are only audible and provide relevant information to the player through radio communication to complete the level.

We conducted a preliminary qualitative within-subject study in which the participants (N=10) played both versions of our VR game. Our qualitative evaluation measures comprised observational notes from the experimenter, participants' comments and reactions during gameplay, and insights gathered through post-gameplay semi-structured interviews. In this work, we wanted to answer the following research question:

RQ: How does voice interaction with multiple interlocutors impact the player experience and perceived team spirit in a VR game?

Game Design

We designed a VR escape room game to assess our proposed multi-agent concept. The players' mission was to infiltrate a secure bank without being noticed while solving various puzzles and challenges. Agents connected via radio provided players with instructions and assistance throughout the game. Obstacles, such as security cameras, lasers, and door lock mechanisms, required players to consult with the agents. Mistakes with regard to the puzzles triggered the bank's alarm system, prompting an agent to disarm it, allowing players to continue the mission. Agent responses were triggered by the player's position in the virtual environment, actions, and voice commands. For instance, approaching a closed door prompted the agent(s) to assist in finding the security code. In the multi-agent

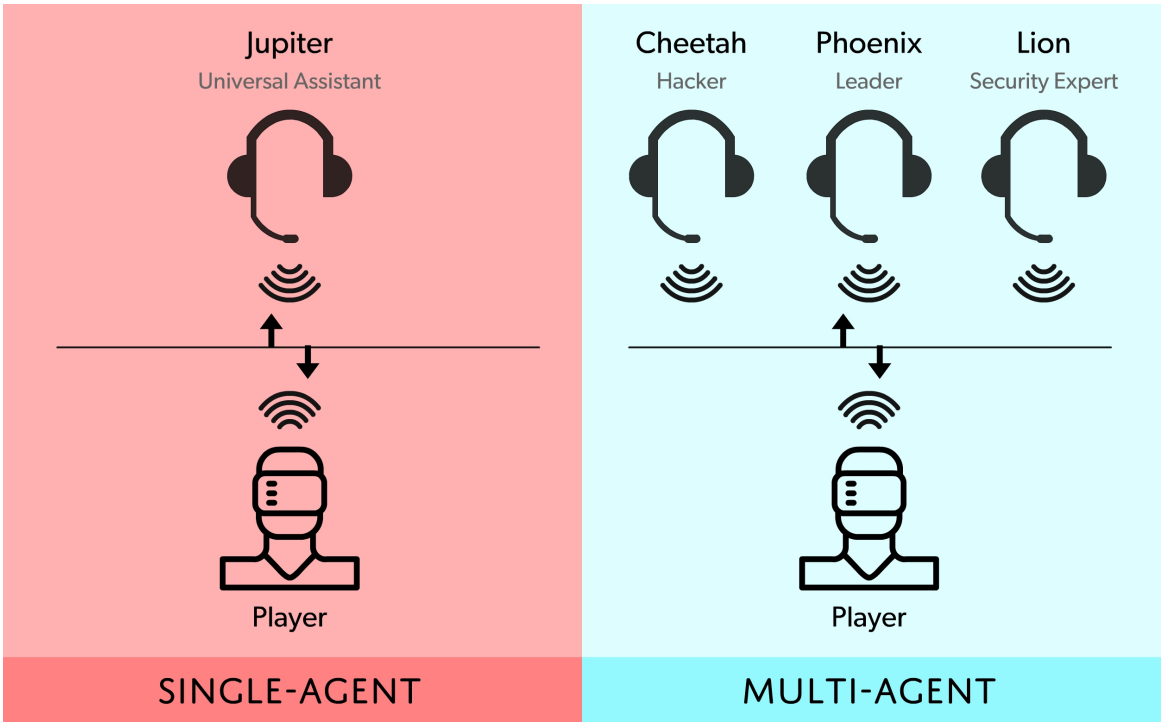


Figure 5.3: Schematic illustration of the two interaction conditions. (left) In the single-agent version of the game, the player interacts with one universal assistant. (right) In the multi-agent condition, a team of three characters supports the player, each with their unique expertise.

condition, we created a team of three characters: Phoenix, the leader; Lion, the security expert; and Cheetah, the hacker. Phoenix and Lion had male voices, while Cheetah had a female voice, each assisting the player in their specialized area. For example, if the player needed help with lasers, Lion responded. In the control condition, a single agent named "Jupiter," with a female voice, assisted the player with all challenges. We utilized a text-to-speech tool [1] to generate all voices for the agents. Agents' responses were limited to those aiding the player in accomplishing the mission, and unrelated inquiries went unanswered. The mission, game environment, and mechanics remained consistent across both conditions. Players were required to experience both conditions sequentially. To ensure equal engagement, we designed distinct puzzle solutions for each condition. Upon starting the game, in-game agents briefly introduced themselves, with characters in the multi-agent condition additionally mentioning their specialized task domain.

Findings

Our findings indicate that participants perceived engaging in conversations with a diverse group of in-game agents as akin to being part of a team. They described multi-agent cooperation as teamwork and found this version of the game more exciting

[1] `https://ttsmp3.com`

and motivating. Overall, players enjoyed the speech interaction with the characters, with nine out of ten participants expressing a preference for the multi-agent version, finding it more entertaining. Similar to the support experienced in multiplayer gaming, the assistance from multiple agents gave players the impression of having their back, fostering a sense of approachability and protection, thereby reducing the feeling of loneliness in the game. Conversely, few players preferred the simplicity of interacting with a single agent, perceiving it as a faster route to success despite the response durations being consistent in both versions. They found the multi-agent interaction too complex and mentally taxing, preferring the simplicity and reliability of conversing with a single character. This suggests that the preference for a multi-agent approach may vary depending on the player's preferred style, with some prioritizing progress over exploration.

Our study revealed that some participants wished to see more human-like behavior from the agents, such as increased exchange among them or the inclusion of humorous discussions. This implies that participants viewed the agents as individuals with distinct opinions and personalities rather than interchangeable entities within a computer system. Notably, contrary to previous literature suggesting users' disapproval of conversations among multiple agents in a task-oriented setting (Luria et al., 2019), players in our game embraced this idea, which we attribute to the hedonic purpose of the game.

In summary, our findings demonstrate that players did indeed perceive interacting with multiple agents as a team, finding it more entertaining, feeling more motivated, and experiencing a sense of protection when conversing with a group of characters.

5.2 | Agent Embodiment

Most voice assistants are designed as voice-only systems, with the agents' personalities primarily conveyed through their voice characteristics, pre-configured personifications such as their names (e.g., Alexa or Siri), and the physical design of the device (Reeves and Nass, 1996; Bickmore and Picard, 2005; Beneteau et al., 2019). Nevertheless, research indicates that the visual characteristics of digital systems play a significant role in human-machine interaction, influencing users' trust, engagement, and perception of the agent's personality (Kiesler et al., 2008; Desai et al., 2009; Schaefer et al., 2016; Hernández-Trapote et al., 2008b). Human communication extends beyond verbal exchange, encompassing facial expressions and body movements that are essential for conveying information. Non-verbal communication, including visual cues, allows people to express information beyond the semantic content of the message, such as emotions and current mood (Castillo et al., 2018). Moreover, these non-verbal factors play a significant role in conveying the personality of the agent. The user experience with VAs is tied to how users perceive the personality of an agent (Reeves and Nass, 1996). Existing literature suggests that the virtual embodiment of agents has the potential to influence their perceived personality through factors such as appearance and behavior (Castillo et al.,

2018). This alteration in perception can consequently impact users' trust and engagement with such devices (Zhou et al., 2019; Cafaro et al., 2016). Additionally, incorporating a visual dimension in communication offers benefits in enhancing accessibility for individuals with hearing impairments.

Currently, the visual representation of home assistants is typically confined to their outer casing and abstract light animations, serving as signifiers to convey the assistant's states to users. To further convey personality and human characteristics, an embodiment can enrich speech agents. In this context, embodiment refers to representing these agents in a physical or virtual form. This includes providing the virtual agent with a visualization, such as a virtual body or avatar, and enabling it to showcase realistic movements, gestures, and expressions. HCI Researchers have emphasized the convenience of embodied virtual agents, noting that interactions with such systems are generally perceived as more natural compared to agents without embodiment (Andrist et al., 2017; Cassell et al., 1999; Wang et al., 2019; Cassell and Thorisson, 1999). One product category of home assistants comes equipped with a screen to display visual output, enabling the virtual agent to be embodied, potentially elevating the interaction experience. These devices are commonly referred to as 'smart displays' (Shalini et al., 2019b). Additionally, we recognize the screen's potential to enhance the visual presence of the virtual agent.

Building on previous research regarding embodied agents and their appearances, this section delves into people's preferences for the virtual embodiment of VA agents. We present two empirical studies investigating users' preferences for virtually embodied agents. Publication 5 explores the degrees of visual realism and the social implications of a continuously present agent in the room, while Publication 6 investigates people's preferences for the visualization of a virtual home assistant through an online survey, taking into account individual user characteristics and their connection to preferences. The following provides insights into these two studies.

5.2.1 | Realism of Rendering

This section is based on **Publication 5**:

Michael Bonfert, Nima Zargham, Florian Saade, Robert Porzel, and Rainer Malaka. **An Evaluation of Visual Embodiment for Voice Assistants on Smart Displays**. In Proceedings of the 3rd Conference on Conversational User Interfaces (CUI '21), Association for Computing Machinery.

My contribution to this work: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, part of software, supervision, validation, visualization, and contribution to all parts of the manuscript.

Smart displays expand on the smart speaker concept by incorporating a touchscreen. Despite the addition of the visual modality in this device variant, the virtual agent is typically represented solely through auditory output and remains invisible in most current products. This work presents an empirical study focusing on user interactions with a smart display, where the virtual agent is embodied with a humanoid representation. In a between-group experiment, we compared three conditions: no agent embodiment, a digitally rendered embodiment, and a photorealistic embodied agent performed by a human actress.

When discussing realistic visualization of agents, the concept of the "uncanny valley" effect inevitably comes to the forefront. This phenomenon arises when the level of realism falls short and observers perceive anomalous features (Mori et al., 1970). Extensive research has explored the uncanny valley effect across various entities, including dolls, masks, facial caricatures, movie characters, avatars, and embodied agents (Seyama and Nagayama, 2007). While realistic humanoids can be appealing to users (McDonnell et al., 2012; MacDorman et al., 2009), achieving this requires a delicate balance of social responsiveness and aesthetic refinement (Hanson et al., 2005). Even abstract faces, common in computer-generated renditions, can evoke an eerie feeling (MacDorman et al., 2009). To avoid the uncanny valley effect, researchers suggest maintaining consistency in realism while allowing intentional stylization (Schwind et al., 2018). To avoid uncanny valley effects in our evaluation, we opted for an actress to perform the agent in the photorealism condition. For most practical applications, this is not an ecologically feasible solution, but it provides clearer results in the context of this study.

This study pursued the following research questions:

RQ1: How does the user experience change if a voice assistant agent is visually embodied on a smart display?

RQ2: How does the degree of visual realism of the embodied agent influence the user experience?

Prototype Design

We developed three versions of a smart display for evaluation: one without an agent visualization, one with a digitally rendered, artificial embodied agent, and one with a prerecorded, photorealistic embodied agent. All versions had the same functionality and only differed in appearance. We selected a female agent to align with the prevalent representation of female assistants in current consumer products to avoid a novelty bias (Hwang et al., 2019).

Disembodied Agent (DEA): This version mirrors the current smart display status quo, presenting no agent embodiment. Users only hear the agent's voice, generated with an online TTS tool.

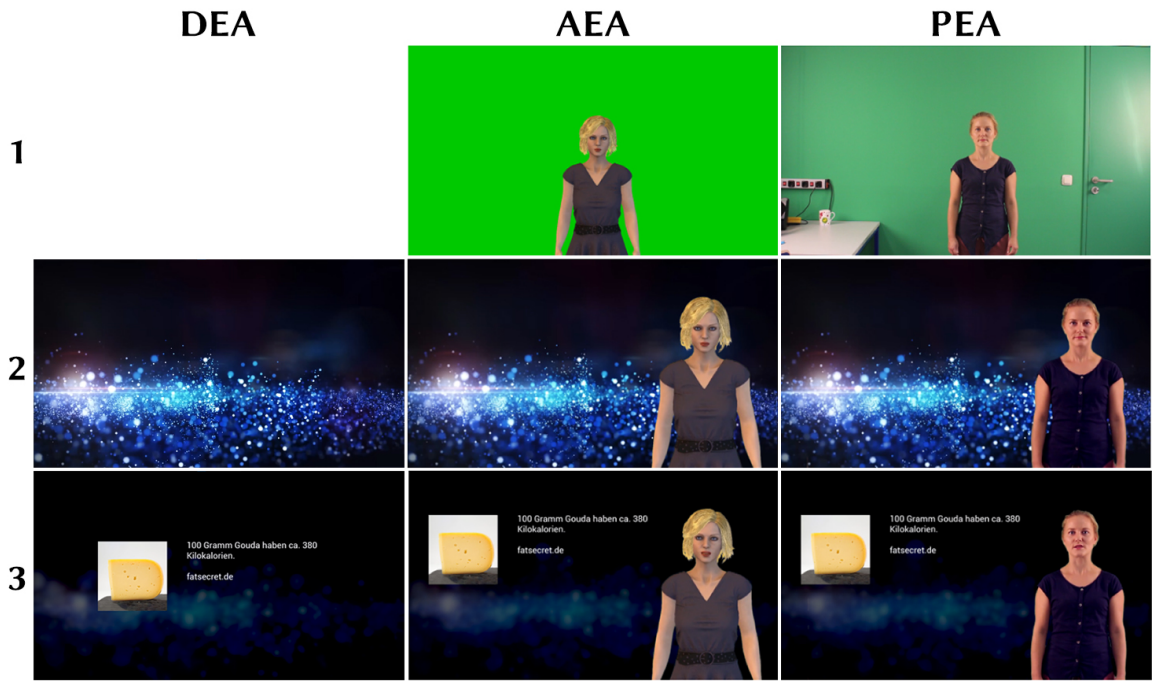


Figure 5.4: The three stages of the video creation process for the three conditions: recording or rendering with a green screen, replacing the background, and augmenting with information cards.

Artificial Embodied Agent (AEA): This version used a digitally rendered, animated agent representation on the smart display. The character, resembling a news anchor in *The Sims* (Maxis, 2000) style, is a female with blonde hair, light-colored skin, and a dark blue dress. An actor's performance, captured via a webcam, served as input for FaceRig[2], animating the virtual character. The video output synchronized with the TTS voice used for the DEA condition.

Photorealistic Embodied Agent (PEA): A theater actress was recorded for this prototype version, embodying a photorealistic character. She was instructed to match the artificial character in intonation, facial expressions, and body language to resemble the AEA visually. Further, the actress and her clothing were selected to resemble the AEA visually. We refrained from using the TTS audio to avoid a mismatch of visual and auditory coherence and complications with lip synchronicity.

We prepared a set of standard tasks representing a morning scenario for users to accomplish with the VA, covering a range of everyday commands based on typical home assistant usage (Kinsella and Mutchler, 2019). Tasks included turning on the light, playing music, retrieving information, setting a timer, or ordering a product online. If users asked questions outside the command list, the system clarified its inability to assist. For smart display output, we created media snippets for each condition, featuring a dynamic

[2] `https://facerig.com`

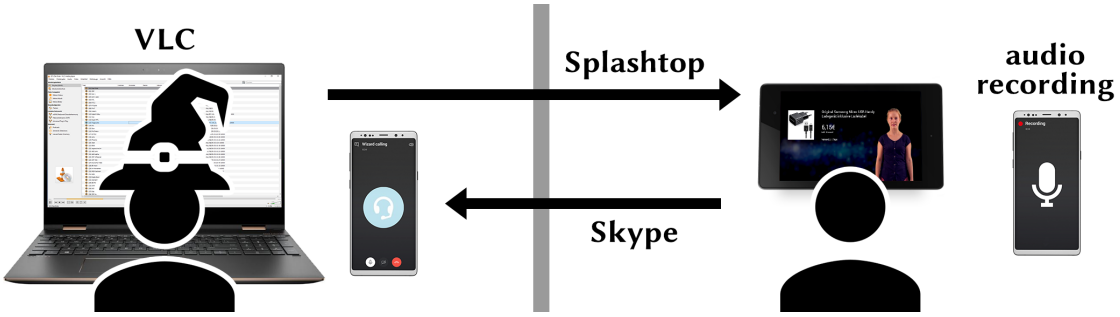


Figure 5.5: Experiment setup: The Wizard listens to the user's commands via Skype and triggers appropriate video snippets in the VLC media player, which is transferred to the smart display. An audio device records the experiment.

background, an information card, an audio track, and agent embodiment where applicable (see Figure 5.4). Information cards displayed text and images relevant to user commands, appearing during the assistant's response and fading out post-task. Dynamic idle videos looped between tasks, with transition smoothness consistent across conditions. Agent embodiments were overlaid on the bottom right without overlapping information cards for AEA and PEA, while the DEA condition centered its screen layout to avoid empty spaces.

For our evaluation, we employed a Wizard of Oz approach (Kelley, 1983; Maulsby et al., 1993) to ensure reliable system operability, with the Wizard controlling the smart display from an adjacent room (see Figure 5.5). We evaluated our prototypes in a between-group design study with 60 participants.

Measures

We asked participants to fill out both the UEQ (Laugwitz et al., 2008b) and the AttrakDiff Short Questionnaire (Hassenzahl et al., 2003), which are validated instruments assessing pragmatic and hedonic qualities and system attractiveness. These measures, with a similar theoretical basis, provided data for reliable comparisons. Subsequently, semi-structured interviews covered topics such as reliability, trust, agent appearance, individual preferences, and on-screen presence. Despite the between-groups approach, we showed the participants the alternative system versions for comparison at the end of the session.

Findings

Approximately one-third of our sample (35.7%) favored the current state of a smart display without an agent depiction. Pragmatic reasons dominated this preference, considering it unnecessary, distracting, and occupying space that could be better utilized for more pertinent content. However, when assessing pragmatic qualities, all systems received similar ratings, indicating no distinct advantages in terms of efficiency, clarity, speed, or

predictability by not displaying an agent. Half of the participants (51.8%) favored the version with a photorealistic agent (PEA). Only one out of eight users (12.5%) favored the artificial visualization (AEA). Users expressed skepticism about the artificial embodied agent, deeming it 'not human enough.' Contrarily, some appreciated the deliberate cartoony realization, finding the humanoid shape conducive to a human-like conversation style while maintaining the interlocutor's artificiality.

The results from the two standardized UX questionnaires (UEQ and AttrakDiff Short) showed no statistically significant differences among the conditions in terms of pragmatic qualities, hedonic qualities, or the attractiveness of the prototypes. However, our qualitative findings contradicted this, revealing influences caused by the embodiment of the agent on UX beyond the measurements of the standardized instruments we applied. The contrast between the results may arise from the broad range of UX aspects the universally applicable questionnaires cover. We witnessed that engaging with an embodied agent offered a distinct advantage, emphasizing higher subjective trustworthiness, supporting prior research (Hancock et al., 2011; Schaefer et al., 2016). The interaction was also perceived as more natural, aligning with existing research (Takeuchi and Naito, 1995; Koda and Maes, 1996). The visual presence of the agent heightened approachability and dependability compared to a voice-only scenario. We also witnessed that the preferences regarding the details of embodiment varied, with some participants emphasizing the importance of visually focusing on the interlocutor, regardless of appearance. Others desired obvious machine-like features to avoid the misconception of interacting with a human. Abstract and non-humanoid representations were suggested, and playful concepts, such as fictional characters, animals, or mythical creatures, were proposed by some participants.

Approximately half of the participants indicated that the agent's gender was unimportant. However, three-quarters of users with a preference leaned towards a female agent, aligning with the literature on gender stereotypes with conversational agents (Brahnam and De Angeli, 2012). A subset expressed a desire for an attractive agent with customizable features, such as preferred hair color, consistent with previous research on preferences for attractive embodied agents (Khan and De Angeli, 2009). Our results also showed that the perceived age of the agent influences the user's reliability assessment, aligning with previous findings by Marin et al. (Marin Mejia et al., 2013). Participants preferred an agent that appears experienced, avoiding an overly young appearance. However, younger users expressed the importance of the agent not looking significantly older than themselves to avoid feeling patronized. The consensus among participants was that the ideal age of the agent should be similar to their own.

Participants perceived the continuous display of the AEA and PEA agents differently. For some, it indicated that the system was online and ready, while others thought the device would be listening to commands non-stop. This led to users omitting the wake word and feeling ignored by the attentive-looking agent. Additionally, users expressed discomfort with the agent staring at them during idle time, creating a feeling of constant

observation. Some even felt the agent was waiting impatiently until assistance was needed again. To address this, we recommend hiding the agent between tasks to avoid social awkwardness and domestic intrusion. Reappearance can serve as feedback, indicating that the system has recognized the wake word and is ready to listen to commands. The transition could be implemented as a fading effect or, for instance, with the agent walking in and out.

In summary, our findings indicate that embodiment is not beneficial, in principle, but rather contingent on its implementation and the preferences of its users. Overall, we identified diverse and conflicting preferences, suggesting that a one-size-fits-all solution may not meet all users' expectations equally. Thus, we propose providing smart display users the flexibility to choose whether an embodied agent is displayed and customize its appearance based on individual preferences.

5.2.2 | Embodiment Preferences

This section is based on **Publication 6**:

> Nima Zargham, Dmitry Alexandrovsky, Thomas Mildner, Robert Porzel, Rainer Malaka. **"Let's Face It": Investigating User Preferences for Virtual Humanoid Home Assistants**. In Proceedings of the 11th International Conference on Human-Agent Interaction (HAI '23), Association for Computing Machinery.

My contribution to this work: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, validation, visualization, and contribution to all parts of the manuscript.

Our previous study highlighted notable differences in user preferences for embodied home assistants. The diversity in individual preferences poses a challenge for systematic agent adaptation to users, making a universally satisfying design impractical. Building on our initial exploration of virtual agent embodiment, we aimed to delve deeper into user preferences regarding specific visualization aspects. Our goal was to investigate the connection between users' own characteristics and their preferences for the agent's embodiment.

One principle transferred from human-human interaction to human-agent interaction is the similarity-attraction principle, suggesting that people are attracted to those they perceive as similar to themselves (Bernier and Scassellati, 2010). This principle, known to positively correlate with initial interpersonal attraction in human-human interaction (Byrne and Nelson, 1965), offers a straightforward and reliable approach for influencing interactions in human-agent scenarios. Nass et al. (Nass and Lee, 2001)

propose leveraging this principle in design to enhance product satisfaction and foster positive impressions toward the producing company. A previous study by Bernier et al. highlighted that people rated a social robot more favorably when it displayed preferences similar to their own (Bernier and Scassellati, 2010). While the similarity-attraction effect has been explored in different contexts, previous research has mainly explored it regarding behavior and personality (Bernier and Scassellati, 2010; Chen and Kenrick, 2002; Condon and Crano, 1988). However, a gap exists in the literature concerning the application of this principle to preferences for the visual embodiment of agents.

In this work, we focus on exploring people's preferences for humanoid visualizations of home assistants with respect to users' perceptions of their own characteristics. We kept our focus on humanoid visualizations, as past studies consistently highlight their benefits in human-computer interaction, enhancing human-likeness, likability, and the perception of shared reality (Salem et al., 2013). It has also been shown to foster a sense of familiarity, credibility, trust, and attachment (Reeves and Nass, 1996; Seymour and Van Kleek, 2021; Nowak and Rauh, 2008; Yuan and Dennis, 2019).

We conducted an online survey ($N = 78$) to gather insights into people's preferences regarding the visualization of humanoid assistants. Our study aimed to address the following research questions:

RQ1: How do users imagine the visualization of their desired humanoid home assistant?

RQ2: What is the relation between users' own characteristics and their preferences for virtual assistants?

Survey Design

We conducted an online survey to analyze users' preferences regarding the embodiment of home assistants and how these preferences relate to users' perceptions of their own characteristics. The survey began with demographic questions covering age, gender, ethnicity, nationality, native language, and accent or dialect. This was followed by questions about participants' appearance, such as hair color and body shape. Participants were then asked to express their desired attributes for their home assistant, encompassing agent gender, age, ethnicity, body shape, hairstyle, hair color, and outfit. They were also asked to identify the most important facial features and rate the significance of the agent's attractiveness to them. Participants also provided insights into the desired personality of the assistant, including emotional expressions and appropriateness of specific emotions. The survey concluded with participants rating the individual importance of the assistant's looks, voice, and personality on a seven-point Likert scale.

Findings

A significant number of participants preferred virtual agents sharing demographics similar to their own. This preference was also evident in correlations between participants'

age, language, accent, and dialect, as well as the characteristics they desired in the virtual agent. Most participants with preferences for the agent's ethnicity and hair color selected options matching their own. These findings suggest a preference for virtual agents that are perceived as more relatable, aligning with previous research indicating a preference for technologies resembling one's own characteristics (Breazeal, 2003). In line with our previous research (Bonfert et al., 2021) (Publication 5), we witnessed a strong preference among participants for a virtual agent with a youthful yet mature appearance. Participants found elderly-looking agents to seem unfit and very young-looking to be immature. We believe that the assumed age of the agent may influence users' perceptions of its capabilities and reliability. Additionally, participants favored an average body shape, associating underweight and overweight with lower fitness. This preference may also align with societal standards of attractiveness. Our findings indicate a preference for an agent that appears healthy, mature, and competent.

Although nearly half of the participants claimed that the attractiveness of the agent was unimportant, no one wanted the agent to appear less attractive than themselves. Participants rated the importance of attractiveness above average. Specific preferences for the agent's hair color and outfit underscore the significance of the agent's attractiveness. Users generally favored a more human-like appearance and realistic rendering over abstraction. These align with prior research on agent visualization (cf. (Yuksel et al., 2017; Khan and De Angeli, 2009; Banakou et al., 2009)), emphasizing that agents should have a minimum degree of visual appeal. Attractiveness is often linked to positive qualities like popularity, competence, and desirability (Lorenzo et al., 2010). Research indicates that attractive communicators tend to achieve greater opinion agreement (Wiedmann and Von Mettenheim, 2020). Societal emphasis on physical attractiveness and its perceived benefits may influence preferences for attractive virtual agents. People may believe that an attractive virtual agent signifies better design, higher quality, or more advanced technology, leading to a preference based on perceived societal standards.

Concerning the agent's gender, participants generally preferred an androgynous presentation, deviating from some prior findings on gender stereotypes of agents (Hwang et al., 2019; Brahnam and De Angeli, 2012). Additionally, we found no significant differences in preferences between male and female participants. In line with the study by Nag et al. (Nag and Yalçın, 2020), we witnessed that gender stereotypes were not as effective as previously assumed for virtual agents. While a few participants sexualized the agent, particularly regarding the outfit, a more significant proportion preferred a standard and modest appearance. Many participants expressed a desire for the agent's outfit to avoid being revealing or sexualized. Over half of the participants had specific outfit preferences, viewing this aspect as playful. Some even suggested it could convey important information such as calendar events or weather forecasts.

Few participants expressed a desire for the agent to display dynamic behavior and autonomous actions, suggesting a preference for a character with a life of its own. Examples include participants wishing for the agent to change hairstyles, colors, or outfits

without user interventions. Participants also provided feedback on the agent's on-screen presence. Many suggested incorporating random actions, like reading a book or taking a nap, to simulate a routine life for the agent. However, this preference might stem from a desire to avoid feeling constantly observed, as discussed in our prior research (Bonfert et al., 2021).

Overall, our findings indicate that users still prioritize the pragmatic aspects of home assistants. Most participants emphasized the importance of clear and understandable language, accent, and dialect in virtual agents, prioritizing effective communication. Participants regarded all three aspects of the agent's voice, personality, and appearance as important. However, they placed a higher priority on the agent's voice and personality over its visual appearance. This underscores that elements such as understandability, approachability, and the agent's character are more vital in shaping the participants' perception of the agent than its physical appearance. Nevertheless, there is a notable appreciation for the agent's hedonic qualities. Users preferred an agent that appeared mature, healthy, relatable, and attractive. We found that users prioritize demographic similarities for relatability rather than direct resemblance. Preferences include age, language, accent, dialect, ethnicity, hair color, and body shape. The diverse preferences underscore the subjectivity in individual participant choices, emphasizing the need for a customizable solution. We argue that granting users the ability to modify various agent characteristics could enhance satisfaction and engagement.

5.3 | Customization and Personalization

This section is based on **Publication 7**:

Nima Zargham, Dmitry Alexandrovsky, Jan Erich, Nina Wenig, and Rainer Malaka. **"I Want It That Way": Exploring Users' Customization and Personalization Preferences for Home Assistants**. In Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22), Association for Computing Machinery.

My contribution to this work: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, validation, visualization, and contribution to all parts of the manuscript.

As observed in the studies discussed earlier, due to the variety in individual preferences for a desired home assistant, the systematic adaptation of these systems is highly challenging. Consequently, commonly available devices often adopt a one-size-fits-all approach, overlooking the potential advantages of tailoring the experience to individual user preferences. Studies have demonstrated that customization and personalization features

contribute to heightened user satisfaction and improved performance (Cowan et al., 2015, 2016; Molnar and Kletke, 1996; Murad et al., 2018; Choi et al., 2020). Customized features are explicitly chosen by the user from specific options, while personalized features are dynamically generated by computers in response to individual user needs (Nielsen, 1998). Moreover, prior research has shown that users tend to unconsciously attribute personalities to conversational agents (Reeves and Nass, 1996) and that users' perception of the agent's personality could greatly impact users' trust and engagement with the device (Zhou et al., 2019; Cafaro et al., 2016).

In this work, we adopted a user-centered approach, employing semi-structured interviews with storyboards featuring everyday domestic scenarios to uncover user preferences for customization and personalization of home assistants, along with their desired personality types for the agents.

More specifically, we looked into the following research questions:

RQ1 How do users imagine the personality of their desired home assistant?

RQ2 In which ways do users want to customize and personalize their home assistants?

Study Design

We conducted an exploratory study to gather users' preferences for customization and personalization preferences, as well as the desired personality of home assistants. Inspired by scenario-based design methods (Carroll, 1999) and vignette experiments (Aguinis and Bradley, 2014), our approach involved presenting participants with a series of hypothetical situations, prompting them to reflect on these scenarios. This method enables the examination of technologies despite current technical limitations. To enhance visualization, we employed graphical storyboards illustrating the spatial configuration of the home environment, the user(s), and the smart speaker within that setting (see Figure 5.6).

Through iterative discussions, we developed a number of scenarios set in the home environment, depicting situations with either a single person or multiple individuals. The comprehensive process of scenario design is outlined in the paper. Ten diverse scenarios encompassing various domestic situations were chosen, and storyboards were developed for each selected scenario. We iteratively refined the design of the storyboards to minimize cultural and ethnic cues. Notably, the characters in the storyboards were intentionally designed without facial expressions to prevent any potential influence from character reactions.

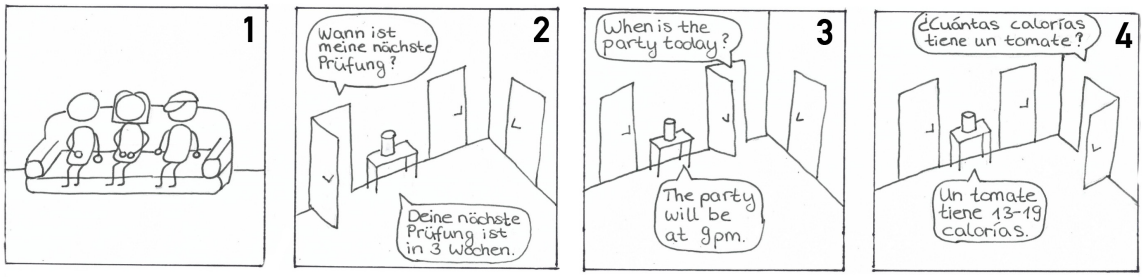


Figure 5.6: An exemplary storyboard that was used in the interview in which the home assistant responds differently to each household member.

Measures

We conducted online interviews via video calls with 15 participants. At the beginning of each session, participants disclosed their personality traits using the Ten Item Personality Inventory (TIPI) (Gosling et al., 2003), a 10-item questionnaire assessing personality based on the Five-Factor Model. Subsequently, the semi-structured interviews commenced. Participants were presented with all ten scenarios in random order and provided comments on each story. Throughout this phase, participants shared their impressions of individual scenarios, highlighting the positive and negative aspects of interactions with the home assistant. They offered general suggestions and recommendations related to customization, personalization, and the desired personality of a home assistant. Following the review of storyboards, participants completed a second TIPI, expressing their desired personality traits for the speech agent.

Findings

Our findings indicate participant dissatisfaction with home assistants' current customization and personalization features, highlighting a demand for additional and enhanced functionalities. We observed notable trends in preferences for customization and personalization features. Similar to previous research (Choi et al., 2020; Abdolrahmani et al., 2018), participants expressed a strong interest in modifying an agent's voice characteristics, including speed, tone, and volume. Participants valued features that could improve accessibility, such as simultaneous support for multiple languages. Creative ideas, such as making the agent sound or look like a celebrity, were suggested. Preferences for a humanoid digital embodiment were expressed, with requests for hairstyle, clothing, and body size customization. Most users favored direct control over these features rather than relying on personalization. However, concerns were raised about the complexity of adjustments, emphasizing the need for a user-friendly interface. Current systems' customization features often require users to navigate a graphical interface, such as a

smartphone app, to configure the VA. Difficulty locating specific customization options within the app was a common challenge, potentially leading to user frustration and complicating the customization process.

Participants generally preferred an agent that matched their personality and adapted to their interests. They also wanted adaptive systems capable of continuous evolution based on changes in user interest. Some users wanted the agent to occasionally change its visualization, suggesting the agent has a daily life in line with Publication 6. While most users desired emotionally expressive, humorous, and mood-sensitive agents that behaved more human-like, some expressed hesitancy toward these features. Concerns about the potential for technology-related mistrust were apparent among users who worried that inaccurately recognized moods or inappropriate responses could have adverse effects. Additionally, a subset of users preferred a clear distinction between a human and software, finding human-like features in agents to be artificial and disliking such interactions.

Users consistently rated their desired agent's personality as equal to or higher than their own across all five TIPI scales. Notably, there were significantly higher ratings in terms of agreeableness, aligning with previous research indicating a preference for agreeable agents among users with higher agreeableness (Völkel and Kaya, 2021). Participants also rated conscientiousness and emotional stability higher for the agent than their self-ratings. These traits are associated with reliability, an essential factor in human-agent communication.

Our participants raised privacy concerns as a significant challenge for home assistants. Users expressed the need for more privacy-related features, including individual user profiles, user roles, and access control, to build trust and prevent inconvenient situations.

Overall, in this work, we identified four categories of features users wish for home assistants: *agent's speech characteristics*, *agent's visualization*, *agent's personality*, and *privacy and security*. Once again, we observed that users' preferences regarding agent representation often vary, and no universal solution could satisfy the expectations of all users equally. Therefore, to improve user satisfaction and enhance user agency, we recommend providing more customization and personalization features on the dimensions identified in order to adjust systems to individual users.

5.4 | Conclusion

This chapter presented five research papers exploring the representation of virtual speech agents. These covered the number of agent characters or personas, the visual embodiment of the agent, the agent's personality, and users' personalization and customization preferences for such systems. We employed various methods, including surveys, interviews, a Wizard of Oz study, and lab experiments, to address the gaps in the literature regarding the dimensions we explored. The introduction of the multi-agent concept represents a novel approach to representing the system, aiming to create the perception that users are assisted by a team of experts rather than a single assistant.

Investigations into the embodiment of agents on smart displays aimed to add a new visual dimension to home assistant interactions. By examining users' preferences for virtual assistant visualization, we provide valuable insights into aligning these systems with user expectations. Moreover, identifying dimensions for agent personalization and customization can assist designers and developers in creating more user-centered systems.

The findings from these studies underscore the significant impact of the appearance and representation of speech systems on users' experiences, even though pragmatism remains a top priority for most people. One crucial observation derived from these works is that a universal design for such systems to satisfy all users is not possible when it comes to the hedonic aspects. Our investigation highlighted the key role that individual user characteristics play in shaping preferences for agent representation. People have very diverse expectations, and our findings repeatedly pointed to the inherent subjectivity in individual choices. User's unique personality traits were shown to contribute to their preferences for agent characteristics. Our research consistently revealed participants' interest in defining specific personality traits for agents aligning with their interests. Designers and developers should acknowledge this diversity and empower users by allowing them to customize various agent characteristics to enhance user satisfaction and engagement with such systems.

Nevertheless, despite these individual differences, certain themes and trends emerged that can inform the design of such systems. For instance, users expressed a strong desire for agents that appeared relatable, reflecting their own demographics. This observation underscores the influence of users' cultural backgrounds on their preferences, emphasizing the importance of incorporating cultural considerations into the design of virtual agents. Moreover, our findings shed light on the impact of societal standards, particularly on the attractiveness of the agent. Users, consciously or unconsciously, favored agents that aligned with societal norms of attractiveness. Acknowledging these dimensions is critical for designers seeking to create virtual assistants that resonate with users on a personal and cultural level.

An interesting finding that emerged from our exploration of the agent's embodiment was that users perceived the agent as more approachable when it had a visual representation. This visual representation served as a reminder of the system's availability for interaction. However, some participants expressed discomfort when the agent was embodied, feeling as if they were being watched in their homes. This highlights a potential downside to visual embodiment, as users may feel added social pressure and raise privacy concerns. A recurring theme across three of our studies was the desire for the agent to possess a sense of autonomy. Users preferred an agent that did not solely focus on them but had their independent activities while being approachable when needed. This may relate to a sense of realism and relatability, as users perceive the agent as more than just a tool but rather as a dynamic entity within their environment.

One of our aims when proposing a multi-agent system was to tackle the issue of gender stereotyping prevalent in voice assistants. While our approach may not completely

eliminate these biases, it presents an opportunity to mitigate them by offering a diverse range of genders within the system. By introducing a variety of agents with different genders, we aim to challenge existing stereotypes and promote inclusivity in technology. We encourage designers and developers to consider leveraging multi-agent systems as a strategy to diversify the voice assistant market, ultimately fostering a more balanced and representative landscape in artificial intelligence.

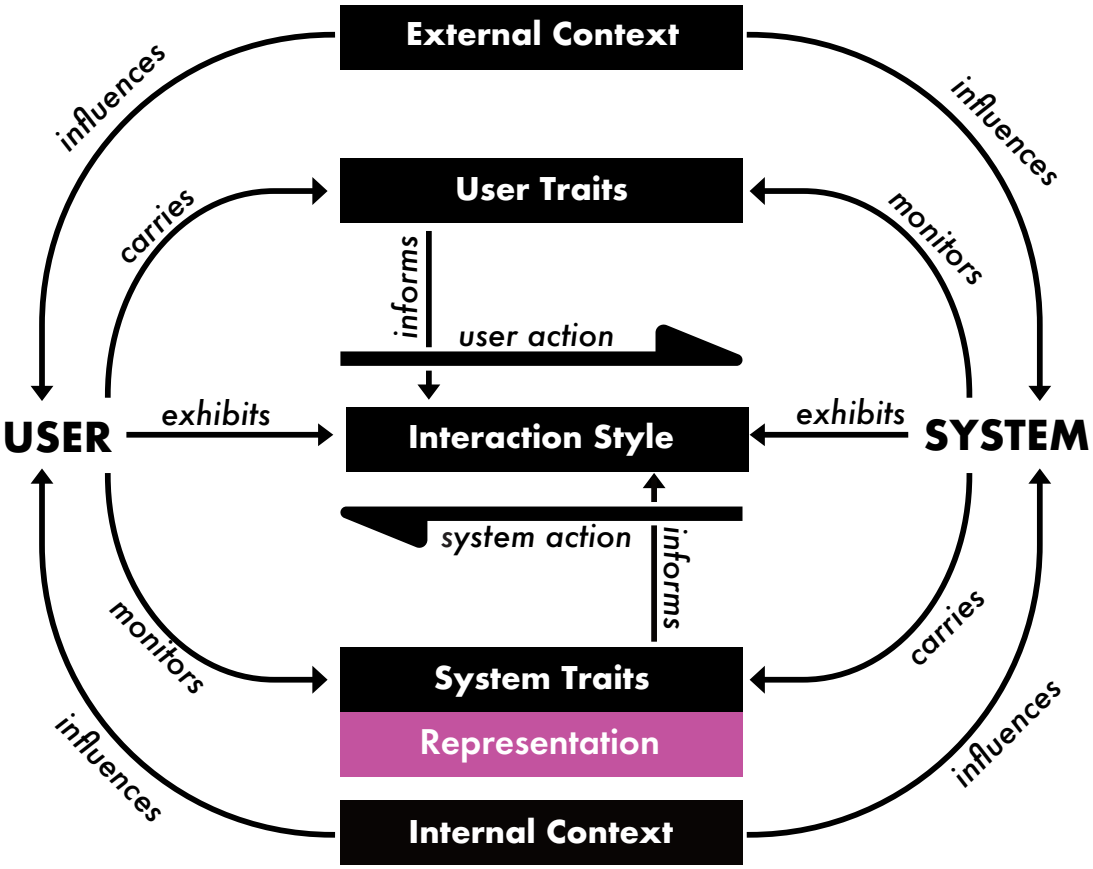


Figure 5.7: The role of system representation in the HASI model.

Reflection on the HASI Model

Looking back at the proposed HASI model, the representation dimension investigated in this chapter also falls under the system traits (see Figure 5.7). Exploring this dimension, we observed several interrelations between the individual factors of the HASI model. Our research highlighted that socio-cultural factors (external context), such as societal standards, influence users' expectations and preferences in agent representation. In certain situations, these lead to the formation of stereotypes, which can have harmful effects. Recognizing this, it becomes critical for designers and developers to navigate

these socio-cultural dynamics thoughtfully and strive for more inclusive and diverse representations within speech agents. Moreover, our findings emphasized the influence of user traits, such as individual user personality or cultural background, on interaction with speech agents. We observed the need for agents to resonate with user characteristics to ensure relatability. Incorporating these dimensions into the design process can lay the foundation for more personalized systems. Speech systems can adjust to individual user traits, such as their personality and preferences, as well as their ethnicity and cultural background, and tailor agent representations to align with users' diverse characteristics. This can potentially enhance users' agency and satisfaction when interacting with VUIs. Such approaches can contribute to a more nuanced and user-centric design, improving the overall user experience with speech systems. Overall, studies in this chapter highlighted that the alignment of system representation with user traits can be beneficial for the overall interaction experience.

In response to the **TRQ3**, we argue that the representation of virtual speech agents significantly contributes to a better interaction experience. While pragmatic aspects remained a priority, the studies presented in this chapter highlighted the users' appreciation of hedonic aspects of speech interaction. This is evident in our study, where participants preferred agents resembling their demographics. This preference extended to age, language, accent, dialect, ethnicity, and body shape, emphasizing the importance of relatability. The findings emphasize that users value agents having characteristics similar to users and complying with societal standards of attractiveness. We witnessed that using a multi-agent system rather than a single agent could give users a higher user experience with such speech systems. People perceived such systems as a group working together to assist the user, leading to a feeling of teamwork and a higher sense of support. The degree of realism with regard to visual rendering was also noted as a factor that could impact the interaction. Users preferred more realistic systems that could better resemble a human. Ultimately, different design decisions regarding the agent's representation can influence user experience with such systems. However, It is crucial to recognize that these influences often operate on an individual basis and cannot be collectively categorized. Designers should acknowledge and navigate this inherent subjectivity, steering away from a one-size-fits-all approach. Instead, the emphasis should be on developing agent representations that align with individual user preferences. This approach calls for the implementation of personalized and customized systems, empowering users to tailor various aspects of the agent to their liking. By prioritizing individuality in design and allowing users agency in customization, designers can enhance overall user satisfaction and engagement with virtual speech systems.

Our investigation primarily focused on exploring dimensions of agent representation within home environments and briefly extended to speech-based video games. While our findings offer insights applicable across various contexts, there remains a need for further studies to delve into agent representation within other domains of use. Examining the nuances of how users prefer their virtual agents in workplace settings, educational

environments, or healthcare scenarios could reveal domain-specific preferences and requirements. Moreover, the necessity for long-term studies becomes evident to comprehend the lasting impact of interacting with different agent representations. Investigating user perceptions, preferences, and experiences over extended periods will provide a more nuanced understanding of how these interactions evolve and whether preferences change over time. Additionally, while we discussed ethical concerns linked to visualizing virtual agents, including the reinforcement of stereotypes, we highly encourage future studies to delve deeper into these issues. It is essential to acknowledge that the studies discussed in this chapter cover only a fraction of the potential design possibilities for virtual agent representation. The realm of possibilities remains vast, and future research continues to explore and innovate in this dynamic and evolving field.

6

Interaction Style

The interaction style of speech systems is a crucial factor impacting user experience in HASI. The design choices in how these systems engage with users directly impact how users perceive the system and its capabilities (Porcheron et al., 2018). How agent comments are formulated, how they are executed, and the system's responsiveness all play integral roles in shaping the overall user experience. The tone and language employed by the speech system are essential in creating a conversational atmosphere, affecting the user's engagement and comfort levels. Additionally, the timing and pacing of agent responses are significant contributors to user satisfaction, with delays or rapid-fire replies potentially disrupting the natural flow of conversation (Cha et al., 2020). In essence, the interaction style of speech systems is multifaceted, incorporating linguistic elements, responsiveness, adaptability, and contextual understanding. This chapter of the dissertation discusses the dimension of interaction style within speech agents. The primary goal is to adopt a human-centered perspective to understand the interaction styles that are most suitable for users by focusing on their preferences and expectations.

Typically, the interaction with speech systems is uni-directional, where users have to initiate the interaction with a command, making these systems reactive (Reicherts et al., 2021). The user interaction with the system starts with initiating a wake word, followed by a user inquiry, after which the agent responds. However, this interaction model restricts the system to user-triggered interactions, thus limiting its capabilities. This chapter explores the potential for enhanced adaptability in these systems by delving into the proactive features of speech agents. We aim to investigate how incorporating proactive elements can expand the system's functionality beyond user-initiated interactions.

This chapter investigates **TRQ4**:

- What is the appropriate *interaction style* in speech systems for domestic activities?

6.1 | Proactivity

As AI, NLP, and sensing technologies advance, speech systems evolve in their ability to comprehend their surroundings and understand users' preferences, activities, and intentions. This progress opens the door to a heightened potential for proactivity in these systems (Miksik et al., 2020; Schmidt and Braunger, 2018; Kraus et al., 2020; Edwards et al., 2021). Proactive behaviors from VAs are considered agent-initiated interactions triggered by events related to the user(s) and their environment. This stands in contrast to user-initiated inquiries or pre-configured actions, such as reminders, alerts, or routines set by the user.

Recent studies have begun to examine proactive behavior in VAs (Miksik et al., 2020; Schmidt and Braunger, 2018; Kraus et al., 2020), exploring the timing of proactive VA interactions (Cha et al., 2020) and how to design such interventions (Edwards et al., 2021). Previous literature underlines the opportunities and benefits that proactive VAs can offer in supporting, probing, or inspiring people (Wei et al., 2021). A study by Schmidt et al. (Schmidt and Braunger, 2018) revealed that users highly favor proactive features. Similarly, an elicitation study by Völkel et al. (Völkel et al., 2021) on users' envisioned dialogues with a perfect voice assistant showed that many participants found proactive voice assistant behavior desirable. Users welcome these interactions if they provide timely and relevant information. However, several challenges with VA proactivity have been highlighted, the most dominant being privacy concerns (Tabassum et al., 2019a). As proactivity entails ongoing data collection from the environment and user activities to maintain context awareness, concerns arise regarding handling this data. Furthermore, users may see it as an intrusive interaction if the VA interrupts an ongoing conversation or the proactive interaction is not deemed helpful at that moment.

Overall, despite the highlighted potential for proactive features in VAs, current research still lacks a comprehensive understanding of how people perceive and feel about this interaction style in domestic activities. This section investigates proactivity for speech agents, aiming to understand people's perceptions of these agent-initiated interactions. We begin our exploration by examining users' perceptions of the usefulness, appropriateness, and invasiveness of proactive interactions initiated by an agent in a domestic setting (Publication 8). Additionally, we investigate the principles for designing desirable proactive interventions (Publication 9) and explore the potential role of humor in supporting proactive agent interactions (Publication 10). Our research aims to contribute to the ongoing discourse on this evolving interaction style by thoroughly examining its nuances and implications.

6.1.1 | Perceptions of Proactive Behaviour

This section is based on **Publication 8**:

> Leon Reicherts, Nima Zargham, Michael Bonfert, Yvonne Rogers, and Rainer Malaka. **May I Interrupt? Diverging Opinions on Proactive Smart Speakers**. In Extended Abstracts of the 3rd Conference on Conversational User Interfaces (CUI '21), Association for Computing Machinery.

My contribution to this work: Conceptualization, data curation, part of formal analysis, investigation, methodology, project administration, resources, validation, visualization, and contribution to all parts of the manuscript.

In our initial step of exploring proactive VAs, we aimed to examine people's perspectives on various everyday domestic scenarios wherein a VA takes a proactive approach by addressing users based on their ongoing activities and conversations. Our goal was to identify which types of proactive interactions are perceived as most useful, pleasant, appropriate, and overall positive. Similar to the approach detailed in Publication 7, we adopted a scenario-based design method (Carroll, 1999) where participants engaged with hypothetical situations presented through eight carefully crafted scenarios involving proactive smart speakers in a domestic setting. These scenarios were illustrated in a comic-style format, featuring two or three panels each (see Figure 6.1). To mitigate gender bias, the fictional proactive agent in these scenarios was assigned the gender-ambiguous name 'Jay.' For further details about the scenario design process, please refer to the publication (Publication 8).

We conducted an online survey involving 47 participants. The participants were presented with the eight scenarios in a randomized order and were tasked with evaluating the agent's proactive interactions, providing ratings for usefulness, appropriateness, pleasantness, and their overall impression. In the following, we will generally refer to these variables as the *'rating dimensions'*. Additionally, open-ended questions invited participants to share specific aspects they liked or disliked concerning the agent's proactive behavior.

Findings

Our investigation revealed positive sentiments among users regarding proactive agent interactions, emphasizing their perceived usefulness. However, it also brought to light concerns about privacy, the timing of interventions, and appropriateness in specific contexts, echoing similar findings in prior studies (Lau et al., 2018; Amershi et al., 2019; Cha et al., 2020). Participants responded more positively to proactive interactions that

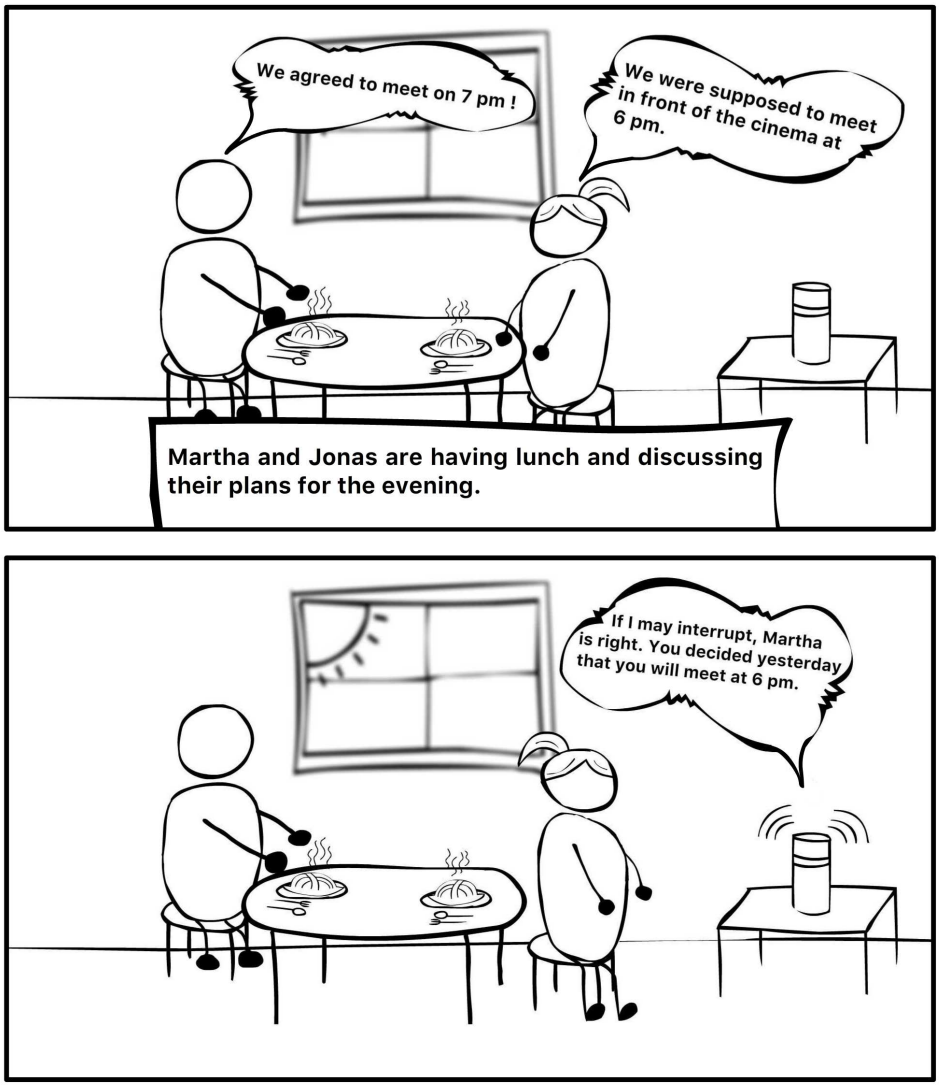


Figure 6.1: One of the storyboards used in the online survey presenting a scenario in which the voice assistant is proactively engaging in a conversation between two people to resolve their disagreement.

were more anticipated. For example, our quantitative and qualitative analyses revealed that reminders were the most favored type of intervention. This preference may stem from participants' familiarity with various types of reminders from existing devices and services they use regularly. Furthermore, proactive instructions related to ongoing tasks or health-related suggestions were perceived as particularly helpful.

A notable observation was the strong correlation among rating dimensions. Interestingly, the overall impression of scenarios demonstrated a stronger correlation with aspects such as *appropriateness* and *pleasantness* compared to *usefulness*. This suggests that designing for social or situational appropriateness should be a primary consideration for proactive VAs. For instance, we observed that scenarios involving a single user were

generally rated more positively than those featuring multiple individuals. Moreover, interfering in personal conversations and presenting evidence from past interactions was deemed inappropriate. Many critiques centered around the social awareness of VAs, questioning their understanding of context and intentions. Participants highlighted the complexity of social skills, such as knowing when to speak or approach others, suggesting that these might be challenging for computer systems to master. Recommendations included a more courteous approach from the agent, such as asking, "May I Interrupt?" before speaking.

As anticipated, several participants voiced privacy concerns, highlighting the inherent challenge of proactive VAs continuously listening and observing users and their surroundings. This concern, previously noted with existing smart speakers (Lau et al., 2018; Tabassum et al., 2019b), becomes more pronounced in proactive scenarios due to heightened data collection requirements for determining opportune moments for VA interactions.

In summary, our findings underscored a generally positive reception of proactive interactions by speech agents, however, with reservations about intervention timing, privacy protection, and user control. The study suggested that a favorable perception of proactive VAs is more closely tied to appropriateness than perceived usefulness. The diverging opinions of the participants emphasize that proactive VAs may be desirable only in particular situations and for specific users.

6.1.2 | Proactivity Dilemma

This section is based on **Publication 9**:

Nima Zargham, Leon Reicherts, Michael Bonfert, Sarah Theres Völkel, Yvonne Rogers, Johannes Schöning, and Rainer Malaka. **Understanding Circumstances for Desirable Proactive Behaviour of Voice Assistants: The Proactivity Dilemma**. In Proceedings of the 4th Conference on Conversational User Interfaces (CUI '22), Association for Computing Machinery.

My contribution to this work: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, validation, visualization, and contribution to all parts of the manuscript.

After our initial exploration, we recognized the need for a more detailed understanding of the contextual factors influencing the appropriateness of proactive behaviors. Hence, we sought to delve deeper into this aspect through a qualitative approach, seeking a comprehensive understanding of the underlying reasons for variations in people's ratings. Once again, we adopted a scenario-based approach to examine users' opinions. This

subsequent study used nine scenarios featuring a proactive VA within domestic settings. Seven scenarios were adapted from our earlier investigation (Publication 8), and two new scenarios were introduced to broaden the spectrum and diversity of situations considered (see Figure 6.2). Employing a qualitative methodology, we delved into these scenarios to determine why specific proactive behaviors are deemed more or less desirable in particular contexts. To facilitate this exploration, we developed an interactive task-based interview procedure. This method involved participants reflecting on the scenarios from diverse perspectives while engaging in various tasks on a virtual whiteboard. Using this approach, we aimed to uncover in-depth considerations surrounding proactive features and collect more nuanced and insightful data.

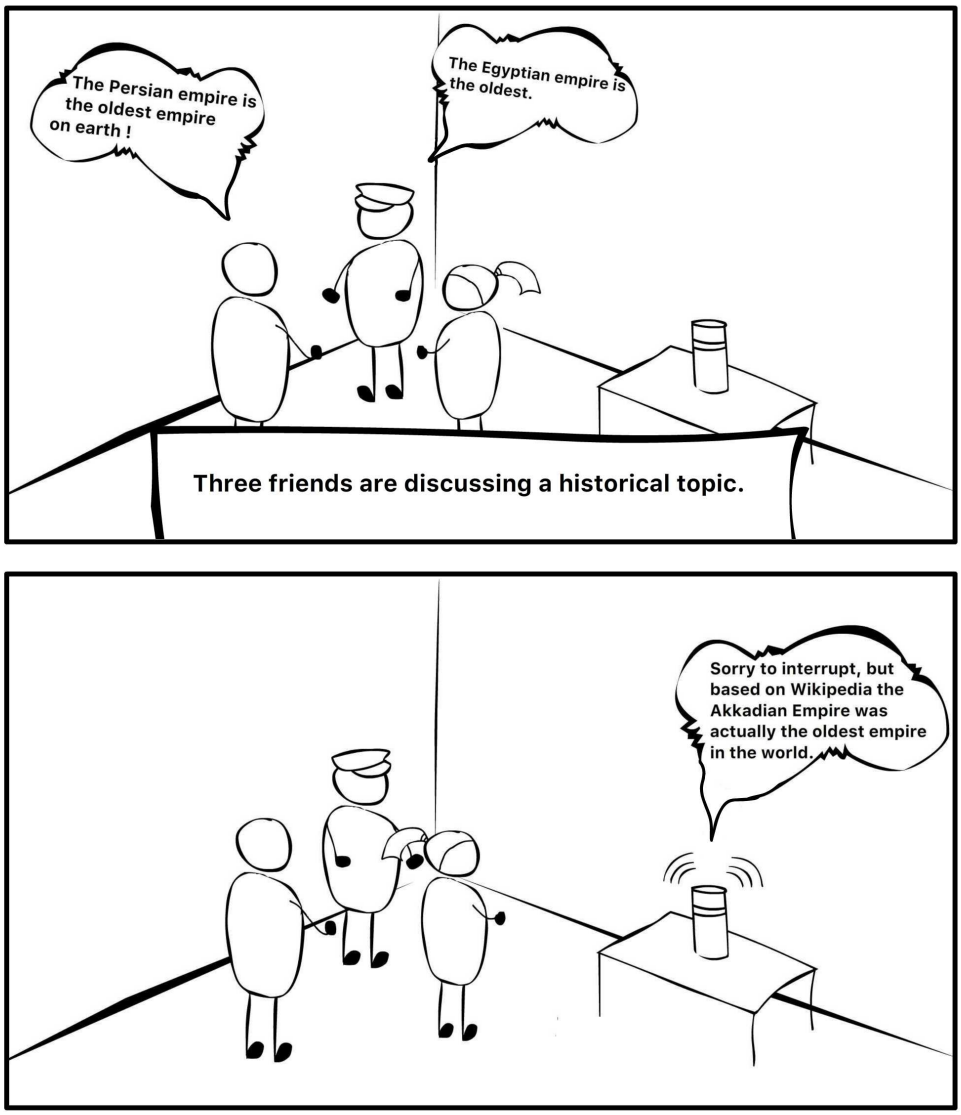


Figure 6.2: An example storyboard used in the study. In this scenario, the agent proactively approaches users based on their conversation.

For this study, we aimed to address the following research questions:

RQ1: Under which circumstances is proactive behavior by a voice assistant perceived as desirable?

RQ2: How should proactive interventions be initiated by the voice assistant?

The study tasks were conducted using a virtual whiteboard tool. In the initial brief interview, we captured participants' initial impressions of the individual scenarios. Subsequently, participants engaged in a card-sorting task wherein they ranked scenarios based on the perceived usefulness, appropriateness, and invasiveness of the agent's interaction. Following this, participants speculated on potential scenario evolutions and character responses to the agent's interventions. In the third task, participants were prompted to identify the most invasive and inappropriate scenarios, proposing improved interventions. The final task involved participants deciding how they preferred the VA to initiate each scenario's interaction and whether a cue should precede speech commencement. After completing all tasks, the session concluded with a short semi-structured interview where participants provided their overall impressions and shared insights into the potentials and challenges associated with proactive smart speakers.

Findings

Participants held diverse opinions regarding the proactive behaviors of a VA. While some participants were more positive towards proactive interventions, appreciating the added features, others held a more pessimistic perspective. Most participants had mixed feelings, acknowledging that proactive interactions can be both beneficial and intrusive, characterizing such interactions as "a double-edged sword." We identified instances where users perceived proactivity as both useful and appropriate. However, a recurring pattern emerged, revealing the *proactivity dilemma*. This dilemma encapsulates the challenge of proactive interventions being seen as helpful but simultaneously overly intrusive. Our findings emphasized the importance of urgency in determining the appropriateness of proactive interventions, aligning with prior research (Cha et al., 2020; Nothdurft et al., 2015; Edwards et al., 2021). We observed that proactive interventions addressing potential health risks were viewed as highly useful, with users prioritizing urgency over privacy concerns, as observed in previous research (Tabassum et al., 2019a). Generally, the more serious and urgent the topic, the more users found proactive assistance valuable. Similar to our previous study (Publication 8), proactively reminding users about upcoming activities or events received positive acknowledgment as an appropriate and useful intervention.

Once again, it became evident that participants' primary hurdle for adopting proactive VAs was the privacy factor. Concerns revolved around companies offering such assistants' potential misuse of personal data. Additionally, participants expressed unease about

introducing an active entity into their homes, shifting from the current passive role of conventional VAs. The fear of paternalism and a perceived lack of control over the device raised concerns, particularly regarding potential negative social impacts in multi-user settings. In multi-user scenarios, interventions that helped users resolve issues and save time were positively received. However, participants emphasized that other than emergency situations, such interventions were considered appropriate only if people could first get a chance to resolve the matter on their own. Participants generally found it annoying and interfering when the agent responded to questions aimed at others in multi-user scenarios. However, if the intended recipient could not provide a satisfactory response, the agent's intervention was considered useful and appropriate.

Participants expressed concern about the agent potentially interfering with interpersonal interactions and bonding opportunities, deeming interventions disruptive in social situations. Understanding the co-located individuals' relationships, conversation seriousness, and intimacy emerged as crucial factors for intervention appropriateness, aligning with prior research (Wei et al., 2021; Miksik et al., 2020). Additionally, participants had mixed feelings about the agent's role in correcting individuals, with some finding it inappropriate, annoying, or insulting. Despite this, some found such corrections useful, highlighting significant individual differences in perceptions. Concerns over potential loss of agency were evident, with participants expressing discomfort at feeling controlled or patronized, especially when the agent suggested healthier behavior like avoiding prolonged binge-watching. Our observations indicated that phrasing and interaction predictability, facilitated through user pre-configuration, could increase the appropriateness of interventions. Participants recommended that the agent's phrasing be polite, calming, and suggestive rather than imposing. Correspondingly, they emphasized the importance of having control over proactive interventions, desiring the ability to configure times and topics for anticipating interactions. Ideally, proactive VAs should be highly customizable and personalized based on individual user needs and preferences.

In most situations, participants expected the agent to request permission before initiating a conversation. This request could be communicated verbally, using high-fidelity cues like addressing the user by name (e.g., "Excuse me, Alex") or polite phrases (e.g., "May I interrupt?") as suggested in our previous work (Publication 8). Alternatively, non-verbal cues, such as abstract audio or light indicators, could offer a more subtle approach. Participant preferences varied based on ongoing activities, with cues designed not to distract unless urgency required a noticeable alert. Verbal cues were perceived as the most distracting, followed by audible cues, while subtle visual cues were considered the least intrusive. Our findings lead to a proposed initiation process model for VAs in non-urgent situations, depicted in Figure 6.3. It begins with an *initial cue*, where the agent signals its intent to speak. Upon user approval, the agent introduces the *topic of intervention*. If the user also approves, the agent can proceed with the intended action. In urgent cases, the second step may be skipped or combined with the first, depending on sensitivity and context, especially in single-user settings.

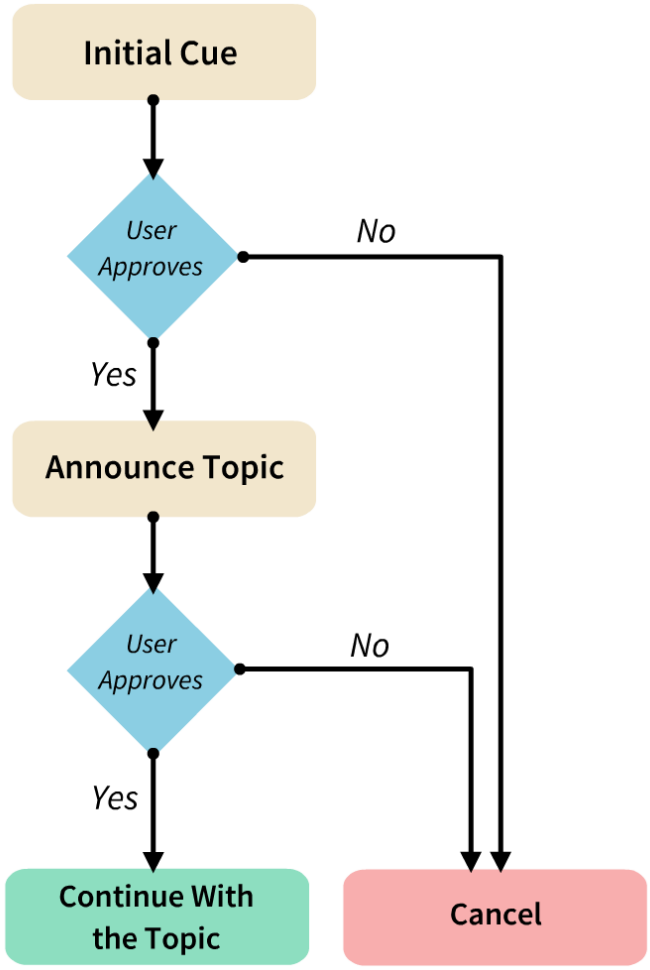


Figure 6.3: The figure shows the proposed initiation process model to proactively interact with people.

In summary, our findings highlight the great potential of proactive interactions, especially for important reminders, time-saving interventions, or emergency support. However, concerns surrounding privacy implications, potential loss of user agency, and interference with social activities may hinder the widespread adoption of such systems. Based on our interpretation, the desirability of proactive interactions hinges on several key factors:

- **Significance:** The urgency or critical nature of a topic could determine the appropriateness of proactive VA interventions. Desirability is heightened in scenarios with a broad scope or severe consequences.
- **Context Awareness:** Proactive VAs must accurately determine the interpersonal and environmental context, considering factors such as the presence of other users, relationships, ongoing activities, and time of day.
- **Agency and Control:** Users should be able to customize proactive features,

specifying times and topics for interventions. They should have control over when the agent listens and observes its environment and when it is allowed to intervene to better anticipate such interactions.

- **Individual User Factors:** Recognizing individual differences in how interventions are perceived is crucial. Proactive VAs should consider individual user factors such as physical and cognitive abilities, current physical and emotional state, and personality and preferences.
- **Form of Execution:** Initiating interactions should involve seeking permission through verbal or non-verbal cues, announcing the intervention's topic—unless time-critical—and phrasing the intent politely, goal-oriented, and concisely. Over time, with user familiarity or consent, the VA may streamline the process.

6.1.3 | Humorous Proactive Agents

This section is based on **Publication 10**:

Nima Zargham, Leon Reicherts, Vino Avanesi, Yvonne Rogers, and Rainer Malaka. **Tickling Proactivity: Exploring the Use of Humor in Proactive Voice Assistants**. In Proceedings of the 22th International Conference on Mobile and Ubiquitous Multimedia (MUM '23), Association for Computing Machinery.

My contribution to this work: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, validation, visualization, and contribution to all parts of the manuscript.

The two previous studies revealed the potential and downsides of proactive agent interactions. However, certain interactions, particularly those involving correction or nudges for positive behavior change, were often perceived as invasive or inappropriate by users. However, we found that one of the factors that could increase the appropriateness of proactive interactions was the style in which the system formulates the interactions with the user. Building on our prior work on proactive interactions of speech agents, we aimed to explore the impact of humor on the desirability of proactive interactions. Speech agents commonly fall short of user expectations as communication partners (Sheehan et al., 2020) and are frequently characterized as machine-like, cold, and socially inept (Feine et al., 2019; Go and Sundar, 2019; Shin et al., 2023). Humor is recognized for its effectiveness in reducing stress (Narula et al., 2011), enhancing well-being (Martin et al., 2003, 1993), and making difficult information easier to 'digest' (Schöpf et al., 2017; Reece, 2014; Lomax and Moosavi, 2002; Gandino et al., 2010). Additionally, studies highlight that incorporating

humor in Conversational Agents (CAs) could enhance service satisfaction (Shin et al., 2023) and potentially improve user engagement (Shum et al., 2018).

In this work, we sought to explore the contextual elements that highlight the appropriateness of humor in agent interactions, considering the specific environments and situations where such interventions are deemed desirable. Recognizing the inherent subjectivity of humor perception, influenced by socio-cultural backgrounds (Braslavski et al., 2018; Yue et al., 2016), we acknowledge the limitations associated with individual interpretation. Nonetheless, there is potential in collectively exploring certain social and environmental aspects of humor to better incorporate it in voice assistants, thereby enhancing user experience.

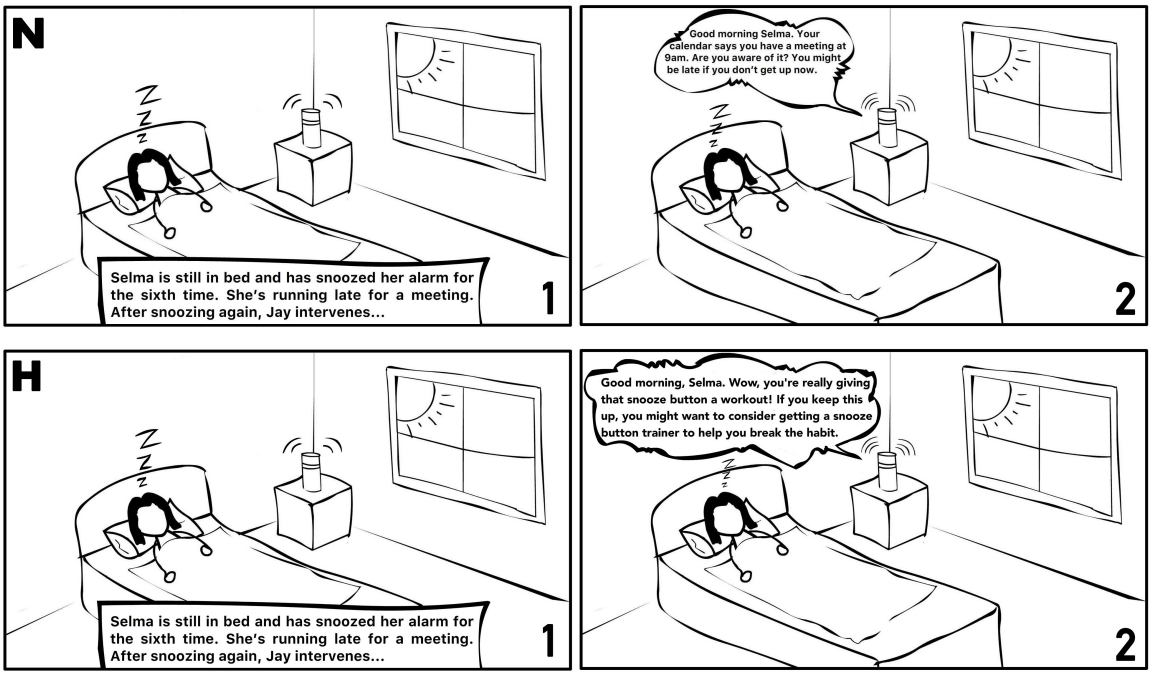


Figure 6.4: Both versions of the scenario *Meeting Reminder*. On the top, the neutral version, and at the bottom, the humorous version is shown. Both versions were evaluated in the survey.

We used scenarios from our prior study (Publication 9) featuring diverse proactive actions by a speech agent in a home setting. Employing a three-step process, we infused humor into the agent's comments. For a detailed explanation of the humorous scenario design process, please refer to the dedicated publication (Publication 10). Given humor's subjective and context-dependent nature, our goal was not universal humorous appeal but to create versions perceived as *more* humorous than the initial/neutral scenarios *on average*. Participants assessed two versions of each scenario—one incorporating humor and the other without (see Figure 6.4). Their task involved rating scenarios for *usefulness*, *appropriateness*, *invasiveness*, and the perceived likelihood of the user *considering* the agent's statements. Participants' responses were collected via the online survey platform

Qualtrics [1]. For this study, we pursued the following research questions:

RQ1: Can the use of humor by a VA increase the desirability of its proactive interventions?

RQ2: In which situations and context can humor be perceived as more appropriate?

Findings

Overall, our findings indicate that the inclusion of humor in our scenarios did not consistently enhance *usefulness*, *appropriateness*, *invasiveness*, or *consideration*. Notably, participant responses varied widely, once again reflecting the inherent subjectivity of humor (Braslavski et al., 2018; Yue et al., 2016; Zargham et al., 2023a). However, we observed that certain factors influenced the desirability of humorous proactive comments by VAs. Half of the participants did not like the humor used in our scenarios, which predominantly resulted in a negative impact on *usefulness*, *appropriateness*, *invasiveness*, and *consideration*. Conversely, another subgroup generally enjoyed the humor, leading to lower *invasiveness* ratings. Taken together, one can assume that when humor fails to resonate with users, it can adversely affect perceptions of the VA's proactive interaction. Conversely, well-received humor can mitigate *invasiveness*, acting as a buffer and increasing receptiveness to proactive interventions.

We found that the effectiveness of delivering a humorous interaction is heightened when users perceive the AI agent as a more socially equal partner. Our evaluation revealed that participants considering the agent to be socially equal tended to rate humorous scenarios as funnier than those perceiving the agent to be inferior. Additionally, participants viewing agents at a similar social level to themselves were more open to humor use by them. These results align with existing literature, underscoring that the perceived characteristics of the individual delivering humor (in this case, the mental model about the AI agent) impact reception, especially regarding social status and perceived authority (Yam et al., 2019; Bitterly, 2022; Romero and Cruthirds, 2006). These findings suggest that speech agents should customize their use of humor based on users' perceptions of their relationship with the agent.

Participants indicated that the agent should avoid using humor in discussions or activities related to serious topics such as health, work, or finances. They stressed the importance of refraining from humor in time-critical and socially tense situations, emphasizing the significance of timing and context in deploying humor. Disapproval arose when humor lacked careful contextualization, leading to perceptions of impoliteness and inappropriateness. Concerns were raised about humor undermining the gravity of sensitive or serious topics, potentially reducing the inclination to address problems due to its association with non-serious contexts (McGraw et al., 2012).

People preferred humor during light-hearted and playful occasions, suggesting its appropriateness when cues like laughter or humorous conversations are detected. They

[1] https://www.qualtrics.com

also noted the relevance of the relationship context, proposing the use of humor in the presence of close friends or family members. Participants favored humorous VA comments that balance entertainment and utility, fostering a motivating and encouraging atmosphere. We observed the potential for using humor to alleviate tension and enhance user experience, especially when combined with factual information and contextual relevance.

In summary, our findings underscore the complexity of integrating humor into proactive voice assistant interactions. It became evident that humor is more than just a supplementary aspect or interactional feature that can be casually incorporated. If not well-received, humor can lead to counterproductive outcomes. However, it could elevate the interaction if it resonates with the user. Additionally, the success of humorous interactions can be enhanced when people perceive the agent as more socially equal. We recommend tailoring humor to individual preferences and sensitivities, recognizing the diverse reactions it can elicit. Designers could view humor as a potential strategy to soften proactive interaction. However, if humor cannot be achieved and tailored to individuals, alternative approaches may be more effective for reaching desirable outcomes.

6.2 | Conclusion

This chapter introduced three research papers, all centered around the interaction style of speech agents, with a particular focus on proactive interactions and their desirability among users. In these studies, we adopted an approach inspired by scenario-based design methods (Carroll, 1999) and vignette experiments (Aguinis and Bradley, 2014). Given the absence of proactive speech agents with comparable capabilities in the current market, we delved into people's opinions on these features through hypothetical scenarios. This method requires participants to engage in speculation, allowing the exploration of interactions with future technologies that would be complicated or expensive to develop. Additionally, it facilitates the evaluation of aspects of the system that are challenging to simulate realistically, such as emergency situations or delicate private settings. In our studies, this method proved effective in capturing people's perspectives and preferences regarding such agents, providing valuable insights into user expectations and interactions.

Our studies revealed the significant potential of proactive interactions as an augmentation to the current interaction style of speech agents. The potential benefits were evident, particularly in urgent situations, time-saving tasks, and reminders. However, concerns regarding privacy implications, potential loss of user agency, and interference with social activities were raised, underscoring the need for careful consideration and mitigation of these issues. Incorporating unique personalities and human-like characteristics, such as humor, demonstrated promise in minimizing invasiveness and enriching interactions. Yet, the user's perception of the agent's social status emerged as a critical factor influencing the success of this approach. We highlight the importance

of carefully crafting and personalizing these characteristics to align with individual preferences, acknowledging their potential to enhance the overall user interaction experience.

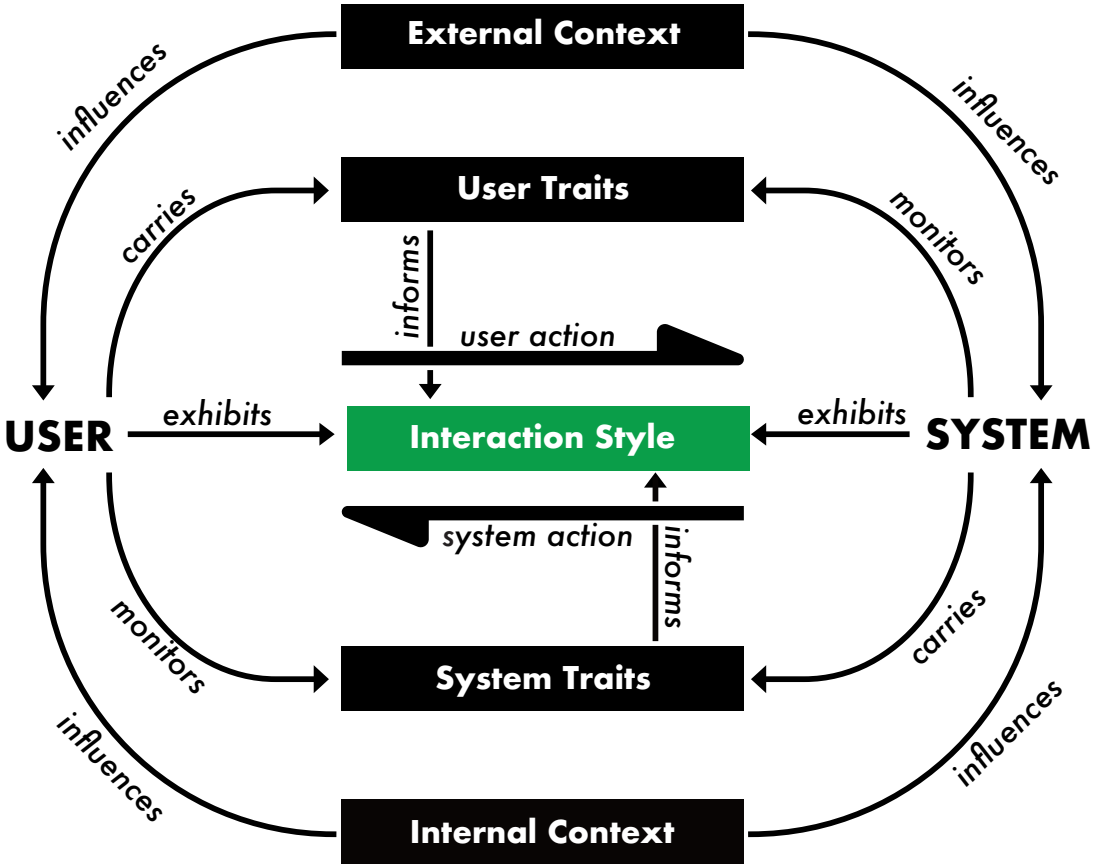


Figure 6.5: The role of interaction style in the HASI model.

Reflection on the HASI Model

In examining the proposed HASI model, we observed several relations between the elements, particularly concerning the interaction style (see Figure 6.5). Notably, the external context emerged as a crucial determinant shaping the interaction. Environmental variables such as the time of day and ambient noise levels could influence the interaction dynamics. The adaptability of the speech agent to these contextual elements was vital in enhancing user satisfaction and effectiveness. Furthermore, interpersonal and socio-cultural contexts played a significant role in the interaction process. Factors such as the user's presence alone or with others, the intimacy of relationships, ongoing activities, and adherence to cultural and societal norms all impacted the user's interaction experience. For instance, participants found proactive interactions more suitable when alone but

deemed them highly inappropriate when the agent interrupted social interactions. Moreover, once again, we witnessed individual user traits emerge as essential factors in shaping interaction perceptions. Physical and cognitive abilities, current emotional and physical states, and unique personality traits and preferences collectively influenced whether an interaction was deemed appropriate. This highlights the importance of tailoring speech interactions to individual user profiles for optimal outcomes. Users' insights about the system, including their familiarity with its functionality and their mental model of AI agents, influenced interaction dynamics. Understanding and accommodating these elements could be highly important for designing speech interactions that resonate with users' expectations. Additionally, the criticality and urgency associated with the subject matter significantly influenced the user's perception, with more crucial and time-sensitive topics aligning with heightened appropriateness. All in all, the studies in this section imply that tailoring the system interaction style to individual user traits while acknowledging broader contextual elements is essential. Factors such as appropriate feedback type, phrasing, and interaction timing could be adjusted based on these factors to enhance user experience effectively.

Responding to the dissertation's **TRQ4**, our findings highlight that several factors regarding the speech agent's interaction style influence its suitability for users. We identified several factors that can enhance the interactions' appropriateness. We witnessed that appropriate phrasing of the interactions could enhance its suitability. Participants expressed a preference for interactions characterized by conciseness and politeness. One crucial dimension influencing successful interaction was the predictability of the interaction. Anticipated interactions were more likely to be perceived as appropriate, while those catching users by surprise risked being deemed intrusive. This underscores the importance of designing interactions that align with users' expectations and provide a certain level of predictability. Repeatedly, our findings underscored the critical importance of user agency and control. Interactions that threatened users' agency and control over the system received significant criticism. Empowering users with heightened control over features and interactions was consistently identified as a key priority in fostering a positive user experience. We also noted that users' mental models of AI agents and perceptions of such systems are highly influential in the interaction process. Users' conceptualizations of these systems regarding their social role (e.g., companions, butlers, or advisors) shape their expectations about interactions with these systems. Understanding users' mental models could be highly beneficial for designing AI agents that align with user expectations and foster trust and satisfaction in the interaction experience. Once again, our exploration underscored the inadequacy of a uniform approach to speech system interactions. To optimize the suitability of interaction for users, designers and developers should advocate adaptability, fostering more personalized and customized systems. This approach empowers users to tailor interactions to their preferences, contributing to a more successful and user-centric experience.

Similar to the previous chapters, it is crucial to acknowledge that the studies discussed

in this chapter represent only a subset of the vast design possibilities concerning the style of interaction for speech agents. The landscape remains broad, underlining the need for future research to further delve into this dimension of speech interaction to improve our understanding of this human-agent speech interaction.

7

Discussion & Limitations

Speech interaction has emerged as a dominant research domain within HCI. Early attempts at leveraging speech interaction were restrained by several factors, contributing to their limited success. Primarily, technological limitations and constrained computational power were significant obstacles. These early systems often lacked the processing capabilities necessary to interpret and respond to speech inputs effectively. Moreover, the absence of access to extensive datasets for training models restricted their ability to adapt and improve over time. Additionally, the complexity of language itself presented a big challenge. Unlike simpler input forms, such as button presses or mouse clicks, language encompasses nuances, dialects, and contextual variations that are difficult to capture and process accurately. Advancements in related fields, such as linguistics and user experience research, were necessary to provide a deeper understanding of language processing and user interaction with speech systems. This broader interdisciplinary approach was crucial for addressing the multifaceted nature of speech interaction beyond purely technological considerations.

Recent years have witnessed significant advancements in this domain, resulting in more sophisticated implementations of speech-based systems. The evolution of speech technology has expanded these systems' capabilities and opened up new avenues for applications across various domains, including healthcare (Latif et al., 2021), education (Terzopoulos and Satratzemi, 2020), and entertainment (Allison, 2020). Consequently, understanding the complexities of the interaction process has become an important topic for HCI researchers and practitioners seeking to create more successful and user-friendly interaction experiences.

This dissertation aimed to enhance our understanding of speech interaction within domestic activities by offering a comprehensive outlook on designing speech systems that can achieve a better user experience. The dissertation posed the following overarching research question:

- **TRQ1:** What factors contribute to successful partner-based speech interaction in domestic activities?

This chapter addresses **TRQ1** by analyzing the findings and insights from the individual research works detailed in previous chapters. Additionally, it discusses the limitations of the dissertation and proposes avenues for future research.

7.1 | Utility Dimension

To ensure successful speech interaction, a solid technical foundation is key. Across several of our investigations, we noted users' primary emphasis on pragmatic considerations and task fulfillment over other factors. Even when discussing hedonic aspects of speech systems, system reliability was raised as a crucial factor that users wish for speech systems. In Publication 7, when evaluating the desired personality traits for an agent, participants consistently gave high ratings for agreeableness, conscientiousness, and emotional stability, all of which contribute to the perception of reliability. Additionally, in Publication 6, we observed that people place a higher priority on the agent's voice and personality over its visual appearance. These underline the importance of optimizing speech systems for maximum efficacy and broad functionality, echoing previous literature highlighting performance challenges in speech recognition and constrained functionality as primary concerns of users (Shrivastava et al., 2023; Wan, 2021; Jentsch et al., 2019). In this dissertation, Publication 1 and Publication 2 were specifically tailored to address this critical dimension of speech systems, aiming to enhance their technical foundation to better align with users' pragmatic needs and expectations. Publication 2 highlighted the importance of prioritizing the intended command by the user over optimal outcomes when evaluating the efficacy of speech systems. The ability to accurately interpret user commands, even if they do not result in the desired response, is crucial for mitigating user frustration and preserving their sense of agency. Previous research on conversational agents has been investigating the prediction of user intents and determining suitable repair strategies in case of conversation breakdowns (Kvale et al., 2019; Ashktorab et al., 2019; Shevat, 2017). Publication 1 emphasized the significance of leveraging contextual data to support speech recognition efficacy. Through our proposed context-aware speech recognition method, we aimed to integrate contextual cues into the speech recognition process to improve accuracy. Initial findings indicate promising results, suggesting that such an approach holds potential for significantly improving speech system performance. This supports previous findings advocating for the use of supplementary data sources, such as audio-visual fusion strategies, to enhance speech recognition performance (Sterpu et al., 2018). Looking ahead, further advancements in deep learning techniques and user modeling have the potential to refine and augment the methods proposed in our research. With this, speech systems can better predict and adapt to individual speech patterns and preferences. Research by Pfau et al. has shown the potential of employing deep learning techniques to create precise models of user behavior (Pfau et al., 2018). Such personalized approach could enhance speech recognition accuracy

and empower users with greater control and agency over their interactions with these systems. The concept of context-aware speech recognition draws inspiration from the dynamics of human-human communication, where contextual factors significantly influence communication effectiveness. Just as in interpersonal interactions, the context in which communication occurs can facilitate or hinder effective communication between individuals. This underscores the importance of considering contextual cues in designing speech recognition systems, as they can play a crucial role in shaping the interaction experience. This insight highlights that valuable inspiration for designing such systems can be derived from our observations and understanding of interpersonal communication dynamics. HCI research is often motivated with such mappings to design more intuitive and user-centered systems.

7.2 | Representation Dimension

The representation of speech agents is a topic that has gained more research interest in recent years. Through our studies, we have highlighted the significance of agent representation in HASI. This dissertation explored this dimension of speech systems through several publications (Publication 3, 4, 5, 6, and 7), aiming to examine various forms of agent representation and user preferences in this regard.

Human communication transcends mere verbal exchanges, encompassing non-verbal cues like facial expressions and bodily gestures that enrich the conversation with contextual information and emotional nuances. These additional layers of communication facilitate a deeper understanding of the message's intent and tone, enhancing the overall communication experience. In our research, we witnessed that integrating these additional elements into speech agents could make interactions more engaging and intuitive for users, consistent with prior studies suggesting that embodied agents can further convey information and emotions, fostering stronger user connections (Wang et al., 2019; Kim et al., 2018). Incorporating visual components in communication holds particular significance for individuals with hearing impairments, as it enhances accessibility by providing alternative channels for information exchange. By embracing embodiment and visual cues in agent representation, speech systems can cater to a wider range of users, fostering inclusivity. The significance of system representation varies depending on the task at hand. For time-sensitive tasks, prioritizing the utility aspect of the speech system is crucial. However, in more casual and social interactions, incorporating additional non-verbal elements can enrich the user experience.

Our studies highlighted that the representational dimension of speech systems can be inherently subjective, with user preferences demonstrating significant variation. Through our studies, we observed a diverse range of opinions regarding system presentation. For example, in the case of multi-agent systems, some participants favored the concept, considering it more professional and engaging, while others deemed it unnecessary and

preferred a single-agent system. Similarly, when it came to visualization of the virtual speech agent, preferences regarding the agent's appearance varied widely. These findings underscore the challenge of adopting a universal design approach to accommodate diverse user preferences.

We observed that the similarity-attraction principle (Bernier and Scassellati, 2010) extends to people's preferences regarding the visual representation of speech agents. We noticed that individuals tend to favor speech agents that bear resemblance to themselves in terms of demographics, fostering a sense of relatability. This highlights the impact of users' cultural backgrounds on their preferences in this domain. We also noted similarities in users' preferences for the agent's personality, specifically in agreeableness, where a strong positive correlation was evident between users' personality traits and their desired attributes for the agent. Such findings contribute to the growing understanding that people view these systems not merely as tools or utilities but rather as communication partners. In human-human interactions, individuals often seek partners who share common interests or traits (Launay and Dunbar, 2015; Reagans, 2005; Sprecher and Regan, 2002). Similar patterns have been witnessed with social robots (Woods et al., 2005; Craenen et al., 2018). Hence, a similar tendency can be assumed for AI agents and their visual representation.

Additionally, we noted a tendency towards attractiveness in these agents, consistent with prior research on agent visualization (Yuksel et al., 2017; Khan and De Angeli, 2009; Banakou et al., 2009), suggesting that agents should possess a minimum level of visual appeal. This preference towards attractive agents could be rooted in the association of physical appeal with positive qualities such as popularity, competence, and desirability (Lorenzo et al., 2010). Consequently, having an attractive virtual agent may give users a sense of elevated social status, fostering feelings of prestige and satisfaction. Research has shown that individuals associated with an attractive counterpart are often evaluated more favorably by others (Sigall and Landy, 1973). Society places significant emphasis on physical attractiveness and its perceived benefits. Societal norms and expectations influence people's preferences for attractive virtual agents. Users may perceive an attractive agent as indicative of superior design, higher quality, or more advanced technology, aligning with perceived societal standards.

Given the inherent subjectivity in preferences with regard to an agent's representation, prioritizing customization and personalization becomes imperative. In Publication 7, we uncovered several aspects that users expressed a desire to modify within these systems through customization and personalization features. Designers and developers should recognize the breadth of preferences in this domain and empower users with the ability to tailor various agent characteristics to enhance user satisfaction and engagement with such systems. This approach acknowledges users' diverse needs and preferences, accommodating different cultural backgrounds, demographics, and tastes.

7.3 | Interaction Dimension

In human-human communication, several communication strategies come into play, including when to initiate conversations, timely responses, sentence formulation, maintaining engagement, and appropriately concluding interactions. These strategies foster effective and meaningful exchanges between people, honed over years of social experience (Meyer et al., 2016). When computers act as communication partners, we naturally expect them to adhere to similar communication norms and behaviors, as argued in the "computers are social actors" paradigm (Reeves and Nass, 1996). In this dissertation, we explored the interaction strategies of speech agents through Publications 8, 9, and 10. Our primary focus centered on proactive interactions, a highly anticipated feature for speech systems (Miksik et al., 2020).

Our studies showed that anticipation plays a crucial role in determining the effectiveness and suitability of interactions for users. We witnessed that interactions that catch users by surprise risk being perceived as intrusive or unwelcome. Anticipated interactions could improve the user experience by allowing individuals to mentally prepare for interactions, thereby acknowledging users' autonomy. Speech agents should provide cues or notifications before initiating the conversation in order to enhance the appropriateness of interactions, as outlined in Publications 9. However, akin to human interactions, the significance of the topic of interaction also influences the dynamics. For instance, we observed that additional cues may be unnecessary in urgent situations, supporting the findings from previous research (Edwards et al., 2021; Nothdurft et al., 2015). This underscores the contextual sensitivity required in designing effective speech interaction systems. Various contextual factors emerged as influential in shaping interactions. These contained physical environmental factors, interpersonal dynamics, and socio-cultural influences, all identified as crucial to the overall effectiveness of interactions.

These arguments collectively emphasize the necessity for speech agents to be highly context-aware in order to achieve successful interaction. The importance of context awareness stems from the dynamic and multifaceted nature of human interaction. By being context-aware, speech agents can better adapt to varying situational factors, enabling them to tailor their responses and behaviors accordingly. However, this also requires these systems to collect more user data, which can raise privacy concerns. Users may be understandably apprehensive about how speech systems access, store, and potentially share their personal information. Several studies have highlighted that people often refrain from adopting voice assistants as they have privacy concerns or distrust the companies offering smart speakers (Malkin et al., 2019; Lau et al., 2018; Cha et al., 2020). This was also a topic that was brought up repeatedly in our studies. Privacy-preserving features are essential when designing speech systems. Designers and developers should consider designing robust privacy measures, such as local data handling, transparent data policies, and anonymization techniques, to address these concerns and ensure users' privacy rights are respected and protected.

Another notable factor we observed in our studies was the influence of individuals' mental models of AI agents on the interaction process. Users' mindset of how AI agents function and behave could impact their expectations, perceptions, and behaviors during interactions with these systems. This is in line with previous literature suggesting that users' mental models of AI agents play a crucial role in their interactions and decisions regarding system use (Schrills and Franke, 2020). Users who perceive AI agents as intelligent and capable may engage with them more confidently and expect them to perform complex tasks effectively. Conversely, users with more skeptical mindsets may approach interactions cautiously. For instance, users concerned about the privacy implications of AI agents and the security of their personal data may be hesitant to engage in interactions with the system, especially those requiring extensive data input. In Publications 10, we observed that the effectiveness of delivering a humorous interaction could be enhanced when the user perceives the VA as a more socially equal partner. People who viewed the VA as socially equal tended to rate humorous scenarios as funnier than those who perceived the VA as inferior. These findings align with existing literature, highlighting that the perceived characteristics of the individual delivering humor influence its reception (Yam et al., 2019; Bitterly, 2022), particularly regarding the social status and perceived authority of the individual delivering humor (Romero and Cruthirds, 2006). Not being able to match users' expectations could lead to frustration. For instance, consider a user who believes speech agents are highly proficient in understanding and executing commands. This user may feel frustrated and disappointed if the agent fails to accurately interpret their requests or perform tasks as expected. Hence, understanding and accommodating users' diverse mental models is essential for designing AI agents that effectively meet users' needs, preferences, and expectations.

It is important to consider that the interaction dynamics in HASI can evolve over time. Early interactions may differ significantly from interactions after some time. Both the user and the system could learn from past experiences and adjust their interaction strategies accordingly. Users tend to adjust their interaction style, such as how they formulate commands (Stent et al., 2008). Similarly, the system could train based on user feedback and usage patterns to better adapt to the users. However, changes in the preferences and needs of the user can also influence interaction dynamics over time. A person's preferences or priorities may shift, leading to different communication patterns with the system. Therefore, designing speech systems that can adapt and evolve with users over time could provide personalized and more successful interactions.

7.4 | Reflecting on the HASI Model

Throughout our evaluation, we observed the dynamic interplay between different components in HASI. Contextual elements had a significant influence on both users and systems alike.

Regarding the external context, physical environmental factors proved crucial to consider. Environmental noise, for example, could affect the system's speech recognition capabilities and the user's attentiveness. Conversely, leveraging contextual factors, such as providing additional information about ongoing activities, could enhance clarity about user inquiries. Additionally, socio-cultural factors played a notable role in shaping interactions. Societal standards impacted user expectations and preferences in agent representation (system traits), an example being system attractiveness. Furthermore, linguistic diversity and cultural norms surrounding politeness and formality influenced interaction experiences. A speech system designed for a global market might need to adjust its language and interaction style to fit various cultural contexts. Interpersonal context emerged as another influential element. Factors like the people present, the closeness of their relationships, and ongoing activity influenced interaction dynamics. A user might interact differently with a speech agent when alone rather than when hosting guests, perhaps using more formal commands in the latter scenario. We repeatedly witnessed that the lack of context awareness in system interactions could risk its suitability and appropriateness, highlighting the importance of this element in the overall interaction.

Regarding the internal context, we observed that the goal and topic of the interaction play pivotal roles in shaping the desired interaction style. For example, urgent inquiries may necessitate concise responses, while social or playful inquiries may allow for more intimate and verbose exchanges, possibly incorporating humor. Moreover, the history of interactions could significantly influence how subsequent interactions are shaped. Users adapt their interaction patterns based on past experiences to optimize system performance. The system could also leverage the interaction history to better adapt to individuals' interaction styles and preferences.

Regarding the traits layer, we witnessed how the users' individual traits can be an influential factor in HASI. We saw that users' insights about the system, including familiarity with the system and their mental models about AI agents, influenced interaction outcomes. Deviations from user expectations could lead to dissatisfaction and concerns about user agency. Our findings further emphasized the influence of users' personalities, as well as their ethnicity and cultural background, on interaction perceptions. Physical and cognitive abilities, as well as emotional and physical states, also shaped interaction appropriateness. The appropriateness of the interaction styles varied based on cultural backgrounds. For instance, a particular form of sentence formulation might seem appropriate in one culture but impolite in another. Concerning the system traits, as discussed in the earlier section of this chapter, the utility aspect of the system must remain robust. The representation and interaction style of the systems should be tailored to align with individual users and their traits, all the while taking into account societal norms and standards.

Overall, the proposed HASI model appears to effectively address the influential factors within the interaction process when considering insights from individual studies in this dissertation. Figure 7.1 visualizes the key influencing factors of HASI pointed out through

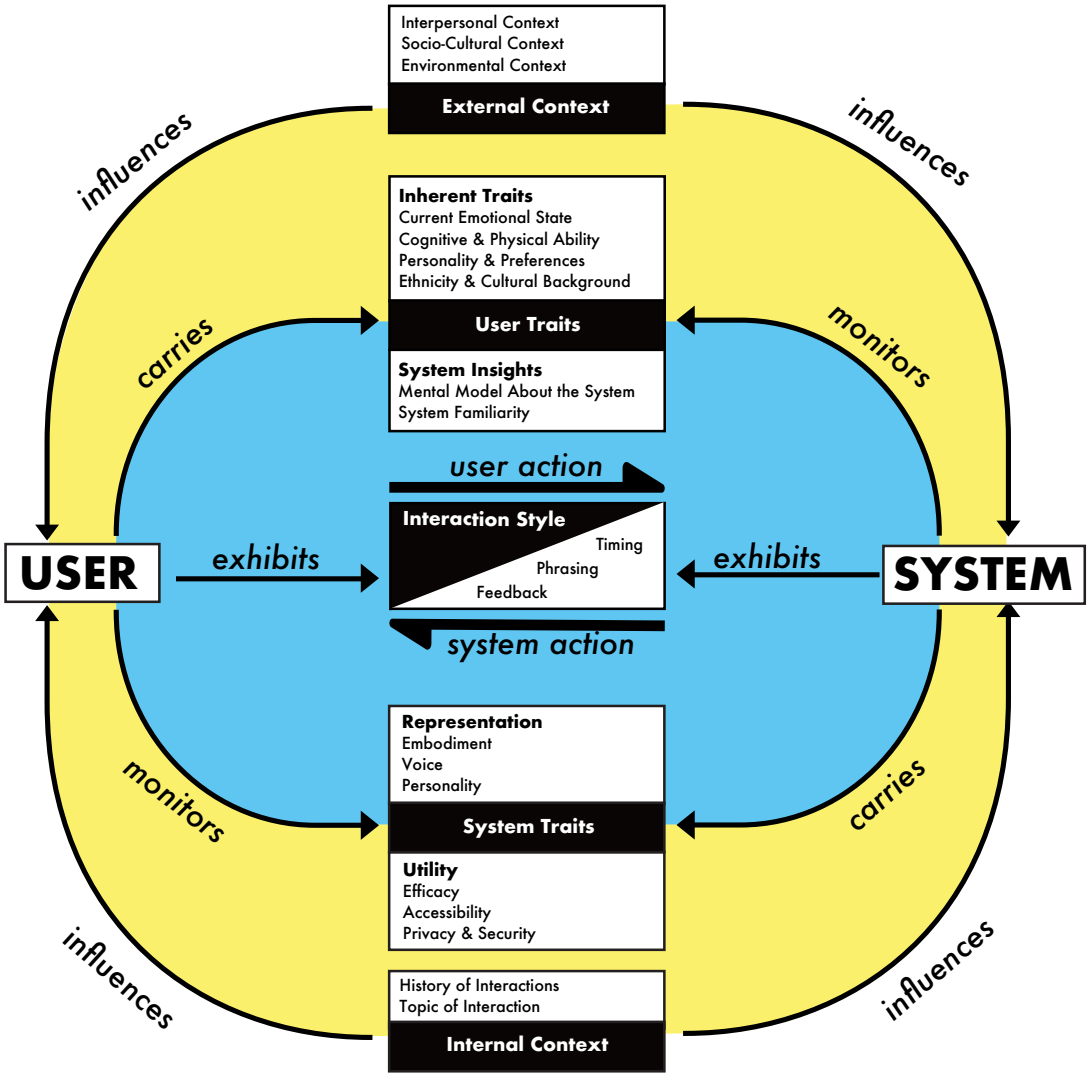


Figure 7.1: The key influencing factors in human-agent speech interaction.

the research undertaken in this dissertation. Just like many other interaction models, the HASI model may have certain limitations. Future studies should continue to examine and validate its applicability and relevance. Additionally, while this model has been primarily examined in the context of domestic activities, it is crucial for future research to explore its applicability across other domains and use cases.

7.5 | Reflecting on Partner-Based Interactions

Earlier in this dissertation, we discussed and argued for the term "partner-based" speech interaction, as discussed in recent literature (Peña et al., 2023; Doyle et al., 2023). Our research findings consistently validated this notion, particularly in domestic settings, where users often perceive speech agents not merely as tools but as distinct entities

and communication partners. In our studies, people often wished for the system to behave more human-like. This perspective reflects a shift from conventional views of technology as passive tools to more dynamic and interactive companions in daily life. The human likeness of speech agents remains a topic of ongoing debate among researchers and practitioners. Despite previous literature consistently highlighting its importance in speech interaction systems (Dubiel et al., 2018; Cowan et al., 2017a, 2015), some researchers argue that the specific dimensions of human likeness people use for defining these interactions are not fully understood (Doyle et al., 2019). In our studies, participants expressed a desire for speech agents to exhibit a sense of autonomy and individuality, suggesting a preference for agents that engage in human-like behavior. This highlights a growing expectation for speech systems to emulate human-like qualities and establish more meaningful and interactive relationships with users. Understanding and incorporating these dynamics will be essential for designing better speech interaction experiences as speech technology evolves.

It is essential to acknowledge that biases are inherent in designing these systems since humans create them. Consider the process of developing a speech recognition system. The data used to train the system come from various sources. This data selection could introduce potential biases based on factors such as language variations. Suppose the training data predominantly consists of speech samples from a specific region. The system may perform more accurately for the demographics of that region but less accurately for users from different linguistic backgrounds (Pyae and Scifleet, 2018). Biases also arise from societal norms, stereotypes, or prejudices. These implicit biases can manifest in the design choices made throughout development. Recognizing and addressing biases in these systems is crucial to ensure fairness, equity, and inclusivity. It requires a concerted effort to mitigate biases at every stage of system development, from data collection and model training to deployment and ongoing monitoring.

The future of speech interaction appears promising and dynamic. This modality is changing how we interact with technology and the world around us. As technological advancements accelerate, speech systems are expected to become increasingly intelligent, intuitive, and personalized. Additionally, improvements in voice synthesis and emotional recognition may facilitate more human-like interactions. These enhancements promise new possibilities for more immersive and interactive experiences.

7.6 | Limitations and Future Work

While this dissertation has made strides in addressing the research questions, it is essential to recognize that the field of designing speech systems is vast and complex, offering numerous avenues for exploration. The research questions posed here are broad and multifaceted, making it challenging to provide comprehensive answers within the scope of this work alone. Claiming to address these questions fully would be overly ambitious.

Instead, this dissertation lays the groundwork for further investigation and refinement in understanding speech interaction. Future research endeavors will undoubtedly contribute additional insights, leading to a more comprehensive understanding of the topic.

Our evaluation focused on speech interaction within domestic environments. This focus was motivated by the accessibility of home settings to the broader population and the growing prevalence of speech systems in everyday households. Additionally, the home environment offers a rich context for interaction, encompassing various interpersonal dynamics and a diverse range of tasks. While many of our findings can potentially extend to other domains, it is important to recognize that the applicability and generalizability of our research findings to different settings require further investigation. Therefore, exploring the transferability of our results to diverse contexts remains an important avenue for future research. As part of our evaluation process, we conducted studies in the domain of speech-based video games. This focus offered a controlled experimental environment and allowed for an easier replication of prototypes with precision. However, one limitation of this dissertation lies in the extent to which the findings from these studies can be applied to other contexts, particularly within home environments. Video games afford researchers a fully observable setting, facilitating the calculation and quantification of various elements that may not be as readily measurable in household settings. While the solutions and methodologies proposed here may be adaptable to some extent to other domains and could inspire similar approaches in diverse settings, it is essential to recognize the inherent limitations in generalizing these findings beyond the gaming context.

Furthermore, in our investigation, the speech systems designed were mainly capable of supporting a limited set of commands. This approach was intentional, aiming to streamline implementation and give more attention to the applied methods and the research objectives. This choice facilitated a structured procedure with high comparability across experiments (Porzel and Baudis, 2004). However, future studies should investigate the transferability of the insights outlined in this dissertation and examine their broader applicability across diverse application domains and larger command vocabularies.

8

Conclusion

This dissertation aimed to broaden the landscape of speech interaction within the context of domestic activities. The overarching objectives revolved around innovation, optimization, and further exploration of this interaction modality. A novel speech interaction model addressing limitations inherent in communication and HCI models was introduced to support a more comprehensive understanding of the complex dynamics in human-agent speech interactions. Moreover, several challenges and shortcomings prevalent in existing systems were identified, and attempts were made to address them in order to enhance the overall user experience with such systems. Additionally, novel features and concepts were designed and explored to observe their potential and limitations for enhancing HASI. While technological advancements address technical concerns with speech systems, more attention should be given to the experiential side. Even though pragmatism is often prioritized in interactions, the studies in this thesis underscore the importance of considering hedonic aspects when designing speech systems. Our findings underlined the importance of user-centered design. Customization and personalization are essential for fostering desirable human-agent speech interaction. Implementing personalized user profiles that consider individual preferences and interaction histories can enable speech systems to tailor their interactions to better align with user expectations. Additionally, systems can enhance the interaction dynamic by leveraging contextual data. This makes these systems more adaptable to the complex and fluid context of real-world interactions. Furthermore, utilizing additional layers of data transmission, such as a visual embodiment, could elevate the interaction and make speech systems more approachable and accessible. It is crucial to prioritize user privacy and data security for ethical and responsible system behavior. Doing so enhances users' trust and confidence in speech systems while promoting transparency and accountability in system design and operation. Considering these factors allows us to unlock the full potential of speech systems, delivering seamless, intuitive, and empowering user experiences. The results of this thesis can provide helpful insights for the future research and development of speech systems. The proposed model can serve as a framework for designing and evaluating speech systems, guiding efforts toward achieving successful interactions.

9

Publications

This dissertation is grounded in ten publications. This chapter contains all these papers in their original form, with content and formatting as published. All research has been conducted collaboratively and reported as such. Each paper is introduced with a cover page that includes the list of all authors and an indication of my personal contributions according to the *CRediT* taxonomy.

Contribution Taxonomy *CRediT*

These are the 14 roles typically played by contributors to research outputs according to the *CRediT* taxonomy.[1] The taxonomy has been refined by Consortia Advancing Standards in Research Administration (CASRAI) and National Information Standards Organization (NISO). I refer to these roles when outlining my contributions.

Conceptualization: Ideas; formulation or evolution of overarching research goals and aims.

Data curation: Management activities to annotate (produce metadata), scrub data and maintain research data (including software code, where it is necessary for interpreting the data itself) for initial use and later re-use.

Formal analysis: Application of statistical, mathematical, computational, or other formal techniques to analyze or synthesize study data.

Funding acquisition: Acquisition of the financial support for the project leading to this publication.

Investigation: Conducting a research and investigation process, specifically performing the experiments, or data/evidence collection.

Methodology: Development or design of methodology; creation of models.

Project administration: Management and coordination responsibility for the research activity planning and execution.

Resources: Provision of study materials, reagents, materials, patients, laboratory samples, animals, instrumentation, computing resources, or other analysis tools.

Software: Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms; testing of existing code components.

Supervision: Oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team.

Validation: Verification, whether as a part of the activity or separate, of the overall replication/reproducibility of results/experiments and other research outputs.

Visualization: Preparation, creation and/or presentation of the published work, specifically visualization/data presentation.

Writing – original draft: Preparation, creation and/or presentation of the published work, specifically writing the initial draft (including substantive translation).

Writing – review & editing: Preparation, creation and/or presentation of the published work by those from the original research group, specifically critical review, commentary or revision – including pre- or post-publication stages.

[1] `https://credit.niso.org/contributor-roles-defined`

Publication 1

"I Know What You Mean": Context-Aware Recognition to Enhance Speech-Based Games

Nima Zargham, Mohamed Lamine Fetni, Laura Spillner, Thomas Muender, and Rainer Malaka

In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24). New York, NY, USA, 2024. Association for Computing Machinery.

Personal contribution to this work: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, part of software, supervision, validation, visualization, and contribution to all parts of the manuscript.

ISBN: 979-8-4007-0330-0/24/05 DOI: 10.1145/3613904.3642426

“I Know What You Mean”: Context-Aware Recognition to Enhance Speech-Based Games

Nima Zargham
zargham@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

Mohamed Lamine Fetni
fetni@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

Laura Spillner
laura.spillner@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

Thomas Muender
thom@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

Rainer Malaka
malaka@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

Figure 1: A screenshot of our escape room game called “Escape the Echo”, displaying the game environment on the third level.

ABSTRACT

Recent advances in language processing and speech recognition open up a large opportunity for video game companies to embrace voice interaction as an intuitive feature and appealing game mechanics. However, speech-based systems still remain liable to recognition errors. These add a layer of challenge on top of the game’s existing obstacles, preventing players from reaching their goals and thus often resulting in player frustration. This work investigates a novel method called context-aware speech recognition, where the game environment and actions are used as supplementary information to enhance recognition in a speech-based game. In a between-subject user study ($N = 40$), we compared our proposed method with a standard method in which recognition is based only on the voice input without taking context into account. Our results indicate that our proposed method could improve the player experience and the usability of the speech system.

CHI ’24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642426

CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**;
• **Applied computing** → **Computer games**.

KEYWORDS

Game Design, Speech Recognition, Speech-Based Systems, Voice-Controlled Game, Voice Interaction

ACM Reference Format:
Nima Zargham, Mohamed Lamine Fetni, Laura Spillner, Thomas Muender, and Rainer Malaka. 2024. "I Know What You Mean": Context-Aware Recognition to Enhance Speech-Based Games. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3613904.3642426

1 INTRODUCTION

Speaking to computers has become part of the daily life of many people in recent years. With the new advances in artificial intelligence and language processing, people can now communicate with their smart devices using natural language. Using speech as an interaction modality has been implemented in cars, homes, and workplaces. Most new smartphones include a voice assistant feature. Moreover, home assistants such as Amazon Echo, Apple HomePod, and Google Nest are some of the many households that support users in everyday tasks such as smart home control, weather forecasts, setting reminders, and more [6, 10].

Voice interaction has also been an interest within the entertainment industry, and several game companies have been adapting voice-activated services to their products. Due to the improvements in speech recognition technology and growth in the availability of microphones in consumer gaming devices, a great opportunity has been created for video game companies to utilize voice interaction in their games [2]. By allowing players to use an intuitive and natural form of interaction, using voice in games can increase social presence within the game [26], provide a higher level of immersion [30, 65], and ultimately enhance player experience [62, 63].

While voice interaction has been used in various game genres, it often remains an optional feature due to ongoing recognition challenges. While the technology has been improving extensively, speech-based systems are still susceptible to recognition failures, and creating a seamless experience for users is still highly challenging. Since VUIs rely on automatic speech recognition (ASR) systems, they still cannot entirely avoid specific usability issues, such as bypassing non-speech conversational cues [36]. Moreover, aspects such as background noises, hardware limitations, and language barriers add to the complexity of this process [53].

A critical element to consider for video games is that they are mainly goal-oriented activities with various challenges. Players enjoy solving these challenges and working towards the goal [29, 47]. When issues with speech recognition occur, it adds a new layer of challenge to the game's existing ones, preventing players from reaching their goals and staying in a state of flow [19], often resulting in players getting frustrated and ultimately abandoning the game [20, 63].

For many speech systems, during the development process, a speech recognizer's vocabulary is created where a specific set of commands is defined. Nevertheless, the recognition accuracy is often decreased due to the acoustic similarity between different commands [64]. A standard method to better predict the intended command and improve the recognition is to compare the recognized output text with all the available commands defined based on their Levenshtein distance [31], which is the minimum number of single-character edits required to change word one into word two and execute the one with the lowest distance [66].

To enhance the player experience, in this work, we designed a speech-based video game that uses a new method for handling speech commands given by the players in the game using the game environment and actions. In a between-subjects user study with 40 participants, we compared a conventional method using the Levenshtein distance with a novel approach implementing context-aware speech recognition within the game. In the control group, when a command did not precisely match any available commands at that level, the recognition system would calculate similarity scores using the Levenshtein distance between the recognized input and the available commands and execute the one with the lowest distance. If the Levenshtein distance were higher than a set minimum, the system would consider that command unrelated and trigger the fallback interaction. We refer to this as the *scope filter* in this work. In the intervention group, we additionally implemented an *environment* filter and an *actions* filter. The *environment filter* takes into account the environment of the player inside the game at the given moment, while the *actions filter* is based on context information about the possible commands at a certain point in the gameplay.

For this study, we aimed to answer the following research questions:

RQ1: Can data derived from the game environment and actions aid command prediction?
RQ2: Does using context-aware speech recognition based on game environment and actions enhance usability and player experience?
RQ3: Does using context-aware speech recognition enhance players' performance in a speech-based game?

Our results indicate that context-aware speech recognition significantly increased the perceived usability of the in-game speech system and led to a significantly higher player experience. Players also perceived a significantly lower number of recognition errors in the intervention group, while their performances did not differ. Moreover, we discuss the insights of employing context-aware speech recognition in speech-based games. The implications of our work can support game designers when adapting speech interaction for their games. Its broader insights can also be valuable for using speech in other virtual environments to enrich voice recognition accuracy. Our study findings can be further transferred to other technologies and applications within the field of human-computer interaction (HCI), where data from other user-related sources, such as gaze or previous user actions, can help accurately identify users' intended commands.

2 RELATED WORK

Voice recognition technology has come a long way in recent years, with advances in machine learning and natural language processing allowing for the development of highly accurate speech recognition systems [23]. This technology has great potential to enhance human-computer interaction by enabling users to interact with computers simply by speaking to them. Researchers in the field of

HCI have been extensively exploring interaction with voice user interfaces. This section discusses previous work on voice-based video games and recognition in speech systems.

2.1 Voiced-Based video games

Voice-controlled video games have been developed since the 1970s [46] in a few experimental games that used voice interaction as a novelty feature. With the release of more modern consoles in the 2000s, game companies embraced this interaction more often. In 2002, Xbox introduced voice interaction through a microphone peripheral in the Xbox Live Headset, allowing players to control certain aspects of the game, such as navigating menus or selecting options with their voice. With the release of the Nintendo Wii in 2006, voice interaction became a more mainstream feature in video games. The Wii's motion-sensing controller was equipped with a built-in microphone that could be used for voice-based input. These early systems paved the way for more sophisticated voice-controlled features in games. In the years that followed, the use of voice controls in video games continued to evolve. Today, voice interaction is a common feature in many video games, particularly on platforms like the Xbox and PlayStation, which have built-in microphones and voice recognition capabilities.

In recent years, along with notable technological advancements, there has been increased research on voice interaction in video games. For example, Hong et al. [27] created a two-player conversational defense game utilizing voice input to enhance person-to-person communication. Their results revealed positive responses, especially regarding empathy and behavioral involvement. Anzai et al. [8] conducted a study comparing the impact of game interfaces with voice dialogue on the development of intimacy with screen characters, as opposed to conventional games. Their findings indicated that incorporating voice dialogue led to heightened enjoyment and increased intimacy with screen characters.

However, despite this growth in research, the exploration of voice-controlled video games—where voice interaction is integral—is still relatively limited [63]. This leaves multiple unexplored factors and questions about voice interaction in games and how they should be appropriately adapted into the gameplay [5, 63]. Allison et al. [4] argue that players' voice commands in a video game are associated with a sense of taking on a character from the game's world. Ignoring such aspects obstructs the player's engagement with the in-game world [5]. Carter et al. [15] note that the effective integration of voice interaction in video games differs from other contexts, requiring careful consideration of the voice's identity—whose voice is recognized and how it is embodied. In video games, where voice interaction is not tied to the virtually embodied experience, this can create dissonance between the user and their character, adversely impacting the gaming experience.

Voice-controlled games have also been developed for educational purposes. For instance, Filimon et al. [21] made a voice-based geography game for the Amazon Echo, while Jung et al. [28] created a voice-controlled game to teach kids programming. Their research found that the games effectively boosted immersion and understanding of the educational concepts. Besides the possibility of adapting intuitive and novel game mechanics, voice interaction for games could be especially important for users with disabilities [60]. People with deficits in motor control or vision are unable to play video games with standard controls such as a mouse and a keyboard; hence, they are excluded from this form of entertainment and social interaction [24, 37]. Moreover, speech-based games have been shown to be practical for speech therapy while enabling remote treatment. Navarro Newball et al. [39] designed a voice-controlled video game to rehabilitate children with early-diagnosed hearing disabilities. Authors found that the narrative and entertainment elements of the game led to an engaging experience, thus favoring the repetitive approach needed for speech mechanization sessions. Ahmed et al. [1] explored the feasibility and user experience of speech-controlled games for children with childhood apraxia of speech and typically developing children. Their findings indicated that these games are enjoyable for both children and speech-language pathologists, suggesting their potential to foster higher-intensity practice for children. To enhance performance and address limitations associated with speech-only interactions, voice interaction in games can be combined with other human modalities [51]. For instance, van der Kamp and Sundstedt [58] employed a combination of gaze and voice commands to enable hands-free interaction with a drawing application. Their study indicated that participants preferred their proposed multimodal approach despite providing less control than the conventional mouse and keyboard setup. Similarly, Hedeshy et al. [25] introduced a hands-free video game interaction method using non-verbal voice interaction and gaze, comparing it with a standard mouse and keyboard. Results showed a preference for their multimodal approach as an exciting, engaging, and fun game interaction method.

Voice interaction in video games is categorized into verbal and non-verbal forms [4]. Verbal forms of voice-controlled games are speech-based games that require recognition technology. In such games, players use complete words or sentences as input to interact with the game [63]. Non-verbal forms are those that use other characteristics of voice, such as pitch and volume. Such games do not require a speech recognition system. Games that use non-verbal forms of interaction have been shown to be more successful than speech-based games as they avoid having recognition problems [3, 4]. Nevertheless, this voice-controlled game category has very limited and restricted game mechanics. Recognition issues are one of the main challenges with speech-based games and a reason for players' reservations about playing such games [41]. It is also one of the main justifications that voice interaction is commonly an optional feature in games rather than a core aspect of the game design. In an attempt to maintain the game's flow and minimize player frustration, Zargham et al. [63] developed anticipatory error handling for a speech-based video game. In their approach, when players' intents were not recognized, the game would perform a locally optimized action considering goal completion and obstacle avoidance. Their results showed that, although their approach could improve the usability of their speech system, it does not necessarily lead to a better player experience if the anticipated action does not follow the user's intention, even when technically optimal decisions were made. Another downside to this approach is the players' potential misuse of the error handling system by purposefully giving unclear commands and being confident that the system would perform the optimal action.

2.2 Recognition in Speech Systems

Researchers believe speech recognition challenges and restricted functionality are the main factors contributing to disapproval or non-adoption of voice systems [9]. Despite notable progress in ASR systems, reaching accuracy levels exceeding 90% [45], speech-based systems are still prone to recognition inaccuracies [63]. When the system can not correctly recognize users' speech input, user dissatisfaction arises [43], often leading to a lack of progress or the inability to complete tasks [34].

Three primary sources of failures in human-agent speech interaction involve instances where the system fails to understand the user's command, the provided command is out of context (not in the vocabulary), or the command is misunderstood [32]. Several reasons could lead to such recognition issues, including users giving complicated or fuzzy input to the system, background noise, the system having a limited vocabulary, or faulty hardware [7]. Recognition issues have driven users to commonly speak differently than speaking to a human when interacting with a VUI. Many expect natural language not to be adequately understood by such systems and adapt particular communication strategies. Reducing the talking pace, reformulating command sentences, speaking loudly, and physically relocating themselves or the system are popular observable patterns when users are confronted with recognition errors [9]. These user tactics are referred to as hyperarticulation [54] and are deployed to resolve recognition problems. Moreover, language barriers can also add to the issue of inaccurate speech recognition. Pyae and Scifleet [44] found that VUIs are more useful and easier for native English speakers to interact with than for non-native speakers.

On the software side, researchers and developers have explored different approaches to enhance ASR. A standard method to enhance recognition is to train the data with extensive voice samples [32]. A study by Rosenberg et al. [48] demonstrated that augmenting training data with synthesized material can enhance speech recognition. Moreover, machine learning techniques, such as deep learning [38], have been proposed to learn the underlying patterns in speech data and improve voice recognition accuracy [22]. Other studies have recommended using multimodal approaches where the system can combine the input from multiple sources to improve recognition accuracy [37, 55]. For instance, if the voice recognition result is uncertain, the system can use the keyboard or gesture input to confirm or correct the recognition result.

All in all, considering the advances in speech technology, technical limitations and recognition issues are still one of the main reasons for user frustration and their skepticism towards using VUIs [33, 59].

In a human-human interaction scenario, imagine someone entering a room and saying, "Where's the charger?" This inquiry gains clarity and precision when contextual information is considered. Observing the room's environment, including where electronic devices are typically located, where people are currently situated, and recalling previous actions and events (such as someone using a laptop earlier) aids in comprehending the request. This holistic understanding allows the respondent to accurately direct the inquirer to the specific location of the charger. Similarly, in human-agent interaction within a virtual setting, incorporating contextual details, such as the virtual environment's layout, the location of objects, and the user's gaze direction, along with a history of the user's actions, could potentially enhance the virtual agent's ability to decipher user commands within the specific context. Just as humans rely on contextual cues to navigate and understand their surroundings, integrating contextual information in virtual interactions aligns with natural communication patterns, which could optimize the virtual agent's responsiveness and overall user experience.

In our approach, we build upon the prior work by proposing a novel method that uses game environment data and actions to handle unrecognized intents to enhance speech recognition for video games. We aim to improve the technical limitations of speech recognition using our context-aware speech recognition method and investigate its effect on players' experience, performance, satisfaction, and usability of the speech system. To the best of our knowledge, this is the first exploration of this kind of recognition in a speech-based video game.

3 GAME DESIGN

In order to address our research questions, we have developed a speech-based video game called "Escape the Echo," an escape room game where players have to communicate with the main character "Sophie" using speech commands. The game begins with Sophie waking up in a closed room where she realizes she is connected to someone (player) via a communication device, and they can see the room using Sophie's handheld video camera. The character then asks the player for help in finding a way out. The player's goal is to support Sophie in escaping various rooms by assisting her inspecting particular objects and using them to exit the room.

The game consists of three levels (rooms) - a jail cell, a bathroom, and a classroom. In each level, there are a series of actions that players can instruct Sophie to take in order to progress and escape. As players advance through the game, they receive hints suggesting that Sophie is dreaming and that the entire escape is happening in her mind. This sets up the game's plot twist ending, in which players must instruct Sophie to wake up from her dream and trigger the game's conclusion.

3.1 Mechanics

Players have to instruct Sophie to perform specific tasks based on the in-game objects (e.g., mirrors, desks, or doors) using voice commands. Sophie could take actions, including finding and inspecting objects, moving objects around, breaking objects, and solving puzzles to unlock new areas. Performing certain actions, such as triggering alarms or making too much noise, could also lead to failure. An end screen would appear in such cases, and players could restart from the last checkpoint. If the player targets an interactable object in the room, the object's name appears in the center of the screen, as shown in Figure 2. An instruction manual was implemented to inform users about the game's procedure, controls, and events. The game starts with a basic intro, where the player is introduced to the voice interaction by replying to Sophie with simple words such as 'Yes,' 'No,' and 'Okay.' When pointing at interactable objects, a list of actions appears on the top-right corner of the screen (e.g., inspect, move, or break) (see Figure 2). Players

Figure 2: When players target interactable objects in the game, the object's name appears in the center of the screen.

could determine how to instruct the character based on the displayed actions. For instance, when the player sees the hint "Break" and is targeting the game object "mirror," they can assume that the voice command for this action is "Break the mirror." For some of the more complex actions, players have the option to receive hints and view the full command for that action (see Figure 3). Every interactable object has an 'inspect' command (e.g., inspect the table) that, after being executed, would reveal other possible actions (hidden actions), if any existed. Nevertheless, the player can still instruct Sophie to perform the hidden actions regardless of their visibility in the "actions" list. The game has a total of 86 unique actions and 36 unique interactable objects distributed between the three levels. The speech system was programmed to handle various phrases for each action. For instance, if the player wanted to tell Sophie to "Break the mirror," saying "Try breaking that mirror" or "Smash the mirror" would also be acceptable. If the command has been executed already, Sophie would reply, "I have already done that." If the command could not be performed on that game state, she would reply, "I cannot do that."

The player controls are limited to the mouse movement to look around the room, which is perceived in-game as the player controlling the handheld camera that Sophie is holding. Players cannot move the character around, as that would contradict the story and the player's identity established in the game. The only way to move in the game was to instruct Sophie to perform an action, and the character would automatically move based on that intended action. Players can communicate with Sophie at any time during the game using speech. They can also mute their mic by pressing the 'M' key and pause the game by pressing the 'Escape' key. To save time on animating, during the execution of an action, most of the time, Sophie would put the camera down, and a sound effect of that action would be played, and when the action was done, Sophie would pick the camera back up. After proceeding to the next room, players could not go back to the previous room unless they started the game from the beginning.

3.2 Implementation

The game was developed using Unity 3D [1]. For speech recognition, we initially used Unity's built-in speech recognizer for Windows, which uses Windows's built-in speech recognition engine. However, this required an internet connection, and the connection quality could impact the response time of the in-game interactions. As the study was planned to be conducted remotely, we could not guarantee a convenient internet connection for every session. As an alternative, we used VOSK Speech Recognition API [2], model `'vosk-model-small-en-us-0.15'`, which works offline. VOSK is an open-source speech recognition toolkit that offers continuous large vocabulary transcription, zero-latency response with streaming API, customizable vocabulary, and speaker identification capabilities [57]. We created a build for Windows only due to compatibility reasons.

3.3 Speech Recognition

When a command exactly matched one of the commands in the vocabulary, the game would execute that command, and no filter would be applied in both study conditions. However, if a command did not precisely match any available commands at that level, the

[1]https://unity3d.com/unity
[2]https://alphacephei.com/vosk/

Figure 3: When pointing at interactable objects, a list of actions appears in the top-right corner of the screen. The already performed actions are crossed out.

recognition system would calculate a confidence score for each possible command. In the control group, this score was based only on the *scope filter*, which used the Levenshtein distance. In the intervention group, the final confidence score was calculated as a weighted sum of the three scores based on the *scope filter*, *environment filter*, and *actions filter*. In the first step, the set of possible commands was limited to those commands in the vocabulary that were similar to what was recognized. We set a maximum Levenshtein distance threshold of 20 - all commands with a distance > 20 were not considered possible commands. This number was chosen empirically after our initial testing sessions of the game, as it showed to be an appropriate number to effectively detect phrases that are too long, too short, or too different from the list of accepted commands (vocabulary). A fallback interaction was triggered if there was no possible command with a Levenshtein distance below this threshold. In such cases, Sophie would respond with a message indicating that she did not understand the player's instruction. Otherwise, the confidence score was calculated for all possible commands with a distance of ≤ 20, and the command with the highest total score was executed.

3.3.1 Scope Filter. The *scope filter* assigns a score to each command based on its similarity to the recognized output text. The system compares the recognized output text with available commands in the level, executing the one with the lowest Levenshtein distance. Levenshtein distance indicates the minimum single-character edits needed to transform the source string into a target string. For instance, suppose the player's recognized intent is "Bake the mirror," and the expected command in the system's vocabulary is "Break the mirror." In this case, the Levenshtein distance would be 3 (add 'r' after 'B,' add 'e' after 'r,' and remove 'e' after 'k'). If there were no other commands with a lower distance, then the "Break the mirror" action would be triggered. Due to the distance threshold, the Levenshtein distance of a possible command would be between 1 and 20. Let N be the Levenshtein distance of a given command to the output text, and M be the maximum possible distance based on the threshold. Then, the scope filter score was calculated as follows: $(M - N + 1) * 4$.

For commands with the minimum possible distance of 1, this score would be 80:

$(M - N + 1) * 4 = (20 - 1 + 1) * 4 = 80$

For those with a maximum distance of 20 (in which case M = N), it would be 4:

$(M - N + 1) * 4 = (20 - 20 + 1) * 4 = 4$

3.3.2 Environment Filter. The *environment filter* assigns scores to each available voice command in the current level based on the game environment at the time the command was given. The maximum possible score for this filter is 30. The environment filter score for a given command is calculated based on the interactable objects visible in the frame (camera view) and which, if any, interactable object the player is currently targeting (meaning that it is at the center of the field of view) (see Figure 4). If the object mentioned in a command is the one that is currently being targeted, this adds 15 points to the environment filter score of this command. Additionally, all those commands corresponding to objects visible in the frame but not necessarily being targeted also get a number of points calculated as $15/N$, with N being the number of objects in view.

Figure 4: All interactable objects within the frame, highlighted in red and orange. The targeted object is highlighted in orange.

This is because players might not necessarily aim at an object as long as it is visible. If three objects are in the frame and one is being targeted, then commands referring to the target object will receive an environment score of 15 + 5 = 20. The other two will receive an environment filter score of 5 each. This approach allows the system to prioritize commands related to more prominent objects in the frame. The presence of several objects in the frame can impact the accuracy of the environment score as it becomes more challenging to determine which object(s) the player is interacting with. Only if an object is being targeted and is also the only object in the frame can it get a maximum environment filter score of 30.

3.3.3 Actions Filter. The *actions filter* assigns scores to possible commands based on the actions that should be performed, that is, how a player wants to interact with an object instead of which object it is. Similar to the *environment filter*, the *actions filter* also has a maximum score of 30 and a minimum of zero. This filter takes into account four facts about the current context: whether or not the action has just been revealed as an option to the player (after the player inspected the same item in the previous step), whether or not the action is known to the player in general, whether or not it is possible in the current game state, and whether or not the action has already been tried in this state. If an action has just been revealed, it must be possible and is now known to the player. However, it can happen that the action has already been tried even before it was revealed. Thus, for a given possible command, one of the following will apply:

- All four facts are true, as the action has just been revealed and has not been tried before. In this case, the command receives 30 points (maximum).
- The action has just been revealed. However, the player has already tried it before in a previous step: 15 points.
- The action has not just been revealed, but it is possible, known to the player, and not tried yet: 15 points.
- The action has not just been revealed, and it is also not possible. However, it is otherwise known to the player and yet to be tried: 1 point.
- In all other situations, this command receives 0 points.

3.3.4 Final Confidence Score. In the control group, the final confidence score equals the *scope filter* score (and is thus highest for the commands with the smallest Levenshtein distance to the recognized command). In the intervention group, the final confidence score is calculated as a sum of all the three previously explained filter scores (see Figure 5). The maximum possible final score was 140, while the minimum was 4. The maximum number of points for each filter (80 for the *scope filter* and 30 each for the *environment* and *actions filter*) was chosen to weigh the filters' importance. The goal of this weighting was that the *scope filter*, which is based on the recognized text compared to the text of the possible commands, should still have a more significant influence on the final score than the two supplementary filters. We conducted preliminary tests based on which we learned that the average Levenshtein distance for possible commands was 11 (with the maximum allowed due to the threshold being 20). Considering this average, the *scope filter* score formula was defined. For a distance of 11, the *scope filter* score is 40 (out of a maximum of 80 for a distance of 1). Based on this, the other two filter scores were set to a maximum of 30 each so that in cases of comparatively small Levenshtein distance, the *scope filter*

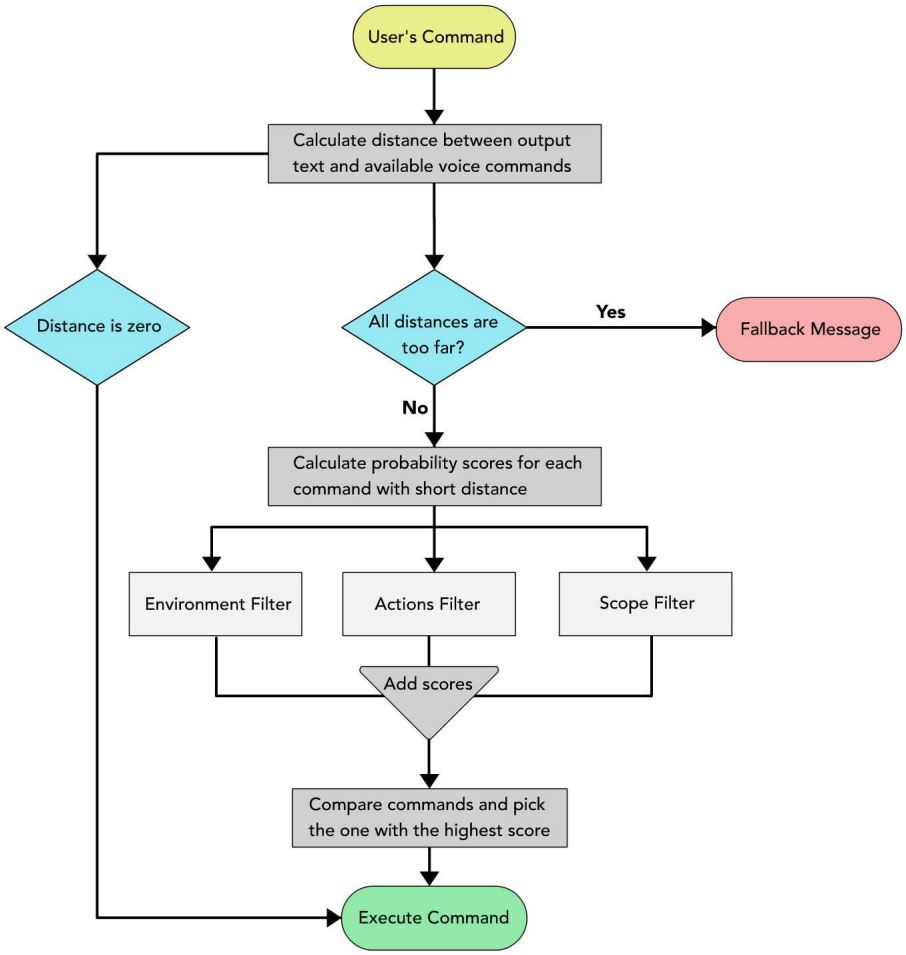


Figure 5: The general process of the command prediction in the intervention group using all three filters.

would outweigh the information from the *environment* and *actions filters*.

After the scores were calculated, a comparison was made between the different commands based on their confidence score, and the command with the highest score was executed. If there were several commands with the same high score, a random selection would choose the executed command.

For instance, in the game situation presented in Figure 6, the player gave the command "Try to open the door." The *environment filter* would calculate a score of 30 for the command "open the door" in the vocabulary as the player is targeting the door, and it is the only intractable object in the frame (15 + 15/1). Assuming the player has just used the 'Inspect' action on the door, the *actions filter* would also give the same command a score of 30 as it is known to the player, the player has just learned it, the command was not used before, and it is possible to execute that command in this game state. If the recognized input by the system was "trial to open the that door" it will result in a similarity distance of 8. Therefore, the *scope* score will result in 52 ($(20 - 8 + 1) * 4$), which ultimately makes "try to open the door" the predicted command, resulting in a high score of 112 (30 + 30 + 52).

4 STUDY DESIGN

To answer our research questions, we conducted a between-subjects user study with ($N = 40$) participants to evaluate and compare our two study conditions. All game aspects, such as the levels, puzzles, game environment, and mechanics, were identical between the two game versions. The only difference between the two was in the recognition method. Group assignment was pseudo-randomized between two equally distributed groups. The participants were instructed to play all three levels of "Escape The Echo" from start to finish - although they had the option to skip to the next level or quit the game earlier if they got stuck or did not want to finish. Each participant played the game on their own Windows PC or laptop. An

Figure 6: A game situation where the player gives an instruction with the intent of "Try to open the door."

executable version of the game (build) was sent to all participants prior to the session. The experimenter made sure the participant had a working microphone and that the game ran without issues before the session. To ensure the reliability of the results, we also controlled for external factors that could impact the game experience, such as the type of computer used, monitor size, and the background noise level in the room.

4.1 Procedure

Each session was held remotely via video calls to ensure the participant's convenience. The experimenter noted verbal statements and in-game observations while assisting when the participants encountered issues. The sessions began with the experimenter briefing the participants about the procedure and explaining the game and its controls while reminding them to read the in-game instructions. After the participants gave informed consent, they would then play through the game in either one of the two conditions. While playing, the experimenter would mute their microphone unless the participant requested help. The calls were mainly voice-only, where the players could ask questions or communicate their needs. When participants had issues or wanted to show something, the players could share their screens with the experimenter. Participants could also take short breaks between levels. After playing the game, participants were asked to fill out the post-exposure questionnaires. A short semi-structured interview, which was audio recorded, was held at the end of the session. The interviews took an average of 6.15 minutes (SD = 3.26). Each session lasted approximately 40 – 60 minutes, with the average gameplay time around 36 minutes (SD = 9.04).

4.2 Participants

We recruited (N = 40) individuals using a convenience sampling approach to participate in our user study. The selection was based on social networks, word of mouth, gaming communities, and university mailing lists. Participation in the study was voluntary and uncompensated. Only 17.5% of participants had played a voice-controlled game before, while 82.5% had experience with voice-controlled applications. Of those who had used a voice-controlled application, 54.5% reported having experienced issues with voice recognition. The control group consisted of 20 participants (12 self-identified as male, seven as female, and one as non-binary) aged between 18 to 35 years (M = 23.75, SD = 4.06). Three participants in this group were native English speakers, while the remaining were fluent non-native English speakers. The intervention group comprised 20 participants (12 self-identified as male, seven as female, and one as non-binary) aged between 18 to 34 years (M = 23.10, SD = 3.41). Four of this group were native English speakers, while the rest were fluent non-native English speakers. Most participants played games regularly, with 85% reporting playing at least once a week (13 every day, 15 a few times per week, six once per week, three once per month, and three never). The experiment sessions, including the interviews, were conducted in English. However, three participants requested to have their semi-structured interviews conducted in Arabic because they felt more comfortable being recorded in their native language. These interviews were later transcribed and translated by the experimenter.

4.3 Measures

We used a series of standardized questionnaires to assess player experience and the perceived usability of the speech system. The post-exposure questionnaires included demographic questions, the System Usability Scale (SUS) [14], and the Player Experience of Need Satisfaction (PENS) [49] throughout the subscales of *Competence*, *Autonomy*, *Presence/Immersion*, and *Intuitive Controls*, while excluding *Relatedness* as it was not relevant to the scope of the study. We chose the SUS as it is an established and reliable tool for measuring a system's perceived usability. In our study, we used it to measure the perceived usability of the speech system. PENS is also a validated questionnaire for determining the player experience within multiple sub-scales.

Additionally, we recorded a series of customized questions regarding players' experience with the game. These were implemented via seven-point Likert scales, including questions about speech recognition and estimating the number of commands they had issues with. Players were also asked about their perceived performance, enjoyment, overall game experience, and willingness to play similar speech-based games in the future.

We also conducted a short semi-structured interview with each participant to evaluate further the qualitative factors of the player experience, usability, and individual preferences [61]. The interview included questions about likes and dislikes concerning the game, the most and least exciting aspects, and the players' thoughts on the voice recognition system and the recognition errors.

4.4 Data Analysis

4.4.1 Quantitative Analysis. Regarding the statistical analysis, the Shapiro-Wilk test was conducted to assess the normality assumption of the data [52]. We conducted unpaired t-tests (when the data was normally distributed) and Mann-Whitney U Tests (when data was not normally distributed) to identify the differences between the two conditions. We applied an alpha level of .05 for all our statistical tests.

4.4.2 Qualitative Analysis. The audio recordings obtained from the interviews were transcribed verbatim. The interview data were then analyzed and coded based on domain summaries [12, 18], where the themes are structured around a shared topic rather than shared meaning, with the goal of capturing the diversity of meaning in relation to a specific subject or area of focus [35]. Broadly, the interview questions centered around players' positive and negative impressions of the game, as well as their perceptions and thoughts regarding the game's speech recognition.

The analysis began with data familiarization and categorization [11]. Initially, two researchers read through the responses to get a sense of the content and context to understand the patterns, ideas, and concepts present in the responses. To develop a coding system, the transcripts of a random selection of 15 interviews were independently coded by two researchers using inductive coding [16, 56], where a single quote could be assigned to multiple codes, including descriptive, conceptual, or emotional codes. The researchers then agreed upon a coding system after a thorough discussion. In cases of disagreements, an additional author was consulted to reach a consensus. Subsequently, an iterative discussion between the two authors led to the establishment of a coding manual. The remaining transcripts were then individually coded by one author, utilizing the coding manual. During this process, noteworthy and unique player statements were also collected. As the evaluation proceeded, some new codes emerged, requiring the coding manual to be adjusted accordingly. The coding manual can be found in the supplementary material. This process resulted in extracting key insights and findings from the analyzed responses, which are presented in subsection 5.4. Two participants (one from the control group and one from the intervention group) did not participate in the semi-structured interview sessions. Therefore, the interview responses were evaluated with 38 participants (control: 19, intervention: 19).

4.4.3 Game Logs. After each gameplay session, a log file was generated containing information on the total number of given commands, the number of directly recognized commands (without using filters), the number of predicted commands (used filters), playtime, average prediction scores of the three filters (*scope*, *environment*, and *actions*), the overall confidence score for commands, and the number of predicted commands that would have had the same outcome if they had been predicted using only the *scope* score (to understand how many commands could have been predicted without the *environment* and *actions* filters). Although the control group's gameplay was only affected by the *scope filter* during the experiment, we still logged the data given by the *environment* and *actions* filter for analysis.

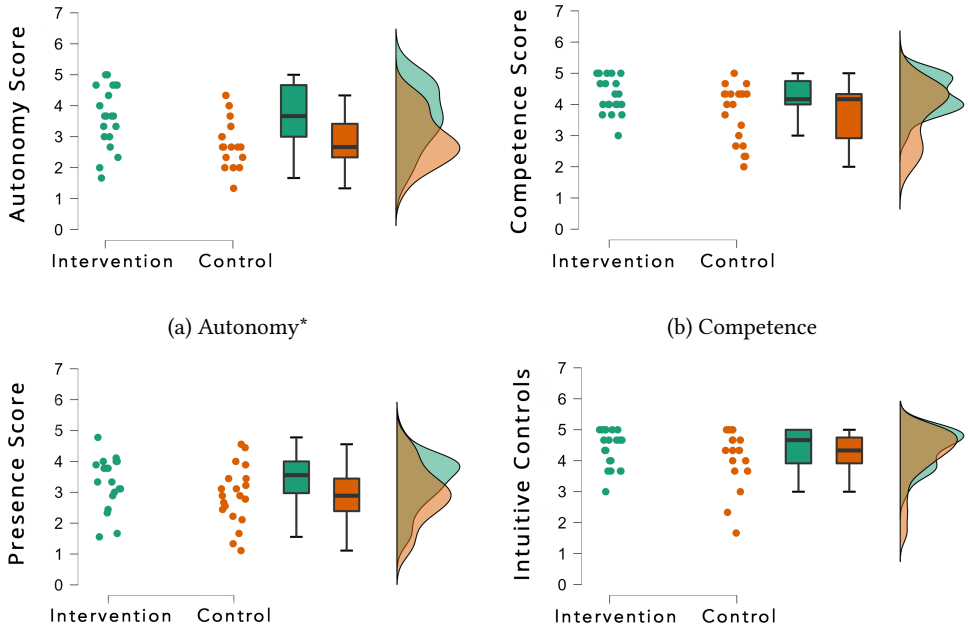
5 RESULTS

Both quantitative results from the questionnaires and qualitative insights from the interviews were gathered in our evaluation. Our results will be presented in this section. Throughout the study, only one participant in the control group did not finish the game and quit after the end of the second level due to time limitations.

5.1 Standardized Questionnaires

To evaluate the results of the SUS scores, we conducted an independent samples t-test, as the Shapiro-Wilk test indicated that the data were normally distributed (see Table 4). The results indicate that the intervention group with context-aware speech recognition outperformed the control group significantly in terms of usability ($t(38) = 2.57$, $p = .014$) with a large effect size ($d_{Cohen} = 0.82$) [17]. When Interpreting SUS scores as percentiles, the mean score of the control group ($M = 73.2$, $SD = 17.22$) results in a good rating (B-), while the intervention group ($M = 85.12$, $SD = 11.31$) results in an excellent rating (A) in terms of usability [50]. As any SUS score above 68 would be considered above average [13], our results indicate above-average usability scores for both conditions.

Regarding the PENS questionnaire, we applied Mann-Whitney U Tests to evaluate the two subscales of *Competence* and *Intuitive Controls*, as the Shapiro-Wilk test highlighted a violation of normal distribution. However, for the *Presence* and *Autonomy* subscales, we employed independent samples t-tests as no deviation from normality was observed (see Table 4). We found a significant difference ($t(38) = 2.52$, $p = .02$) for *Autonomy* in favor of the intervention group ($M = 3.65$, $SD = .99$), in comparison to the control group ($M = 2.92$, $SD = .84$), displaying a large effect ($d_{Cohen} = 0.8$). No significant differences were witnessed for the subscales *Competence*,



(a) Autonomy*

(b) Competence

(c) Presence

(d) Intuitive Controls

Figure 7: The distribution of variables and the mean and confidence intervals of the PENS results between the control and intervention groups.

Table 1: Descriptive statistics and results of the Mann-Whitney U test for the post-exposure customized questions. Statistically significant results are marked with asterisks.

	Control Mean (SD)	Intervention Mean (SD)	*p−value*	*U−value*	Effect size
Game Experience	5.35 (1.18)	6.30 (0.80)	0.004 *	98.5	0.508
Game Enjoyment	4.85 (1.38)	5.75 (1.25)	0.031 *	122.0	0.681
Perceived Error Rate	3.83 (2.78)	1.65 (1.69)	0.009 *	104.5	0.478
Performance Rating	5.05 (1.23)	5.45 (1.23)	0.312	163.5	0.390
Willingness to Play Similar Games	5.25 (2.29)	6.50 (0.83)	0.089	142.5	0.288

Intuitive Controls, and *Presence* between the two conditions ($p > .05$) (see Figure 7).

5.2 Customized Questions

In terms of our customized questions, we also used the Mann-Whitney U test to identify significant differences due to the violation of normality detected by the Shapiro-Wilk test (see Table 5). All customized questions were asked on a seven-point Likert scale (see Table 1).

For the overall game experience (extremely bad to extremely good), players of the intervention group rated it as 6.3 ($SD = 0.801$), significantly higher than the control group ($M = 5.35$, $SD = 1.183$; $U = 98.5$, $p = .004$), revealing a medium effect ($r = .508$). Similarly, assessing the enjoyment ratings of participants, we found significantly higher scores in favor of the intervention group ($M = 5.75$, $SD = 1.25$) in comparison to the control group ($M = 4.85$, $SD = 1.387$; $U = 122$, $p = .031$) with a medium effect ($r = .681$). When participants were asked about the number of commands they had recognition issues with while playing the game, the number in the intervention group ($M = 1.6$, $SD = 1.69$) was significantly lower than the control group ($M = 3.83$, $SD = 2.78$; $U = 104.5$, $p = .009$), revealing a medium effect between conditions ($r = .478$). Regarding players' perceived performance and willingness to play similar games, we did not observe significant differences between the two conditions ($p > .05$).

5.3 Game Logs

We analyzed the logs retrieved from the gameplay sessions. Due to technical issues with log generation, six log documents could not be retrieved after the sessions (three from the control group and

Table 2: Descriptive statistics and results of the Mann-Whitney U test for the game logs.

	Control Mean (SD)	Intervention Mean (SD)	$p-value$	$U-value$	Effect size
Play Time (seconds)	2188.88 (640.82)	2126.69 (463.69)	0.897	124.0	0.031
Total Inputs	101.13 (27.51)	93.44 (24.11)	0.180	92.0	0.281
Commands Predicted %	68.85% (32.02)	76.19% (21.31)	0.308	155.5	0.215
Average Confidence Score	89.1 (9.82)	88.0 (10.72)	0.539	111.0	0.113

Table 3: Descriptive statistics and results of the independent samples t-tests for the game logs. Statistically significant results are marked with asterisks.

	Control Mean (SD)	Intervention Mean (SD)	$p-value$	$T-value(32)$	Effect size
Direct Recognition %	26.64% (13.13)	18.80% (12.50)	0.047 *	2.068	0.611
Environment Score	23.9 (2.00)	23.7 (1.68)	0.735	0.341	0.108
Actions Score	16.8 (2.43)	15.8 (3.22)	0.318	1.013	0.350
Scope Score	48.4 (6.24)	48.5 (6.31)	0.947	0.067	0.015
Same Outcome With Only Scope	40.9 (13.85)	45.06 (9.78)	0.353	0.941	0.346

three from the intervention group). Therefore, the total number of logs analyzed was 34, equally divided between the two conditions. Due to the observed deviation from normality indicated by the Shapiro-Wilk test, we employed the Mann-Whitney U test to identify significant differences in log values between groups for variables *playtime*, *total number of voice inputs*, *total predicted commands*, and *average confidence score*. For *directly recognized commands*, as well as the *environment*, *actions*, and *scope* filter scores, we conducted independent samples t-tests since the Shapiro-Wilk test revealed a normal distribution of the data (see Table 6).

The average playtime measured for the intervention group was 2126 seconds (SD = 463.69), while it was 2188 seconds for the control group (SD = 640.82), showing no significant differences ($U = 124$, $p = .89$) (see Table 2). In terms of the total number of voice inputs given in a gameplay session, no significant differences were observed between the intervention group ($M = 93.44$, $SD = 24.11$) and the control group ($M = 101.13$, $SD = 27.51$; $U = 92$, $p = .18$). The intervention group had 76.2% of commands predicted through filters, while the control group had a 69.9% rate, showing no significant difference. The speech recognition system directly recognized 18.8% of commands for the intervention group and 26.6% for the control group (intents where no filters were applied). This shows a significantly higher direct recognition rate in the control group ($t(32) = 2.068$, $p = .047$), revealing a medium effect ($r = .611$). Considering both the total predicted and the directly recognized commands, the rate of unrecognized commands (with a similarity distance above 20) stood at 5% in the intervention group and 4.5% in the control group.

Although we only used the *scope filter* to predict commands in the control group, in this group, the overall confidence scores, including the *environment* and *actions* scores, were still logged for analysis. The average command confidence score for the executed commands in the control group was 89.1, while the intervention group's average was 88. The average *environment*, *actions*, and *scope* scores were 23.9, 16.8, and 48.4, respectively, in the control group, and 23.7, 15.8, and 48.5 in the intervention group, showing similar scores for the individual filters in both conditions (see Table 3).

Furthermore, from the logs, we also witnessed that, on average, 45.06 of the predicted commands in the intervention group (SD = 9.78) would have had the same outcome if there were no effects from the *environment* and *actions* filters. This number is not significantly higher than in the control group ($M = 40.9$, $SD = 13.85$; $t(32) = 0.941$, $p = .353$). This highlights that, on average, 63% of the commands predicted by our context-aware speech recognition in the intervention group would have had the same result if they were predicted with the *scope filter* only, meaning that 37% of predicted commands were chosen differently due to context-aware speech recognition.

5.4 Qualitative Findings

We analyzed interview responses to extract qualitative results. The presentation of our findings encompasses participants' positive and negative impressions, as well as their perceptions of the game's speech recognition.

5.4.1 Positive Impressions. Participants generally held a positive impression of the game, describing it as engaging, fun, immersive, and exciting. They stated comments such as: "The game was really fun!" (P4) or "I felt engaged with the game. It was quite exciting!" (P29). A number of participants (18.4%) requested to play more levels of the game. Several players (26.3%) praised the game's narrative. 42.1% found the game's ending exciting: "most interesting part of the story." (P22). Four participants (control: two, intervention: two) mentioned that the game was highly immersive, with one participant saying: "It was a story-driven game, and you were included in the story." Another player mentioned: "When you talk to Sophie, you forget it is a game. You feel like she is there, and she needs your help." Many participants (42.1%) expressed that they

liked the speech-based aspect of the game in particular: "I like using voice commands for the game. It is not something we usually see in games." (P16). Several participants (44.7%) indicated they liked the game's design and aesthetics.

5.4.2 Negative Impressions. When participants were asked what they disliked about the game, 31.6% (control: 4, intervention: 8) said they did not like when the character put the camera down to perform an action. They mentioned that the animations of these actions could have been more exciting, and only playing the sound effects felt like "missing out on the ongoing in-game events." (P10). Seven players (18.4%) noted that they did not like that they could not freely roam around the room and control the character's movements. They suggested adding the possibility of free movement in the game using the keyboard to give them more autonomy. Four participants found this aspect the biggest downside of the game. Although many players (26.3%) praised the game's underlying story, 13.1% (control: 1, intervention: 4) found the story rather uninteresting. Four participants found the game challenges too easy, and four others found having to inspect objects repeatedly frustrating.

5.4.3 Voice Recognition. In both groups, most participants (94.7%) responded positively when asked about voice recognition in the game. 39.5% of the participants (control: four, intervention: 11) mentioned that they did not experience any recognition issues throughout the game. Two players from the intervention group said the voice recognition was better than voice-controlled games they had played before. One participant said: "I would like to see a similarly good voice interaction in triple-A games" (P34). However, 26.3% (control: nine, intervention: one) of the players mentioned that they experienced difficulty with recognition.

When participants were asked about the instances where the system did not recognize their commands, and the character would respond with "I do not understand" or similar statements, 21% (control: five, intervention: three) reported negative feelings toward this fallback method. Four participants (control: two, intervention: two) mentioned that these instances made them feel frustrated, annoyed, and confused. On the other hand, 23.7% (control: three, intervention: six) stated that this non-recognition fallback method was fair or made sense to them. One participant compared it with other VUIs, such as Google Assistant, saying, "It was normal. I mean, it is expected from something like this. Even Google does not pick up what you say all the time." (P26).

When asked about instances in which the game did not perform their intended actions by performing a different action, ten participants (control: six, intervention: four) reported a negative feeling. Players mentioned that such instances made them frustrated, angry, and irritated. One participant from the intervention group reported that such instances felt like a bug in the game. Two players from the control group mentioned that a recognition mismatch made them fail since the game went against their initial intention. Two participants from the control group found such occurrences funny.

While two participants (5.2%) in the intervention group praised the fact that they could phrase the commands differently, 23.7% (control: six, intervention: three) criticized the variability of the accepted commands for performing a particular action. One player mentioned: "Sometimes you have to form the command only in a specific way for it to be recognized" (P7). Participants recommended adding more supporting commands and expanding the accepted vocabulary. 13.2% (control: two, intervention: three) of players noted that while the voice recognition was decent, they sometimes struggled with the recognition due to their accents.

6 DISCUSSION

This work aimed to investigate the impact of using game environments and actions in the form of a context-aware speech recognition technique on player experience and usability in a speech-based video game. Players generally enjoyed playing "Escape the Echo" in both groups and gave positive feedback. They found controlling the game with their voice exciting and novel. After the experiment, participants asked if there would be new rooms where they could play, and several expressed a desire to replay the game to discover everything else. Our interviews revealed that players felt immersed in the game when speaking with the main character and felt they were in the game's world, supporting previous literature that the player's in-game voice commands can be associated with a feeling of taking on a character in the game's world [4, 63]. This finding also aligns with research on voice-controlled games, which suggests voice interaction in games can provide higher levels of immersion [30, 40, 65].

Eventually, we interpreted the results of this experiment to provide answers to the following overarching research questions:

RQ1: Can data derived from the game environment and actions aid command prediction?
RQ2: Does using context-aware speech recognition based on game environment and actions enhance usability and player experience?
RQ3: Does using context-aware speech recognition enhance players' performance in a speech-based game?

6.1 Supporting Command Prediction

The results from the game logs generated after the sessions show that, overall, our proposed method impacted 37% of the predicted commands. Even though the *scope filter* had a higher weight than the supplementary filters of *environment* and *actions*, they still had an impact on the outcome of around one-third of the total given commands. These findings provide strong evidence for the significant influence of our proposed method on command prediction. However, it is important to acknowledge that RQ1 cannot be fully answered due to the absence of ground truth and insight into the players' intended actions for each command. Nevertheless, considering the fact that the perceived error rates were significantly lower in the intervention group, one can assume that the utilization of context-aware recognition holds promising potential in effectively aiding command prediction.

In our experiment, recognition failures were classified into two distinct groups. The first group comprised commands with a Levenshtein distance exceeding 20, prompting the initiation of fallback interactions (non-recognition). The second group contained recognition failures where the user's intended command deviated from the system's executed action (misrecognition). While our experiment allowed for measuring non-recognition instances through

game logs, assessing misrecognition instances was challenging. Identifying the player's intended action was not always feasible, thus hindering a complete measurement of this type of recognition failure. While our method might not impact non-recognized intents, incorporating additional environment and action filters holds particular promise in mitigating misrecognition instances. This efficacy can be especially beneficial for non-native English speakers or those with pronounced accents and dialects who face heightened challenges in dealing with misrecognition. The supplementary information these filters provide can guide the system to better predict users' intended actions, thus reducing the likelihood of misrecognition.

6.2 Effect on Usability and Game Experience

We found a significant difference in *Autonomy* in favor of the intervention group. The reason for this could be the higher flexibility of command formulation in the intervention group, as highlighted by our qualitative findings, which showed that players in the intervention group felt less restricted by command variability. This finding could imply that context-aware speech recognition can lead to higher perceived freedom of control and flexibility within the game. Players in the intervention group also rated significantly higher in enjoyment of the game and overall experience. Moreover, our findings indicate higher usability scores for the intervention group. Our customized questionnaire further supports this, as players in the intervention group perceived a significantly lower number of errors, even though the game logs showed that the total number of correctly recognized intents (no filters applied) was significantly higher in the control group. The high usability ratings could suggest that the game was more convenient to play as the system could accurately interpret players' intended commands, reducing the need for the player to repeat their intents. This result supports the notion that it can make the system more usable and engaging for the player, as they can focus on the game itself rather than worrying about recognition errors [63]. Thus, we conclude that context-aware speech recognition can improve the usability and player experience in speech-based games (RQ2).

6.3 Effect on Player Performance

Based on the game logs, we observed no significant difference in playtime, the number of voice commands invoked, and the number of times the filters were used. There was also no difference in the prediction scores as well as the filter scores between the groups. This implies that both groups had similar playing conditions and experienced similar recognition errors and interactions with the environment and game state. The lack of difference in the number of commands and playtime suggests that players from both groups had the same level of performance regardless of the recognition method. Players also observed this as they rated their performance similarly in both conditions in the customized questions. Therefore, we conclude that the context-aware recognition method did not necessarily enhance the players' performance (RQ3).

The interpretation of our results leads us to the following conclusions: data from the game environment and game actions can be used in video games or other virtual environments to assist or enhance voice recognition accuracy and error handling. This method could be further enhanced using deep learning and player models to predict the intended commands better. Context-aware speech recognition can significantly improve the usability of a speech-based video game and enhance the player experience, particularly concerning the degree of autonomy offered by the system and the player's enjoyment of this type of game. While context-aware recognition could improve the accuracy of the recognition system, it does not necessarily improve player performance.

The approach used in our study holds applicability to other speech-based video games. While customizing the *environment* and *action filters* based on a specific game may be necessary, the fundamental principles and techniques introduced in this work provide a groundwork for designing similar systems in diverse gaming contexts. Furthermore, this method can be applied to other areas of human-computer interaction (HCI) beyond just games. The broader insights of this research can be used to enhance speech interaction in other virtual environments, such as virtual reality. In non-virtual settings, a common approach is to use multimodal systems, which utilize data from multiple sources to improve command prediction accuracy. Users' intended commands can be more accurately identified by incorporating different input modalities, such as gesture or gaze, in addition to speech.

6.4 Limitations and Future Work

The findings of this work provide important considerations concerning recognition error handling in speech-based games. However, there are still certain limitations to this study that need to be addressed.

We recruited ($N = 40$) participants for our study. Although this sample size suffices for a first exploration of context-aware speech recognition, future research can validate the results by investigating a wider population. In our experiment, most participants were non-native English speakers. Throughout the sessions, we observed that a few participants (non-native English speakers) with strong accents had more complications with speech recognition, as the system was not trained with data from non-native English speakers. Although this is a common issue with speech-based systems in general, we acknowledge this limitation and encourage future research to address this by incorporating training data from non-native speakers, enhancing the system's ability to recognize diverse accents. Additionally, we had a limited sample of seven native English speakers, preventing us from conducting statistically significant tests between native and non-native English speakers. Future research could delve into our method and explore its impact on both groups individually.

In our game, we used a specific formula to calculate the scores for each filter as well as the final confidence scores. This formula was calculated empirically based on our initial play-testing sessions with the game and the recognition system. Future research can explore further possibilities to set these values to find an optimal formula and enhance recognition. Moreover, other approaches could be considered to enhance the context-aware recognition filters. For instance, in this work, as a first attempt, the environment filter used interactable objects visible in the frame for its scoring to

enhance recognition. Future work can investigate other environmental sources from the game that could be used to enhance this filter, such as the distance between the player and objects in the game environment or the audio cues and sound effects in the game environment. In addition, speech recognition might be further improved by incorporating context information from the environment and action filters as features into the speech recognition system itself, instead of using this information for post-processing, as we have done in this study.

One important critique point highlighted by our participants was the lack of animation while certain actions were being done in the game. This might have impacted the players' experience and engagement with the game.

While playing "Escape the Echo," players could use a limited set of pre-defined voice commands. This enabled us to have a structured procedure with high comparability [42]. However, we recommend future studies to extend the scope of the potential actions and the command vocabulary to examine the scalability of our findings for broader application domains. AI-based dialog systems can be infused with pre-scripted dialog systems to avoid repetitive responses and expand the scope of accepted intents.

In this study, we focused on a specific type of video game by testing our hypotheses with an escape room 3D game. Thus, the study findings may not apply to other types of games, such as 2D games, fast-paced games, or puzzle games that do not use interactable game objects. Further research is needed to observe the applicability of context-aware speech recognition for other types of video games. Moreover, our study was conducted on Windows computers. To explore applicability, future research can replicate this study for other gaming platforms, such as mobile devices and virtual reality games.

We explored the context-aware recognition method in the context of a video game. Although the broader insights of our findings can be applied to other speech-based systems, future work could use this method to evaluate it in other domains, such as smart homes and cars, to explore diverse and more complex settings.

All in all, even though our work contains certain limitations, they do not invalidate the study's implications. We witnessed that the majority of participants were highly enthusiastic about the game and found it interesting and exciting. Considering that only 17.5% of our participants have previously played a voice-controlled game and were mainly not familiar with such a type of game, we believe part of the interest in our game comes from the unconventionality of speech-based video games. The findings of this work highlight that employing speech-based interaction in games can result in high levels of immersion. We highly encourage researchers to explore this specific category of video games further.

7 CONCLUSION

In this work, we examined context-aware speech recognition for a speech-based game. We developed "Escape the Echo," a 3D escape room game where players use speech commands to control the game events. We conducted a between-subjects design study to compare a standard error handling method where the system would compare the recognized output text with the available commands and execute the one with the lowest similarity distance, with the context-aware speech recognition where the game environment and actions were used as supplementary information to enhance recognition. Our results indicated that our proposed method could increase the usability of a system while enhancing the player experience. The findings of this work can contribute valuable insights for researchers and developers on how to enhance speech recognition in speech-based video games and other application domains in the field of voice user interfaces.

ACKNOWLEDGMENTS

This work was partially funded by Klaus Tschira Foundation, the FET-Open Project 951846 "MUHAI – Meaning and Understanding for Human-centric AI" funded by the EU program Horizon 2020, as well as the German Research Foundation DFG as part of Collaborative Research Center (Sonderforschungsbereich) 1320 "EASE – Everyday Activity Science and Engineering", University of Bremen (http://www.ease-crc.org/).

REFERENCES

[1] Beena Ahmed, Penelope Monroe, Adam Hair, Chek Tien Tan, Ricardo Gutierrez-Osuna, and Kirrie J Ballard. 2018. Speech-driven mobile games for speech therapy: User experiences and feasibility. *International journal of speech-language pathology* 20, 6 (2018), 644–658. https://doi.org/10.1080/17549507.2018.1513562 arXiv:https://doi.org/10.1080/17549507.2018.1513562 PMID: 30301384.

[2] Fraser Allison, Marcus Carter, and Martin Gibbs. 2017. Word Play: A History of Voice Interaction in Digital Games. *Games and Culture* 15, 2 (2017), 91 – 113. https://doi.org/10.1177/1555412017746305

[3] Fraser Allison, Marcus Carter, Martin Gibbs, and Wally Smith. 2018. Design Patterns for Voice Interaction in Games. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play* (Melbourne, VIC, Australia) *(CHI PLAY '18)*. Association for Computing Machinery, New York, NY, USA, 5–17. https://doi.org/10.1145/3242671.3242712

[4] Fraser Allison, Joshua Newn, Wally Smith, Marcus Carter, and Martin Gibbs. 2019. Frame Analysis of Voice Interaction Gameplay. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300623

[5] Fraser John Allison. 2020. *Voice interaction game design and gameplay*. Ph. D. Dissertation. University of Melbourne. http://hdl.handle.net/11343/240857

[6] Tawfiq Ammari, Jofish Kaye, Janice Y. Tsai, and Frank Bentley. 2019. Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Transactions on Computer-Human Interaction* 26, 3, Article 17 (apr 2019), 28 pages. https://doi.org/10.1145/3311956

[7] M. A. Anusuya and S. K. Katti. 2010. Speech Recognition by Machine, A Review. arXiv:1001.2267 [cs.CL]

[8] Saki Anzai, Tokio Ogawa, and Junichi Hoshino. 2021. Speech Recognition Game Interface to Increase Intimacy with Characters. In *Entertainment Computing – ICEC 2021*, Jannicke Baalsrud Hauge, Jorge C. S. Cardoso, Licínio Roque, and Pedro A. Gonzalez-Calero (Eds.). Springer International Publishing, Cham, 167–180. https://doi.org/10.1007/978-3-030-89394-1_13

[9] Maresa Biermann, Evelyn Schweiger, and Martin Jentsch. 2019. Talking to Stupid?!? Improving Voice User Interfaces. In *Mensch und Computer 2019 - Usability Professionals*, Holger Fischer and Steffen Hess (Eds.). Gesellschaft für Informatik e.V. Und German UPA e.V., Bonn. https://doi.org/10.18420/muc2019-up-0253

[10] Michael Bonfert, Nima Zargham, Florian Saade, Robert Porzel, and Rainer Malaka. 2021. An Evaluation of Visual Embodiment for Voice Assistants on Smart Displays. In *Proceedings of the 3rd Conference on Conversational User Interfaces* (Bilbao (online), Spain) *(CUI '21)*. Association for Computing Machinery, New York, NY, USA, Article 16, 11 pages. https://doi.org/10.1145/3469595.3469611

[11] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597. https://doi.org/10.1080/2159676X.2019.1628806 arXiv:https://doi.org/10.1080/2159676X.2019.1628806

[12] Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. 2019. *Thematic Analysis*. Springer Singapore, Singapore, 843–860. https://doi.org/10.1007/978-981-10-5251-4_103

[13] John Brooke. 2013. SUS: a retrospective. *Journal of usability studies* 8, 2 (feb 2013), 29–40. https://dl.acm.org/doi/abs/10.5555/2817912.2817913

[14] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[15] Marcus Carter, Fraser Allison, John Downs, and Martin Gibbs. 2015. Player Identity Dissonance and Voice Interaction in Games. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play* (London, United Kingdom) *(CHI PLAY '15)*. Association for Computing Machinery, New York, NY, USA, 265–269. https://doi.org/10.1145/2793107.2793144

[16] Yanto Chandra and Liang Shang. 2019. *Inductive Coding*. Springer Nature Singapore, Singapore, 91–106. https://doi.org/10.1007/978-981-13-3170-1_8

[17] Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Erlbaum, Hillsdale, NJ.

[18] Lynne M Connelly and Jill N Peltzer. 2016. Underdeveloped themes in qualitative research: Relationship with interviews and analysis. *Clinical nurse specialist* 30, 1 (2016), 52–57.

[19] Mihaly Csikszentmihalyi. 1990. *Flow: The psychology of optimal experience*. Vol. 1990. Harper & Row, New York, NY, USA.

[20] Steven Dow, Manish Mehta, Ellie Harmon, Blair MacIntyre, and Michael Mateas. 2007. Presence and engagement in an interactive drama. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 1475–1484. https://doi.org/10.1145/1240624.1240847

[21] Marta Filimon, Adrian Iftene, and Diana Trandabăţ. 2019. Bob - A General Culture Game with Voice Interaction. *Procedia Computer Science* 159 (2019), 323–332. https://doi.org/10.1016/j.procs.2019.09.187 Knowledge-Based and Intelligent Information and Engineering Systems: Proceedings of the 23rd International Conference KES2019.

[22] Reinhold Haeb-Umbach, Shinji Watanabe, Tomohiro Nakatani, Michiel Bacchiani, Bjorn Hoffmeister, Michael L Seltzer, Heiga Zen, and Mehrez Souden. 2019. Speech processing for digital home assistants: Combining signal processing with deep-learning techniques. *IEEE Signal processing magazine* 36, 6 (2019), 111–124.

[23] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: Scaling up end-to-end speech recognition.

[24] Susumu Harada, Jacob O. Wobbrock, and James A. Landay. 2011. Voice Games: Investigation Into the Use of Non-speech Voice Input for Making Computer Games More Accessible. In *Human-Computer Interaction – INTERACT 2011*, Pedro Campos, Nicholas Graham, Joaquim Jorge, Nuno Nunes, Philippe Palanque, and Marco Winckler (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 11–29. https://doi.org/10.1007/978-3-642-23774-4_4

[25] Ramin Hedeshy, Chandan Kumar, Mike Lauer, and Steffen Staab. 2022. All Birds Must Fly: The Experience of Multimodal Hands-Free Gaming with Gaze and Nonverbal Voice Synchronization. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (Bengaluru, India) *(ICMI '22)*. Association for Computing Machinery, New York, NY, USA, 278–287. https://doi.org/10.1145/3536221.3556593

[26] Kieran Hicks, Kathrin Gerling, Patrick Dickinson, Conor Linehan, and Carl Gowen. 2018. Leveraging Icebreaking Tasks to Facilitate Uptake of Voice Communication in Multiplayer Games. In *Advances in Computer Entertainment Technology*, Adrian David Cheok, Masahiko Inami, and Teresa Romão (Eds.). Springer International Publishing, Cham, 187–201.

[27] Minki Hong, YoungJun Choi, and Sihun Cha. 2021. "Anyway,": Two-Player Defense Game via Voice Conversation. In *Extended Abstracts of the 2021 Annual Symposium on Computer-Human Interaction in Play* (Virtual Event, Austria) *(CHI PLAY '21)*. Association for Computing Machinery, New York, NY, USA, 345–349. https://doi.org/10.1145/3450337.3483509

[28] Hyunhoon Jung, Hee Jae Kim, Seongeun So, Jinjoong Kim, and Changhoon Oh. 2019. TurtleTalk: an educational programming game for children with voice user interface. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–6.

[29] Jesper Juul. 2007. Without a goal: on open and expressive games. In *Videogame, player, text*, Barry Atkins and Tanya Krzywinska (Eds.). Manchester University Press Manchester, England, 191–203. http://www.jesperjuul.net/text/withoutagoal/

[30] Kwan Min Lee, Wei Peng, Seung-A Jin, and Chang Yan. 2006. Can Robots Manifest Personality?: An Empirical Test of Personality Recognition, Social Responses, and Social Presence in Human–Robot Interaction. *Journal of Communication* 56, 4 (11 2006), 754–772. https://doi.org/10.1111/j.1460-2466.2006.00318.x arXiv:https://academic.oup.com/joc/article-pdf/56/4/754/22325856/jjnlcom0754.pdf

[31] Vladimir I Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. In *Soviet Physics Doklady*, Vol. 10. 707. https://api.semanticscholar.org/CorpusID:60827152

[32] Toby Jia-Jun Li, Igor Labutov, Brad A Myers, Amos Azaria, Alexander I Rudnicky, and Tom M Mitchell. 2018. An end user development approach for failure handling in goal-oriented conversational agents. In *Studies in Conversational UX Design*, Robert J. Moore, Margaret H. Szymanski, Raphael Arar, and Guang-Jie Ren (Eds.). Springer, Berlin, Germany.

[33] Xiaoliang Ma, Congjian Deng, Dequan Du, and Qingqi Pei. 2023. An enhanced method for dialect transcription via error-correcting thesaurus. *IET Communications* 17, 17 (2023), 1984–1997. https://doi.org/10.1049/cmu2.12671 arXiv:https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/cmu2.12671

[34] Lina Mavrina, Jessica Szczuka, Clara Strathmann, Lisa Michelle Bohnenkamp, Nicole Krämer, and Stefan Kopp. 2022. "Alexa, You're Really Stupid": A Longitudinal Field Study on Communication Breakdowns Between Family Members and a Voice Assistant. *Frontiers in Computer Science* 4 (2022), 791704. https://doi.org/10.3389/fcomp.2022.791704

[35] Hani Morgan. 2022. Understanding thematic analysis and the debates involving its use. *The Qualitative Report* 27, 10 (2022), 2079–2090. https://doi.org/10.46743/2160-3715/2022.5912

[36] Christine Murad, Cosmin Munteanu, Benjamin R Cowan, and Leigh Clark. 2019. Revolution or Evolution? Speech Interaction and HCI Design Guidelines. *IEEE Pervasive Computing* 18, 2 (2019), 33–45. https://doi.org/10.1109/MPRV.2019.2906991

[37] Moyen Mohammad Mustaquim. 2013. Automatic speech recognition-an approach for designing inclusive games. *Multimedia tools and applications* 66, 1 (2013), 131–146. https://doi.org/10.1007/s11042-011-0918-7

[38] Ali Bou Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, and Khaled Shaalan. 2019. Speech recognition using deep neural networks: A systematic review. *IEEE access* 7 (2019), 19143–19165. https://doi.org/10.1109/ACCESS.2019.2896880

[39] Andrés Navarro Newball, Diego Loaiza, Claudia Oviedo, Andrés Castillo-Saavedra, A Portilla, Diego Linares, and Gloria Alvarez. 2014. Talking to Teo: Video game supported speech therapy. *Entertainment Computing* 5, 4 (2014), 401–412. https://doi.org/10.1016/j.entcom.2014.10.005

[40] Hunter Osking and John A Doucette. 2019. Enhancing emotional effectiveness of virtual-reality experiences with voice control interfaces. , 199–209 pages.

[41] Tony Di Petta and Vera E Woloshyn. 2001. Voice recognition for on-line literacy: Continuous voice recognition technology in adult literacy training. *Education and Information Technologies* 6, 4 (2001), 225–240.

[42] Robert Porzel and Manja Baudis. 2004. The Tao of CHI: Towards Effective Human-Computer Interaction. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*. Association for Computational Linguistics, Boston, Massachusetts, USA, 209–216. https://www.aclweb.org/anthology/N04-1027

[43] 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* (Denver, Colorado, USA) *(CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 2853–2859. https://doi.org/10.1145/3027063.3053246

[44] Aung Pyae and Paul Scifleet. 2018. Investigating differences between native english and non-native english speakers in interacting with a voice user interface: a case of google home. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction* (Melbourne, Australia) *(OzCHI '18)*. Association for Computing Machinery, New York, NY, USA, 548–553. https://doi.org/10.1145/3292147.3292236

[45] K. Radzikowski, R. Nowak, Le Wang, and O. Yoshie. 2019. Dual supervised learning for non-native speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing* 2019 (2019), 1–10. https://doi.org/10.1186/s13636-018-0146-4

[46] D.Raj Reddy, Lee Erman, and Richard Neely. 1973. A model and a system for machine recognition of speech. *IEEE Transactions on Audio and Electroacoustics* 21, 3 (07 1973), 229–238. https://doi.org/10.1109/TAU.1973.1162456

[47] Gavin Reid. 2012. Motivation in video games: a literature review. *The computer games journal* 1, 2 (2012), 70–81.

[48] Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Ye Jia, Pedro Moreno, Yonghui Wu, and Zelin Wu. 2019. Speech recognition with augmented synthesized speech. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*. IEEE, New York, NY, USA, 996–1002. https://doi.org/10.1109/ASRU46091.2019.9003990

[49] Richard M Ryan, C Scott Rigby, and Andrew Przybylski. 2006. The motivational pull of video games: A self-determination theory approach. *Motivation and emotion* 30, 4 (2006), 344–360.

[50] Jeff Sauro and James R Lewis. 2016. Quantifying the user experience: Practical statistics for user research.

[51] Katie Seaborn, Norihisa P. Miyake, Peter Pennefather, and Mihoko Otake-Matsuura. 2021. Voice in Human–Agent Interaction: A Survey. *ACM Comput. Surv.* 54, 4, Article 81 (may 2021), 43 pages. https://doi.org/10.1145/3386867

[52] S. S. Shapiro and M. B. Wilk. 1965. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* 52, 3/4 (1965), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

[53] Aaron Springer and Henriette Cramer. 2018. "Play PRBLMS": Identifying and Correcting Less Accessible Content in Voice Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA,

1–13. https://doi.org/10.1145/3173574.3173870

[54] Amanda J Stent, Marie K Huffman, and Susan E Brennan. 2008. Adapting speaking after evidence of misrecognition: Local and global hyperarticulation. *Speech Communication* 50, 3 (2008), 163–178. https://doi.org/10.1016/j.specom.2007.07.005

[55] Bernhard Suhm, Brad Myers, and Alex Waibel. 2001. Multimodal error correction for speech user interfaces. *ACM transactions on computer-human interaction (TOCHI)* 8, 1 (mar 2001), 60–98. https://doi.org/10.1145/371127.371166

[56] David R Thomas. 2006. A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation* 27, 2 (2006), 237–246.

[57] Asma Trabelsi, Sébastien Warichet, Yassine Aajaoun, and Séverine Soussilane. 2022. Evaluation of the efficiency of state-of-the-art Speech Recognition engines. *Procedia Computer Science* 207 (2022), 2242–2252. https://doi.org/10.1016/j.procs.2022.09.534 Knowledge-Based and Intelligent Information and Engineering Systems: Proceedings of the 26th International Conference KES2022.

[58] Jan van der Kamp and Veronica Sundstedt. 2011. Gaze and Voice Controlled Drawing. In *Proceedings of the 1st Conference on Novel Gaze-Controlled Applications* (Karlskrona, Sweden) *(NGCA '11)*. Association for Computing Machinery, New York, NY, USA, Article 9, 8 pages. https://doi.org/10.1145/1983302.1983311

[59] Jing Wei, Benjamin Tag, Johanne R Trippas, Tilman Dingler, and Vassilis Kostakos. 2022. What Could Possibly Go Wrong When Interacting with Proactive Smart Speakers? A Case Study Using an ESM Application. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 276, 15 pages. https://doi.org/10.1145/3491102.3517432

[60] Tom Wilcox, Mike Evans, Chris Pearce, Nick Pollard, and Veronica Sundstedt. 2008. Gaze and Voice Based Game Interaction: The Revenge of the Killer Penguins. In *ACM SIGGRAPH 2008 Posters* (Los Angeles, California) *(SIGGRAPH '08)*. Association for Computing Machinery, New York, NY, USA, Article 81, 1 pages. https://doi.org/10.1145/1400885.1400972

[61] Chauncey Wilson. 2013. *Interview techniques for UX practitioners: A user-centered design method.* Elsevier, Waltham, Massachusetts, USA.

[62] Nima Zargham, Michael Bonfert, Georg Volkmar, Robert Porzel, and Rainer Malaka. 2020. Smells Like Team Spirit: Investigating the Player Experience with Multiple Interlocutors in a VR Game. In *Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play* (Virtual Event, Canada) *(CHI PLAY '20)*. Association for Computing Machinery, New York, NY, USA, 408–412. https://doi.org/10.1145/3383668.3419884

[63] Nima Zargham, Johannes Pfau, Tobias Schnackenberg, and Rainer Malaka. 2022. "I Didn't Catch That, But I'll Try My Best": Anticipatory Error Handling in a Voice Controlled Game. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 153, 13 pages. https://doi.org/10.1145/3491102.3502115

[64] Andrej Zgank and Zdravko Kacic. 2012. Predicting the acoustic confusability between words for a speech recognition system using Levenshtein distance. *Elektronika ir Elektrotechnika* 18, 8 (Oct. 2012), 81–84. https://doi.org/10.5755/j01.eee.18.8.2628

[65] Rui Zhao, Kang Wang, Rahul Divekar, Robert Rouhani, Hui Su, and Qiang Ji. 2018. An immersive system with multi-modal human-computer interaction. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, IEEE, New York, NY, USA, 517–524. https://doi.org/10.1109/FG.2018.00083

[66] Bartosz Ziółko, Jakub Gałka, T Jadczyk, and D Skurzok. 2010. Modified weighted Levenshtein distance in automatic speech recognition. In *Proceedings of the XVI National Conference Applications of Mathematics to Biology and Medicine*. Citeseer, 116–120. https://api.semanticscholar.org/CorpusID:15188678

A APPENDIX

Table 4: Test of Normality (Shapiro-Wilk) for SUS and PENS questionnaires. Significant results suggest a deviation from normality.

	Condition	W	p-value
SUS	Intervention	0.929	0.145
	Control	0.904	0.059
PENS - Intuitive Controls	Intervention	0.870	0.012 *
	Control	0.824	0.002 *
PENS - Presence	Intervention	0.925	0.124
	Control	0.980	0.932
PENS - Autonomy	Intervention	0.944	0.281
	Control	0.954	0.431
PENS - Competence	Intervention	0.906	0.054
	Control	0.880	0.018 *

Table 5: Test of Normality (Shapiro-Wilk) for the customized questions. Significant results suggest a deviation from normality.

	Condition	W	p-value
Game Experience	Intervention	0.760	< .001 *
	Control	0.877	0.015 *
Game Enjoyment	Intervention	0.829	0.002 *
	Control	0.930	0.153
Perceived Error Rates	Intervention	0.852	0.006 *
	Control	0.949	0.359
Willingness to Play Similar Games	Intervention	0.661	< .001 *
	Control	0.742	< .001 *
Performance Rating	Intervention	0.895	0.033 *
	Control	0.891	0.028 *

Table 6: Test of Normality (Shapiro-Wilk) for the game logs. Significant results suggest a deviation from normality.

	Condition	W	p-value
Play Time	Intervention	0.875	0.032 *
	Control	0.751	< .001 *
Total Inputs	Intervention	0.754	< .001 *
	Control	0.895	0.068
Direct Recognition	Intervention	0.933	0.276
	Control	0.971	0.866
Commands Predicted	Intervention	0.884	0.045 *
	Control	0.972	0.881
Average Confidence Score	Intervention	0.895	0.033 *
	Control	0.891	0.028 *
Environment Filter Score	Intervention	0.946	0.437
	Control	0.948	0.470
Actions Filter Score	Intervention	0.921	0.179
	Control	0.94	0.348
Scope Filter Score	Intervention	0.935	0.292
	Control	0.988	0.997
Same Outcome With Only Scope	Intervention	0.918	0.158
	Control	0.968	0.809

Publication 2

"I Didn't Catch That, But I'll Try My Best": Anticipatory Error Handling in a Voice Controlled Game

Nima Zargham, Johannes Pfau, Tobias Schnackenberg, and Rainer Malaka

In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22). New York, NY, USA, 2022. Association for Computing Machinery.

Personal contribution to this work: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, part of software, supervision, validation, visualization, and contribution to all parts of the manuscript.

ISBN: 978-1-4503-9157-3/22/04 DOI: 10.1145/3491102.3502115

“I Didn’t Catch That, But I’ll Try My Best”: Anticipatory Error Handling in a Voice Controlled Game

Nima Zargham
zargham@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

Johannes Pfau
jpfau@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

Tobias Schnackenberg
tschnack@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

Rainer Malaka
malaka@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

ABSTRACT

Advances in speech recognition, language processing and natural interaction have led to an increased industrial and academic interest. While the robustness and usability of such systems are steadily increasing, speech-based systems are still susceptible to recognition errors. This makes intelligent error handling of utmost importance for the success of those systems. In this work, we integrated anticipatory error handling for a voice-controlled video game where the game would perform a locally optimized action in respect to goal completion and obstacle avoidance, when a command is not recognized. We evaluated the user experience of our approach versus traditional, repetition-based error handling ($N = 34$). Our results indicate that implementing anticipatory error handling can improve the usability of a system, if it follows the intention of the user. Otherwise, it impairs the user experience, even when deciding for technically optimal decisions.

CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**;
• **Applied computing** → **Computer games**.

KEYWORDS

Voice User Interfaces, Game Design, Error Handling, Speech-Based Systems, Voice-Controlled Game, Voice Interaction

ACM Reference Format:
Nima Zargham, Johannes Pfau, Tobias Schnackenberg, and Rainer Malaka. 2022. “I Didn’t Catch That, But I’ll Try My Best”: Anticipatory Error Handling in a Voice Controlled Game. 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, 13 pages. https://doi.org/10.1145/3491102.3502115

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/10.1145/3491102.3502115

1 INTRODUCTION

Voice user interfaces (VUIs) are gaining more and more attention in recent years due to the intuitive nature of their interaction. Speaking is a natural way of communication amongst humans and people find it easier to interact with technology that resembles some of their own characteristics [15]. Voice input is now a feature in many devices such as mobile phones, cars and home assistants. In their early days, VUIs were designed for handling few specialized tasks [55], but due to the advancements in the technology, they now can have a broad range of capabilities in performing various functions in different settings. Current VUIs are used for various purposes such as smart home control, scheduling, navigation, education, and entertainment. The technical aspects of the VUIs, as well as their usability and user experience (UX), have been the subject of extensive research in the recent years [21, 23, 25, 43, 44].

In order to integrate speech recognition, developers need to have a large repository of collected voice data so that the system has enough information to process different inflections and variations in different voices. If the product is aimed at the global market, different languages, accents, and dialects need to be considered to assure a better recognition system. On top of that, different forms of phrasing for a single command should be incorporated to allow for a more natural experience, underlining the issue that designing a satisfying experience with speech-based systems is a complex and difficult process.

Although this technology is steadily improving in various aspects, speech-based systems are still prone to recognition failures. Several elements such as hardware limitations, background noises and language barriers make designing voice interfaces a very complex and time-consuming task. Researchers believe that problems with speech recognition and limited functionality are the main reasons for disliking or not using voice systems [11]. Users have frequently reported that they find voice interaction disappointing or embarrassing, which lets such systems appear as unintelligent and immature [5, 11, 18, 42, 48]. This makes error handling a critical part of designing VUIs, which includes the situations where the system does not understand the user’s command, the given command is out of context, or the command is misunderstood [39]. Several guidelines for designing fallback strategies have been proposed,

such as asking the user to repeat the command, redirecting the user to the tasks that the system can support, or presenting user options to correct their commands [12, 39, 51]. In some cases, the voice assistant (VA) falls back on humor in response to complex conversational input and commands that cannot be handled otherwise, which might be seen as sarcastic or entertaining [56].

Recently, this technology has gained considerable attention in the entertainment industry and video game companies have been adopting voice-activated services to their games. As speech recognition technology is improving rapidly and the number of available microphones in consumer gaming devices is growing every day, it leaves a great potential for using VUIs in games [2]. This allows voice-control to be used as an appealing and intuitive feature in video games to enhance the experience of the players. Speaking is a natural and enjoyable way of interacting, which can increase social presence within the game and make them more immersive [38, 76]. With the release of Microsoft Kinect in 2010, Xbox games in various genres such as *Mass Effect 3* [31], *FIFA 14* [30], *Forza Motorsport 5* [67], and *Ryse: Son of Rome* [27] took advantage of the voice interaction that was provided by Kinect. However, in most cases, voice input is an optional feature and not a core element of the game design.

Voice-activated games attempted to provide natural language input, but this experience has been frequently described as "uncomfortable" and "awkward" by players [29]. Video games are mainly goal-oriented activities, and players find enjoyment when they work towards this goal [35]. If the challenge is right, the players are in a state of flow [28]. The misrecognition of voice input in video games adds another layer of challenge on top of the game's existing obstacles, preventing players from reaching their goals and staying in the state of flow, which often results in player frustration. Moreover, studies have shown that once a recognition error occurs, the likelihood of having an error in the next intent increases [13, 60, 66]. One of the reasons for this is that, as more errors occur, user's patience runs out and frustration increases, which can lead to acoustic and language mismatches [13]. Previous research has shown that human operators often do not signal non-understandings, but rather try to advance the task by asking different questions, which generally led to a speedier recovery [63]. Similarly, for speech-based systems, researchers suggest that when non-understandings happen, instead of trying to repair the current problem, use an alternative dialog plan to advance the task [13].

On this basis, we designed a voice-controlled video game with the aim of investigating user experience with two different error handling methodologies. In this game, players control the game protagonist using voice commands. A between-subjects user study was conducted to compare traditional repetition-based error handling with a novel approach implementing anticipatory error handling within the game. In the control group, the game would notify the player of the recognition failure so that the player could repeat the command once again. With anticipatory error handling, if a command was not recognized, the game would proceed by performing a locally optimized action in respect to goal completion and obstacle avoidance without notifying the player about the recognition failure. In the scope of this work, when we refer to recognition failure, our interest lies in command recognition, which is a subset of natural language understanding (NLU). Nonetheless, the insights of this work might also hold for certain NLU issues.

In this study, we pursue the following research questions:

RQ1 Does performing a locally optimized game action in times of misrecognition lead to a measurably improved usability in a speech-based video game?

RQ2 What are the effects on player experience in terms of competency, autonomy, presence, and intuitive control, if error handling mechanisms decide for unintended actions?

Based on our design space and the existing literature, we developed the following hypotheses:

- *H1: Participants will observe a lower number of recognition errors in case of anticipatory error handling.*
- *H2: The anticipatory error handling will lead to a higher rating regarding:*
  (a) *players' perceived competence.*
  (b) *players' perceived autonomy.*
  (c) *players' perceived presence.*
  (d) *intuitiveness of the game controls.*

Our results showed significantly higher usability ratings for the anticipatory error handling, as well as a significantly lower number of perceived errors for this condition. Furthermore, this study contributes useful insights and implications on the user experience with recognition error handling in speech-based systems, most importantly the users' aversion to error handling that opposed their intention – even in cases of goal-directed and anticipated solutions.

2 RELATED WORK

Since the early success of voice and gesture in an interface with the "Put that There" system [14], voice user interfaces have been largely investigated by researchers in the field of HCI. In this section, we provide a summary of the previous literature on speech-based systems, voice interaction in video games, and complications with VUIs.

2.1 Research on Speech-Based Systems

Developing speech-based systems requires techniques, methodologies, and development tools that are capable of flexible and adaptive interaction, bearing in mind the need of different user groups and different environments [68]. In recent years, natural language processing (NLP) has become much more sophisticated and reliable [19]. Apart from technical development, interaction research tackled multitudes of novel voice interfaces, investigating how people use these devices and how they respond to different kinds of speech from computers [5, 9, 23, 41].

Speech-based systems have been evaluated for various purposes and professional fields. In the medical domain, Austerjost et al. presented a VUI for controlling laboratory instruments [7], while Miehle et al. presented a concept for voice assistants (VAs) as a support in surgical operating rooms [46]. For the purpose of teaching, Jung et al. [34] developed a voice-controlled educational game to teach children computer programming, concluding that their game led children to be more immersed in the game and understand the elements of programming with ease and confidence. Winkler et al. [73] compared groups who either used a human or a VA tutor when solving a problem. Their results indicated that groups

interacting with VA showed significantly higher task outcomes and higher degrees of collaboration quality compared to groups interacting with human tutors. Another prominent application area resides in entertainment. Zargham and Bonfert et al. [75] investigated voice interaction in a single-player VR game where they compared a version of the game in which the players could talk to multiple characters using natural language to a version where they verbally interact with a single character. The study showed that the participants preferred conversing with a group of interlocutors, found it more entertaining, and felt like being part of a team.

Although the functionality and ease of use of VUIs are frequently researched and enhanced, research suggests that the reliability of these systems is not more important than their attractiveness [74]. In a study by Lopatovska et al. [40], the authors explored user interactions with the popular VA Amazon Alexa. They report that people were still satisfied with the system even when Alexa did not produce desirable outcomes. Authors suggest that the UX might be more important to the users than the quality of the output.

One particular challenge with VUIs is that it can lead to unrealistic expectations from the system's intelligence, what it can do, and how well it can keep a natural and fluid conversation [43]. Users tend to test the capabilities of VUIs by asking different questions and in many cases, their expectations tend to exceed the agent's capabilities [41, 42]. This also applies to children. In a study by Lovato et al. on children's experience with Siri, authors found that children predominantly ask Siri personal questions, to get to know the agent and test its potential [41]. When users' initial expectations from such systems are not met, it can lead to disappointment and a generally negative experience [55].

Overall, a great deal of the design research is focused on narrow application areas and specific interface components. This in turn leads to the lack of more generalizable design guidelines [22]. In our work, we seek to advance the state of the art by exploring methodologies of recognition error handling.

2.2 Voice Interaction in Video Games

The intuitive nature of voice user interfaces allowed them to become an increasing trend, not only in assisting function within smart homes, phones or cars, but also for the advancement of mechanics within the entertainment industry. Although the rate of VUI studies has increased in recent years, research on voice interaction in games – those where voice control has a fundamental role in the game – is rather limited [20]. Using alternative means of interaction for games such as voice can not only expand the possibility space for novel in-game mechanics, but can also be especially important for users with disabilities, where traditional controls are not feasible [71]. Speech-controlled video games have also proven effective in enhancing speech therapy and facilitating remote treatment [1]. Other human modalities can be combined with speech to optimize players' performance and overcome the drawbacks of using only speech [50]. Nonetheless, there are still essential aspects and questions regarding voice interaction in games that have gone largely unexplored [5].

Voice interaction in video games is rather distinct from the other contexts. Research shows that in-game voice commands are associated with a sense of taking on a character in the game's world [4]. Allison et al. suggest that voice interactions which creates a conflict with the social world can impede the player's engagement with the in-game world [5]. Early research on voice interaction in digital games roots back to the 1970s, where *VoiceChess*, a game which could support standardized chess instructions using a speech recognition system, was developed [2, 59]. Since then, numerous video game titles have embraced the use of voice as input. In a successive study by Allison et al., the authors surveyed 449 video games and 22 audio games in which players use their voice to affect the game state [3]. They observed that academic research has focused on a narrow subset of design patterns, especially pronunciation, and recommend game designers to consider non-verbal forms, which have proven to provide enjoyable game experiences with fast and discrete input possibilities [32, 52, 64, 70].

Although there are plenty of examples of video games that use speech-based voice interaction, those which use non-verbal forms of voice input have been more successful. The reason behind the success of such games is that they avoid recognition errors entirely [3, 4]. However, due to the limited controls, these games are usually restricted to relatively simple mechanics. In this work, we simulate an environment that enables fast and reliable calculation of technically optimized actions so that gaps in recognition can be handled and the resulting experience investigated. To the best of our knowledge, no previous video game has used a similar technique, thus our approach of an anticipatory error handling method is original.

2.3 Complications with Voice Interaction

A large portion of research about voice interaction is concerned with speech recognition and its accuracy rates [3]. These systems are commonly trained with a large sample of voice data, connected with ontologies and knowledge graphs, in order to identify and understand users' commands and respond with a reasonable and satisfying answer [39]. Nevertheless, the given commands by the users can be fuzzy, personal, and complicated, resulting in the system not being able to understand them, which often leads to user frustration, disappointment, and dissatisfaction [10, 26, 42]. These issues are not likely to be overcome by soft- or hardware advancements in recognition alone. To conquer the difficulties inherent in processing the commands, users usually need to put more effort in formulating the command so that it is recognized by the system.

When interacting with a VUI, users typically speak differently than they would speak to a human. Many expect natural language not to be understood by such systems and adapt special communication strategies therefore. Reducing the talking pace, re-formulating command sentences and physically relocating themselves and/or the system are popular observable patterns when users are confronted with recognition errors [11]. Jentsch et al. observed that users took a considerable amount of time to formulate their prompts before commanding them to a VUI [11]. In their study, authors also witnessed that even when the users are not instructed to use keywords, they are still likely to restrict themselves to a set of words or commands when addressing a speech assistant. This has led users to refrain from speech-based systems to perform difficult tasks. In a study by Luger et al. [42], authors interviewed frequent users of conversational agents and found that the study participants did

not trust the system to do complex tasks – like writing emails or making phone calls – down to an apprehension that the system would not get the task done correctly. Authors also note that the interaction with the agent was generally considered as a secondary task.

On the other hand, when errors occur, the system should give an appropriate response. In her book about designing VUIs, Cathy Pearl suggests that, if the error handling is done well, it will not derail users, and you can get them back on the track and have them successfully complete a task [53]. If it's done poorly, not only the user will fail to complete a task, but they actually might refuse to use the system again. A study by Suhm et al. explored multimodal error correction methods that allows the user to correct the recognition errors in speech user interfaces [65]. The authors found that although users preferred speech as an input modality, if the accuracy of recognition was low, they learned to avoid it with experience. Vertanen et al. explored different techniques such as silence filtering to improve the recognition of spoken corrections when a system fails to recognize the command in the first try [69]. Their study showed that by combining multiple techniques, the percentage of correctly recognized spoken corrections increased by more than 30%. Bohus et al. subdivide speech recognition errors into two types of misunderstandings and non-understandings [13]. Misunderstandings are referred to those cases in which the system misinterprets the user's input, where in non-understanding events, the system fails to obtain any interpretation. In their study, the authors looked at ten non-understanding recovery strategies and compared their performance. Their results showed that advancing the conversation by ignoring the non-understanding and trying an alternative dialog plan performed best [13].

Although the technical aspects of VUIs have been largely investigated, researchers agree on the stance that the user side of speech interaction is relatively less explored [8, 22, 47, 49]. Above that, language barriers pose a further common problem with VUIs. A study by Pyae et al. showed that VUIs are easier to use, friendlier and potentially more useful for native English speakers than non-native speakers [58]. The complex and expensive process of implementing a reliable speech-based system, impels researchers in this field to often use a Wizard of Oz approach [36, 45].

Eventually, technical limits, unnatural assumptions, and lack of faith in the system's technological capabilities still make up the major reasons for users' reservations against using VUIs. To build upon the prior work regarding the error handling of speech systems, we came up with an approach to avoid unrecognized commands as well as repeating the command in order to correct it which could result in user frustration and ultimately abandoning the system entirely [53, 65]. In our approach, we focus on overcoming innate technical limits of speech recognition with anticipatory error handling and examine the impact of this intervention on the perceived intelligence, appraisal and usability of the system.

3 PROTOTYPE DESIGN

To evaluate our hypotheses, we designed and implemented "Listen, Sparky!", a speech-controlled arcade game. In this game, players are in control of the sheepdog "Sparky" who has to guide a sheep through restricted courses and keep away hazardous encounters. Using speech-controlled commands, players impersonate a shepherd that gives directions to his sheepdog. The game consists of eight levels. In every level, players have to safely navigate and return the sheep that escaped from a meadow, up to a designated goal location (gate).

The first four levels of the prototype served as a tutorial. In these, players were taught about the game controls and the commands to use. Every level would introduce one new command to the players, with the exception of the fourth level that would introduce two commands. The participants were able to access an overview of the available commands at any time in the game menu (see Figure 2). After going through the first two levels, a hostile wolf character was introduced that threatened the survival of the escorted sheep. If the sheep would get too close to the wolf, the level failed and had to be restarted. With increasing progression of the levels, the challenge of the game would similarly increase (see Figure 1). For instance, in the early levels, the wolf is standing still and does not move and the player has to simply avoid those areas of the game. In higher levels, the wolf would start moving or even chase the sheep to make the game more demanding for the player and enforce quick acting. At the end of each level, the game would display a screen indicating that the level was successfully completed while presenting performance feedback throughout a classic star rating system. This rating was given based on the number of commands used in that level and the time taken to finish it.

Level	*Is there a wolf?*	*Is the wolf moving?*	*Is the wolf chasing the sheep?*
1st	No	--	--
2nd	No	--	--
3rd	Yes	No	--
4th	Yes	No	--
5th	Yes	Yes	No
6th	Yes	Yes	No
7th	Yes	Yes	No
8th	Yes	Yes	Yes

Figure 1: The increasing complexity of the levels with the players' progression.

In order to start the speech recognition and have Sparky listen to the commands, players had to press and hold the spacebar. As long as the space bar was pressed, the default computer microphone was used to record the players' voice. If the space bar was released too fast, the system would not process that command. While holding the space bar, the player's voice input was recorded, processed and (if possible) interpreted as one of the following actions:

- "Walk towards": Sparky walks straight towards the sheep, navigating the sheep to the same direction.
- "Flank Left": Sparky flanks the sheep from the left side, navigating the sheep to the right side (relative to the fixed view angle of the participant).

Figure 2: Voice commands making up the core game controls, assessable anytime during gameplay.

- "Flank Right": Sparky flanks the sheep from the right side, navigating the sheep to the left side.
- "Back": Sparky goes back to the position where it began the level.
- "Bark at wolf": Sparky moves towards the wolf and barks. This results in paralyzing the wolf for some seconds and making it harmless to the sheep.

The system was able to handle multiple phrases per action. For instance, if players wanted to command Sparky to "flank right", they could also use phrases such as "go right!", "right side" or "move right". If a command was recognized by the voice recognition system, Sparky would execute the corresponding command. If no matching command was found, the system would consider that as a failed attempt. In such cases, the game would refer to the error handling system based on the respective experimental group. For every participant, the system recorded the error rates, which was the number of commands that were not recognized by the system throughout the session. In order to evaluate different error handling methods, we needed to ensure noticeable instances of recognition failure. To achieve this, both game versions were programmed to have a minimum overall error occurrence of 15% after the first ten commands. This means, if a player managed to get lower than the target error rate, the next request was intentionally misrecognized by the system (even if this turned out to only rarely occur). At the end of the session, all participants were told about the planted errors (minimum 15% overall errors). This was done last in the interviews to not influence any prior assessments.

The environment of the game and the game logic have been built with Unity 3D[1]. For speech recognition, the Google Cloud Speech-To-Text service[2] was used. The requests were directly sent to the Google services. We chose this service as it does not require any native library to run and makes the prototype compatible with any available platform. We created builds for Windows, Mac OS and Linux.

3.1 Anticipatory Error Handling

The anticipatory error handling was implemented to pick the best available option based on the current game state. In effect, if a command was not recognized, the game would perform a locally optimized action regarding obstacle avoidance and goal completion without letting the player know that the recognition failed. The game would first prioritize not getting eaten by the wolf (obstacle), and then would consider the action which would position the sheep closest to the gate (see figure 3). In the following section, we will give an example of how this procedure worked within the context of "Listen, Sparky".

4 EVALUATION

4.1 Study Design

We conducted a between-subjects design user study with ($N = 34$) participants to compare and evaluate our two conditions. In the control group, participants played a version that employed traditional

[1]https://unity3d.com/unity
[2]https://cloud.google.com/speech-to-text

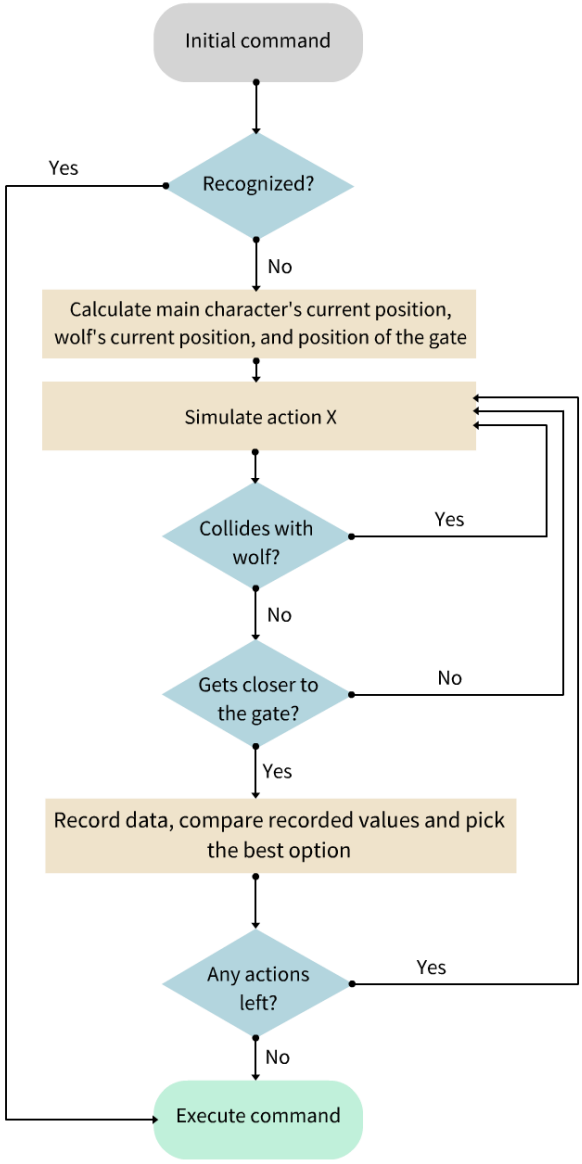


Figure 3: General process of the anticipatory error handling.

error handling, i.e. in the case of non-recognition, the character would not react but only indicate that the command was not recognized by displaying some question marks above its head (see Figure 4).

In the intervention group, players played a version that implemented anticipatory error handling, based on the underlying game state. For instance, considering the game situation in figure 5, the player commands sparky to "bark" but the intent was not recognized. The game would then refer to the error handling system that would then decide which action would be most optimal at

Figure 4: In the control group, when a command is not recognized, the game displays question marks over Sparky's head.

that moment so that the sheep can avoid getting eaten by the wolf and/or can get closer to the gate. The system then chooses "flank left" as the anticipated solution since it would have the best possible outcome where the sheep stays away from the wolf, and it gets close to the gate.

Among both conditions, levels, game environment, and mechanics remained equal, leaving the error handling method as the single manipulated variable. Group assignment was pseudo-randomized between two equally distributed groups. Participants were asked to play all eight levels of "Listen, Sparky" – yet, if they became stuck on a specific level after multiple tries, they were allowed to skip it. The execution took place on the subjects' own PC or laptop device. We sent an executable format of the game (build) to the participants prior to the session and made sure that every player had a functional microphone to use for the game.

4.2 Procedure

Every experimental session was held remotely via video calls. The experiment and interview were recorded acoustically and transcribed for later analysis. Furthermore, the experimenter noted verbal statements and in-game observations while providing assistance in cases of issues. Before starting the session, participants were briefly informed about the experiment procedure. Although the game contained an explanatory tutorial, the interview conductor would shortly explain the game and the controls. After the participants gave informed consent, they would share their screen with the experiment conductor. Participants would then play through the game in either one of the two conditions. They could also take a short break in between the levels. After finishing the game, participants completed the post-exposure questionnaires. At the end of the session, we held a short semi-structured interview with each participant. Each session took approximately 40 – 50 minutes, with an average of 18.4 minutes game-play time (SD = 5.16).

4.3 Measures

In order to evaluate our hypotheses and to understand how players experience the error handling in both conditions, we used standardized questionnaires to assess the player experience and the

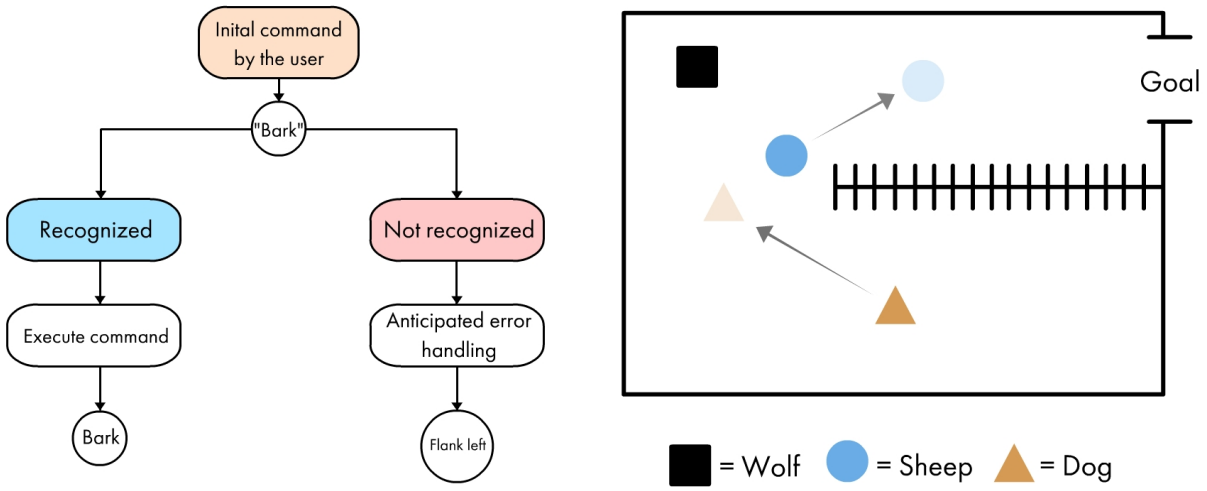


Figure 5: Displaying a specific game situation where the recognition fails and the system chooses to flank left as it would have the best possible outcome (right). The flowchart showing the process of the anticipatory error handling in the intervention group (left).

perceived usability of the system. Our post-exposure questionnaires included demographic questions, the System Usability Scale (SUS) [17], as well as the Player Experience of Need Satisfaction (PENS) [61] throughout the subscales of *Competency*, *Autonomy*, *Relatedness*, *Presence/immersion*, and *Intuitive controls*. Both questionnaires are validated and established measurement instruments. We chose SUS as it is a reliable tool for measuring usability of a system, which ensures high comparability. The PENS is also a validated questionnaire which determines the player experience. In our evaluation, we did not consider the sub-scale of *Relatedness* as it was not relevant to the scope of this study.

Additionally, we recorded a series of customized questions regarding their experience with the game. These were executed via 5-point Likert scales and concerned the extent with which Sparky behaved as the participant expected him to do so, Sparky's perceived intelligence and the overall experience with the game. Above that, players were asked to estimate the approximate number of commands that were not recognized, and to explicate what Sparky did when the commands were not recognized by the system. For all statistical tests, we applied an alpha level of .05. We concluded the session with a brief, semi-structured interview to further evaluate qualitative aspects of player experience, usability, and individual preferences for both conditions [72]. The interview recordings were systematically examined. For this, two researchers agreed on a coding system that was generated from a random selection of ten interviews. Subsequently, all recordings were analyzed, coded along this categorization, and summarized. Additionally, we collected insightful and unique statements.

4.4 Participants

A quota sampling approach was used to recruit participants for this study in which the selection was based on mailing lists, social networks, word-of-mouth and gaming forums. Participation was voluntary and uncompensated. ($N = 34$) people participated in the experiment. In the control group, 17 participants (5 self-identified as female, 12 as male) between 22 and 43 years of age ($M = 29.64$, $SD = 5.42$) played a version of the game with traditional error handling. In the intervention group, 17 players (5 self-identified as female, 12 as male) which were mutually excluded from the first group, between 22 and 38 years of age ($M = 27.7$, $SD = 4.87$) played a version that implemented anticipatory error handling. 85% of our participants had previous experience with voice assistants (18 rarely, 11 often). Only 17% of the participants have previously played a voice-controlled video game. We conducted the experiment in English with international participants. The sample consisted of two native English speakers and the rest were fluent non-native English speakers.

5 RESULTS

In order to identify possible differences between both conditions, we applied Mann-Whitney U Test as well as qualitative content analysis towards our issued research questions.

Four participants did not fill in an item within the Autonomy subscale of PENS. These missing values were imputed by the average value. In our study, we focused on the four sub-scales of *Competency*, *Autonomy*, *Presence/Immersion*, and *Intuitive Controls* (cf. Figure 6). Consequential, we found a significant effect for *Intuitive Controls* in favor for the intervention group ($M = 5.96$, $SD = 1.29$), compared to the control group ($M = 4.7$, $SD = 1.98$), $U = 84.5$, $p = .040$, displaying a medium effect ($d_{Cohen} = 0.75$) [24]. In contrast, *Competency*, *Autonomy*, and *Presence/Immersion* did not show significant differences between the two conditions ($p > .05$).

Regarding usability, SUS scores reached an average of 63.23 ($SD = 20.47$) within the control group, whereas the intervention

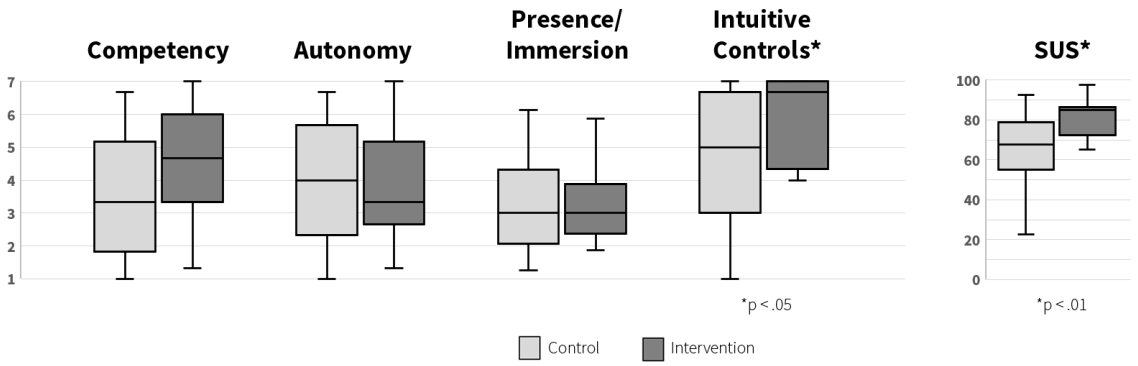


Figure 6: Boxplot indicating significant results from PENS-subscales and SUS between control and intervention group. Includes median (–), standard deviation (box) and range (whiskers).

group resulted in 80.88 ($SD = 8.96$). The subsequent Mann-Whitney U Test indicates that anticipatory error handling outperformed the control group significantly in terms of usability ($U = 65.5$, $p = .0069$, cf. Figure 6), revealing a large effect between conditions ($d_{Cohen} = 1.055$). Any SUS score higher than 68 would be considered above average, and anything lower is below average [16]. Therefore, the results indicate an above average usability score for the intervention conditions and a below average one for the control condition. For the overall game experience, players of the control group rated it as 3.411 ($SD = 1.18$) on average, not significantly different from the intervention group ($M = 3.889$, $SD = 0.93$; $U = 111$, $p = .254$). Assessing to what extent Sparky followed the users' expectations, no significant differences between the control ($M = 3.12$, $SD = 0.99$) and intervention group ($M = 3.24$, $SD = 0.90$) could be found ($U = 134.5$, $p = .74$). Similarly, no significant effect on Sparky's perceived intelligence emerged ($U = 134$, $p = .72$), with an average of 3.06 ($SD = 0.97$) under the control condition, and 2.94 ($SD = 1.14$) within the intervention group.

Overall, the participants in the intervention group had a mean error rate of 42.94% ($SD = 18.41$), while the control group resulted in 33.53% ($SD = 11.31$) errors on average. This showed no significant differences between the two conditions in terms of error rates ($U = 98$, $p = .114$).

However, when participants were asked to write down the approximate number of commands that were not recognized by the system, the mean number of perceived errors in the control group resulted in 34.863 ($SD = 36.882$) which is significantly higher ($U = 63.5$, $p = .0056$) than that of the intervention group ($M = 6.438$, $SD = 5.501$), revealing a large effect ($d_{Cohen} = 1.09$).

We also asked the participants to explain Sparky's behavior in cases where commands were not correctly recognized by the system. In the intervention group, 59% believed it did something wrong, 23% said it did something random, 12% said it always understood the commands, and 6% thought, it helped to perform the right action. Among the participants under the control condition, 76% said Sparky did not react when the command was not recognized, 12% said it did something wrong, one participant (6%) said it did something random and another stated that the commands were always recognized.

5.1 Qualitative Results

Interpreting the post-exposure interview sessions, qualitative insights could be extracted with respect to the different error handling methodologies and the overall game experience itself.

5.1.1 Overall Impressions. Participants generally enjoyed playing the game and attributed it as entertaining. Some (21%) even asked for repeating the game levels to get more playtime. In both groups, players liked the idea of playing a speech-based video game in general and found it novel. Many players (15 of 34=44%) stated that they perceived the game's controls as intuitive and several (41%) mentioned that they especially liked the game's aesthetics. Four participants fancied the background music and sound effects used in the game while one user found it distracting. Three participants gave suggestions regarding the use of speech-based games in serious contexts such as teaching and therapy. One participant specifically mentioned that speech-based games such as this game could be an interesting medium to teach foreign languages to children. Two participants stated that they felt like an actual shepherd in control of a dog. One participant said, "I had a feeling of control because the dog behaved as I intended. I was in command of the dog."

5.1.2 Progressive Enhancement. In both groups, participants mentioned that they got better at controlling Sparky after some playing time. One participant stated: "I felt that I learned how to speak for the game to understand me". However, some (10 of 34=29%) believed that with their improvements, the game's challenges also got more complex. Eight players (29%) stated that they specifically enjoyed the progressive enhancement of the game's difficulty. One user suggested using different difficulty levels, where the recognition gets worse when you increase the difficulty.

5.1.3 Voice Command For Game Control. During the sessions, we observed that all participants looked at the commands list more than once. Even though the controls were rather limited, participants were struggling to memorize them all. We also noticed that our participants would look at this list to use the exact phrase suggested in that screen. Although, the recognition system was able to handle different styles of commands in the same context that were likely

to be given by participants, and thus not limited to the particular commands from the tutorial. This was observed by several players. One of the participants stated: "The commands were intuitive. I did not use exactly the game's commands, and it still worked. I liked that". On the other hand, players wished for fewer restrictions regarding the commands for the game controls. One player said, "I'd expect all the normal replacement phrases to work as well". Some (12%) participants shared an opinion that more controls would be helpful, e.g. one participant stated that "it would be nice to have a command that repeats the previous one". One participant said that single-word commands would be better for such games. Few (9%) believed that using phrases felt more natural and interesting. On multiple occasions, we witnessed that the participants' voices were raised, or they spoke faster when they were under time pressure and had to make quick decisions, which likely led to a higher error rate.

5.1.4 Recognition and Error Handling. Both groups equally (five participants in each group) reported the disliking of the occurrence of voice recognition malfunctioning, as well as the delay between the command and execution. Two participants (both none-native English speakers) expressed their struggle with the recognition due to their accent and mentioned that it would have been nice if the system could learn their voice and accent. One player in particular found it entertaining that Sparky could not understand all the commands: "It felt more realistic this way". Three participants (two from the intervention group, one from the control group) believed that 100% of their commands were recognized by the system, although none of the participants had a smaller error rate than 17%.

Seven participants in the intervention group mentioned that they sometimes found the behavior of Sparky unexpected. Only one player in the control condition mentioned something similar. During the interviews, we revealed both conditions and their difference in error handling to the participants. Four of them mentioned that they would prefer to have anticipatory error handling as an optional feature that they could activate in the game's settings. One participant stated, "When the recognition is not working, that means there is a problem. If I don't see the errors, I don't see the problem. So I think the errors should be seen to acknowledge the problem and improve the recognition". Another mentioned "I would personally choose this version [repetition-based] as I want to have full control of the game." Multiple participants (15%) of the intervention group shared the opinion that they like that the game's flow is not being disturbed by recognition errors. One of them stated: "I really like the idea of this game since it does not disturb the flow when there is an issue with the recognition technology". One participant said, "I would prefer that the game performs an action randomly. That way, it makes the game more exciting and challenging".

6 DISCUSSION

This evaluation aimed at exploring the impact of recognition error handling techniques on the user experience by contrasting traditional to anticipated handling within a speech-controlled video game. Overall, users' feedback about "Listen, Sparky!" were rather positive and supporting. Players in both conditions generally enjoyed playing our voice-controlled game. During the experiment, participants asked for repeating the levels even after successfully finishing that level. They also wanted to continue playing after the experiment was done. Three of our participants specifically pointed out that controlling Sparky with voice made them feel more immersed as they felt 'like an actual shepherd', supporting the findings by Allison et al. [4], that the player's in-game voice commands can be associated with a feeling of taking on a character in the game's world. Moreover, we witnessed a significant difference in terms of intuitive control between the two conditions. This can imply that implementing optimal error handling can lead to a higher perceived intuitiveness of a system.

Many players expressed their struggle with the recognition of their commands, especially in the beginning of the game. We observed that participants improved in understanding how the recognition system works after spending some time in the game. They learned how to formulate their commands and to speak clearly in order to be recognized by the system. Additionally, they also developed their ability to play the game by adopting the game mechanics over the various levels. Furthermore, we saw that many participants looked at the controls screen multiple times during the game to use the exact phrases suggested in that screen, even though the recognition system was able to handle different types of commands for the same action. This was inline with the previous work by Jentsch et al. [11] who also mentioned that the users adapt special communication strategies to speak to the system.

Additionally, we observed that players often perceived time pressure, leading to more complications with command recognition. This was mainly due to the change in the talking pace and fast decisions, which at times led to unclear and incorrect inquiries. We also recorded a higher error rate for non-native speakers. This led to more frustration for these players during the game, aligning with the results of the study by Pyae et al [58].

Eventually, we interpreted the results of this experiment to provide answers to the following comprehensive questions:

RQ1: Does performing a locally optimized game action in times of misrecognition lead to a measurably improved usability in a speech-based video game?
RQ2: What are the effects on player experience in terms of competency, autonomy, presence, and intuitive control, if error handling mechanisms decide for unintended actions?

Regarding **RQ1**, results indicate a significantly higher usability, as well as higher ratings of intuitive control for the version employing anticipatory error handling. Yet, qualitative statements underline that this increase of usability is mainly due to the cases where the error handling actually followed the user's intention, which was not always the case, even when deciding for the technically optimized solution. In cases of mismatch, participants perceived it as a different kind of error, even if the performed action was the technically optimized choice. As soon as doubts about the system were raised, the learning curve of the users was also impacted. Thus, we argue that error handling can improve the user experience of speech-based games, though the major objective of the handling technique should not approximate technically optimized decisions, but individually tailored predictions. Supplementary to the usability analysis, quantitative findings of the recorded error observations

confirm the former results: Although participants of the intervention group committed more errors on average, they in fact reported a significantly lower amount of perceived errors, compared to the control group. In effect, we accept our first hypothesis:
H1: Participants will observe a lower number of recognition errors in case of anticipatory error handling.
Even though this was partially caused by the fact that in the intervention group, a certain number of unrecognized commands by the users were in fact the optimized action, therefore no recognition failure was perceived. Nonetheless, even when a misrecognition was handled by an (optimized) action that deviated from the intended command, users were still less likely to detect this intervention.

Furthermore, based on the results of the PENS questionnaire, we accept $H2_d$, while rejecting $H2_a$, $H2_b$, and $H2_c$. Concerning **RQ2**, we observed differences between both groups and interpreted users' reactions and responses to error handling that conflicted with their original intention. Players of the intervention group were repeatedly confused by Sparky acting against their original intention, resulting in a misleading learning experience that impaired in-game progress and proficiency attainment. Since the control group was not affected by automatically handled actions, this issue did only occur in the former condition. Even if quantitative insights suggest a higher usability through the anticipatory error handling intervention, qualitative statements reflect the dissatisfaction in situations where the handling deviates from the user's intention. Above that, since correctly handled errors were not perceived as errors in the first place, participants rated the intervention version as not more intelligent than the without handling.

After we revealed both conditions to our participants, we witnessed a mixture of opinions regarding the different error handling methodologies. Some were in favor of the anticipatory error handling as it helped to keep the flow of the game. Some didn't like it as they believed it hides the problem rather than solving it. One participant also proposed performing a random action rather than an optimized one to make the game more challenging and add an element of surprise. Considering all the differences in the opinions, one can assume that the optimal solution could differ from one player to another. Game developers can consider equipping different methods as optional features of the game, where the players can choose their desired methods based on their own preferences.

In our study, we used a limited set of commands to focus on the error handling methodologies. In cases of larger command sets, the system can eliminate those commands which are out of the current context and between those left, choose the most suitable action based on the situation and previous user behavior. Depending on the application, one can take the action with the highest probability or present the users with a number of top possible actions to choose from. Previous research on repair strategies with chatbots has shown that system-repair where the chatbot provides possible options to users was arguably favored by the users as it required less effort from the user to repeat their inquiry [6]. This can likely be enhanced with machine learning techniques and user models.

Predicting user's intent to improve usability and user experience is not a new topic. In terms of conversational agents for instance, there has been extensive research on predicting user intents and deciding for an appropriate repair strategy in case of a conversation breakdown [6, 37, 62]. In the context of video games, however, this area is still under investigated. There have been attempts to employ deep learning to provide adequate models of individual player behavior with high accuracy [54], or opponent modeling to predict different strategy patterns of opponents [33]. However, to the best of our knowledge, this is the first work aiming at alternative strategies of error handling in times of command recognition failures in voice-controlled video games. In this study, we witnessed that players did not necessarily favor the cases that the anticipatory error handling was used if the action did not match their initial intent. One could assume that having full control over the game and perceiving a feeling of agency could be rather preferred, even if their actions are not the optimal ones towards level completion. Although, repeating the game actions when they are not recognized was even more frustrating. Therefore, more effort should be put on understanding the user's initial actions rather than finding the optimal action. Nonetheless, more aspects such as player types, mood, and game genre's need to be investigated in order to gain a deeper insight in this regard.

Based on the interpretation of the results regarding both research questions, we conclude with the following implications: Error handling can significantly improve the usability of a speech-controlled video game and aid in bridging the technological gap of speech recognition. Yet, ideal error handling should model (and predict) the individual user's intention, be equipped with an internal likelihood estimation whether the handled decision is appropriate or follow similar methods to ensure user satisfaction. Otherwise, false handling can impair both the experience as well as the learning progress and raise doubts about error handling in general. This work successfully demonstrated a first approach of anticipatory error handling, but these "optimal decisions" from a heuristic can still deviate from the user's intention. Future work will extend this by approximating the users' intention even further (e.g. creating user models).

6.1 Limitations and Future Work

While the findings of this study present significant steps forward in exploring recognition error handling methodologies in speech-based games, there are still some limitations that should be addressed. In this work, we investigated anticipatory error handling in a speech-based video game. Although the broader insights of this evaluation can apply to the use and error handling of VUIs in general, in future work, these methods could be transferred and evaluated in other domains such as navigation, medicine, education, and smart homes, to explore conversationally more complex settings.

The anticipatory error handling used in this study could also raise certain ethical concerns in terms of misleading the user into thinking that no error has occurred. Although this may not be a big concern in the context of most video games and the approach may help players with speech recognition, developers implementing this method should transparently make information about the history of errors and the commands that lead to the error handling available to the users. Furthermore, it may also be necessary that the method is explained to the players.

Another concern that could be raised is that certain players may abuse such features by purposefully giving unclear commands,

being certain that the system would perform the optimal action. Although we did not observe such behavior during the experiment, we recommend designers and developers to consider methods to prevent players from misusing this feature, for instance, by observing odd behavior from the player such as repeated unrecognized commands.

During the experiment, we noticed that some participants had difficulties learning the game controls and game mechanics. For future studies, we recommend longer tutorials as well as gaming sessions to counter influences on individual learning rate. Apart from this, differences in player types and players' current emotional and social states could lead to different experiences, which should be incorporated and reflected in further studies. The implemented voice recognition system for the game has not been trained with data from non-native English speakers, yet the majority of participants fell under this condition. The recognition with those who spoke a strong accent was therefore not optimal and could have been improved by training the system differently. Although our game controls were limited to a predefined set of commands, this helped us to have a structured procedure with high comparability [57]. The focus of this work was to study occurrences of recognition failures and the subsequent handling and not to engineer a solution for a large-scale complex system. In order to yield scalable insights for broader application fields and cover large command vocabularies, future studies will expand the scope of the potential actions.

This was a first exploration on anticipatory error-handling in video games. Our experiment sample consisted of ($N = 34$) mostly male users (70.6%). An influence of such bias on the results can not be excluded. Moreover, future studies could validate our findings by investigating a wider population.

The positive feedback and enthusiasm towards our game can be partially affiliated by the unconventionality of speech-based video games in general. The demographics data as well as the perceived novelty of the game by the participants shows that voice-controlled games are still an unfamiliar category. In this paper, we demonstrated that utilizing speech-based interaction in games can help to increase inclusion and as well as immersion as you are actively communicating with in-game characters instead of just pressing buttons. We further encourage researchers in this field to investigate the area.

7 CONCLUSION

In this paper, we investigated anticipatory error handling for a speech recognition system and explored its potentials and challenges. We designed a voice-controlled video game called "Listen, Sparky!" to evaluate our concept. In a between-subjects design study, we compared our anticipatory error handling model to a traditional repetition-based version. Our results showed that implementing anticipatory error handling can improve the usability of a system, if it follows the intention of the user. Otherwise, it can impair the user experience, even when making technically optimized decisions. Ideal error handling should therefore model the individual user's intention, be equipped with an internal likelihood estimation whether the handled decision is appropriate, or follow similar methods to ensure user satisfaction. Our findings contribute useful insights for researchers and developers on how to address, display and handle recognition errors in speech-based video games and the greater application field of voice user interfaces.

ACKNOWLEDGMENTS

This work was partially funded by Klaus Tschira Foundation, by the FET-Open Project 951846 "MUHAI – Meaning and Understanding for Human-centric AI" funded by the EU program Horizon 2020, as well as the German Research Foundation DFG as part of Collaborative Research Center (Sonderforschungsbereich) 1320 "EASE – Everyday Activity Science and Engineering", University of Bremen (http://www.ease-crc.org/) conducted in subproject H02.

REFERENCES

[1] Beena Ahmed, Penelope Monroe, Adam Hair, Chek Tien Tan, Ricardo Gutierrez-Osuna, and Kirrie J Ballard. 2018. Speech-driven mobile games for speech therapy: User experiences and feasibility. *International journal of speech-language pathology* 20, 6 (2018), 644–658.

[2] Fraser Allison, Marcus Carter, and Martin Gibbs. 2017. Word Play: A History of Voice Interaction in Digital Games. *Games and Culture* 15, 2 (2017), 91 – 113. https://doi.org/10.1177/1555412017746305

[3] Fraser Allison, Marcus Carter, Martin Gibbs, and Wally Smith. 2018. Design Patterns for Voice Interaction in Games. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play* (Melbourne, VIC, Australia) *(CHI PLAY '18)*. Association for Computing Machinery, New York, NY, USA, 5–17. https://doi.org/10.1145/3242671.3242712

[4] Fraser Allison, Joshua Newn, Wally Smith, Marcus Carter, and Martin Gibbs. 2019. Frame Analysis of Voice Interaction Gameplay. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300623

[5] Fraser John Allison. 2020. *Voice interaction game design and gameplay.* Ph.D. Dissertation. University of Melbourne.

[6] Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300484

[7] Jonas Austerjost, Marc Porr, Noah Riedel, Dominik Geier, Thomas Becker, Thomas Scheper, Daniel Marquard, Patrick Lindner, and Sascha Beutel. 2018. Introducing a virtual assistant to the lab: A voice user interface for the intuitive control of laboratory instruments. *SLAS TECHNOLOGY: Translating Life Sciences Innovation* 23, 5 (2018), 476–482.

[8] Matthew P Aylett, Per Ola Kristensson, Steve Whittaker, and Yolanda Vazquez-Alvarez. 2014. None of a CHInd: relationship counselling for HCI and speech technology. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, USA, 749–760.

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

[10] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the long-term use of smart speaker assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–24.

[11] Maresa Biermann, Evelyn Schweiger, and Martin Jentsch. 2019. Talking to Stupid?!? Improving Voice User Interfaces. https://doi.org/10.18420/muc2019-up-0253

[12] Dan Bohus and Alexander I Rudnicky. 2005. Constructing accurate beliefs in spoken dialog systems. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.* IEEE, IEE, New York, NY, USA, 272–277.

[13] Dan Bohus and Alexander I Rudnicky. 2008. Sorry, I Didn't Catch That! In *Recent trends in discourse and dialogue*. Springer, New York, NY, USA, 123–154.

[14] Richard A Bolt. 1980. "Put-that-there" Voice and gesture at the graphics interface. In *Proceedings of the 7th annual conference on Computer graphics and interactive techniques*. Association for Computing Machinery, New York, NY, USA, 262–270.

[15] Cynthia Breazeal. 2003. Toward sociable robots. *Robotics and autonomous systems* 42, 3-4 (2003), 167–175.

[16] John Brooke. 2013. SUS: a retrospective. *Journal of usability studies* 8, 2 (2013), 29–40.

[17] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[18] Raluca Budiu and Page Laubheimer. 2018. Intelligent assistants have poor usability: A user study of Alexa, Google assistant, and Siri.

[19] Erik Cambria and Bebo White. 2014. Jumping NLP curves: A review of natural language processing research. *IEEE Computational intelligence magazine* 9, 2 (2014), 48–57.

[20] Marcus Carter, Fraser Allison, John Downs, and Martin Gibbs. 2015. Player Identity Dissonance and Voice Interaction in Games. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play* (London, United Kingdom) *(CHI PLAY '15)*. Association for Computing Machinery, New York, NY, USA, 265–269. https://doi.org/10.1145/2793107.2793144

[21] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, IEEE, New York, NY, USA, 4960–4964.

[22] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, et al. 2019. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers* 31, 4 (2019), 349–371.

[23] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, and et al. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article 475, 12 pages. https://doi.org/10.1145/3290605.3300705

[24] Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Erlbaum, Hillsdale, NJ.

[25] Erica Cooper, Alison Chang, Yocheved Levitan, and Julia Hirschberg. 2016. Data Selection and Adaptation for Naturalness in HMM-Based Speech Synthesis. In *Proc. Interspeech 2016*. ISCA, France, 357–361. https://doi.org/10.21437/Interspeech.2016-502

[26] 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*. ACM, New York, NY, USA, 1–12.

[27] Crytek. 2013. *Ryse: Son of Rome*. Game [XBox One]. Microsoft Studios, Redmond, Washington, U.S.

[28] Mihaly Csikszentmihalyi. 1990. *Flow: The psychology of optimal experience*. Vol. 1990. Harper & Row, New York, NY, USA.

[29] Steven Dow, Manish Mehta, Ellie Harmon, Blair MacIntyre, and Michael Mateas. 2007. Presence and engagement in an interactive drama. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, New York, NY, USA, 1475–1484.

[30] EA Sports. 2013. *Fifa 14*. Game [XBox One]. Microsoft Studios, Redwood City, California, U.S.

[31] Electronic Arts. 2012. *Mass Effect 3*. Game [XBox 360]. Electronic Arts, Redwood City, California, U.S.

[32] Susumu Harada, Jacob O Wobbrock, and James A Landay. 2011. Voice games: investigation into the use of non-speech voice input for making computer games more accessible. In *IFIP Conference on Human-Computer Interaction*. Springer, Springer, New York, NY, USA, 11–29.

[33] He He, Jordan Boyd-Graber, Kevin Kwok, and Hal Daumé III. 2016. Opponent modeling in deep reinforcement learning. In *International conference on machine learning*. PMLR, PMLR, New York, New York, USA, 1804–1813.

[34] Hyunhoon Jung, Hee Jae Kim, Seongeun So, Jinjoong Kim, and Changhoon Oh. 2019. TurtleTalk: an educational programming game for children with voice user interface. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–6.

[35] Jesper Juul. 2007. Without a goal: on open and expressive games. , 191–203 pages.

[36] John F Kelley. 1983. An empirical methodology for writing user-friendly natural language computer applications. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 193–196.

[37] Knut Kvale, Olav Alexander Sell, Stig Hodnebrog, and Asbjørn Følstad. 2019. Improving Conversations: Lessons Learnt from Manual Analysis of Chatbot Dialogues. In *International Workshop on Chatbot Research and Design*. Springer, Springer, New York, NY, USA, 187–200.

[38] Kwan Min Lee, Wei Peng, Seung-A Jin, and Chang Yan. 2006. Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human–robot interaction. *Journal of communication* 56, 4 (2006), 754–772.

[39] Toby Jia-Jun Li, Igor Labutov, Brad A Myers, Amos Azaria, Alexander I Rudnicky, and Tom M Mitchell. 2018. An end user development approach for failure handling in goal-oriented conversational agents.

[40] Irene Lopatovska, Katrina Rink, Ian Knight, Kieran Raines, Kevin Cosenza, Harriet Williams, Perachya Sorsche, David Hirsch, Qi Li, and Adrianna Martinez. 2019. Talk to me: Exploring user interactions with the Amazon Alexa. *Journal of Librarianship and Information Science* 51, 4 (2019), 984–997. https://doi.org/10.1177/0961000618759414 arXiv:https://doi.org/10.1177/0961000618759414

[41] Silvia Lovato and Anne Marie Piper. 2015. "Siri, is This You?": Understanding Young Children's Interactions with Voice Input Systems. In *Proceedings of the 14th International Conference on Interaction Design and Children* (Boston, Massachusetts) *(IDC '15)*. ACM, New York, NY, USA, 335–338. https://doi.org/10.1145/2771839.2771910

[42] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf Between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[43] Michal Luria, Samantha Reig, Xiang Zhi Tan, Aaron Steinfeld, Jodi Forlizzi, and John Zimmerman. 2019. Re-Embodiment and Co-Embodiment: Exploration of Social Presence for Robots and Conversational Agents. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (San Diego, CA, USA) *(DIS '19)*. Association for Computing Machinery, New York, NY, USA, 633–644. https://doi.org/10.1145/3322276.3322340

[44] 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* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376311

[45] David Maulsby, Saul Greenberg, and Richard Mander. 1993. Prototyping an intelligent agent through Wizard of Oz. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 277–284.

[46] Juliana Miehle, Daniel Ostler, Nadine Gerstenlauer, and Wolfgang Minker. 2017. The next step: intelligent digital assistance for clinical operating rooms. *Innovative surgical sciences* 2, 3 (2017), 159–161.

[47] Cosmin Munteanu, Matt Jones, Sharon Oviatt, Stephen Brewster, Gerald Penn, Steve Whittaker, Nitendra Rajput, and Amit Nanavati. 2013. We need to talk: HCI and the delicate topic of spoken language interaction. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 2459–2464.

[48] Christine Murad and Cosmin Munteanu. 2019. "I don't know what you're talking about, HALexa" the case for voice user interface guidelines. In *Proceedings of the 1st International Conference on Conversational User Interfaces*. Association for Computing Machinery, New York, NY, USA, 1–3.

[49] Christine Murad, Cosmin Munteanu, Benjamin R Cowan, and Leigh Clark. 2019. Revolution or Evolution? Speech Interaction and HCI Design Guidelines. *IEEE Pervasive Computing* 18, 2 (2019), 33–45.

[50] Moyen Mohammad Mustaquim. 2013. Automatic speech recognition-an approach for designing inclusive games. *Multimedia tools and applications* 66, 1 (2013), 131–146.

[51] Aasish Pappu and Alexander Rudnicky. 2014. Knowledge acquisition strategies for goal-oriented dialog systems. In *Proceedings of the 15th annual meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. Association for Computational Linguistics, Philadelphia, USA, 194–198.

[52] Jim R Parker and John Heerema. 2008. Audio interaction in computer mediated games.

[53] Cathy Pearl. 2016. *Designing voice user interfaces: principles of conversational experiences.* " O'Reilly Media, Inc.", Sebastopol, California.

[54] Johannes Pfau, Jan David Smeddinck, and Rainer Malaka. 2018. Towards deep player behavior models in mmorpgs. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. ACM, New York, NY, USA, 381–392.

[55] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 640, 12 pages. https://doi.org/10.1145/3173574.3174214

[56] Martin Porcheron, Joel E. Fischer, and Sarah Sharples. 2017. "Do Animals Have Accents?": Talking with Agents in Multi-Party Conversation. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. ACM, New York, NY, USA, 207–219. https://doi.org/10.1145/2998181.2998298

[57] Robert Porzel and Manja Baudis. 2004. The Tao of CHI: Towards Effective Human-Computer Interaction. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*. Association for Computational Linguistics, Boston, Massachusetts, USA, 209–216. https://www.aclweb.org/anthology/N04-1027

[58] Aung Pyae and Paul Scifleet. 2018. Investigating differences between native English and non-native English speakers in interacting with a voice user interface: A case of Google Home. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction*. ACM, New York, NY, USA, 548–553.

[59] D Reddy, Lee Erman, and R Neely. 1973. A model and a system for machine recognition of speech. *IEEE Transactions on Audio and Electroacoustics* 21, 3 (1973), 229–238.

[60] Mihai Rotaru, Diane J Litman, and Katherine Forbes-Riley. 2005. Interactions between speech recognition problems and user emotions.

[61] Richard M Ryan, C Scott Rigby, and Andrew Przybylski. 2006. The motivational pull of video games: A self-determination theory approach. *Motivation and emotion* 30, 4 (2006), 344–360.

[62] Amir Shevat. 2017. *Designing bots: Creating conversational experiences.* " O'Reilly Media, Inc.", Sebastopol, California.

[63] Gabriel Skantze. 2003. Exploring human error handling strategies: Implications for spoken dialogue systems.

[64] Adam J Sporka, Sri H Kurniawan, Murni Mahmud, and Pavel Slavík. 2006. Non-speech input and speech recognition for real-time control of computer games. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*. ACM, New York, NY, USA, 213–220.

[65] Bernhard Suhm, Brad Myers, and Alex Waibel. 2001. Multimodal error correction for speech user interfaces. *ACM transactions on computer-human interaction (TOCHI)* 8, 1 (2001), 60–98.

[66] Marc Swerts, Diane Litman, and Julia Hirschberg. 2000. Corrections in spoken dialogue systems.

[67] Turn 10 Studios. 2013. *Forza Motorsport 5.* Game [XBox One]. Microsoft Studios, Redmond, Washington, U.S.

[68] Markku Turunen, Jaakko Hakulinen, K-J Raiha, E-P Salonen, Anssi Kainulainen, and Perttu Prusi. 2005. An architecture and applications for speech-based accessibility systems. *IBM Systems Journal* 44, 3 (2005), 485–504.

[69] Keith Vertanen and Per Ola Kristensson. 2010. Getting it right the second time: Recognition of spoken corrections. In *2010 IEEE Spoken Language Technology Workshop*. IEEE, IEEE, New York, NY, USA, 289–294.

[70] Marco Filipe Ganança Vieira, Hao Fu, Chong Hu, Nayoung Kim, and Sudhanshu Aggarwal. 2014. PowerFall: a voice-controlled collaborative game. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play*. ACM, New York, NY, USA, 395–398.

[71] Tom Wilcox, Mike Evans, Chris Pearce, Nick Pollard, and Veronica Sundstedt. 2008. Gaze and voice based game interaction: the revenge of the killer penguins. *SIGGRAPH Posters* 81, 10.1145 (2008), 1400885–1400972.

[72] Chauncey Wilson. 2013. Interview techniques for UX practitioners: A user-centered design method.

[73] Rainer Winkler, Matthias Söllner, Maya Lisa Neuweiler, Flavia Conti Rossini, and Jan Marco Leimeister. 2019. Alexa, Can You Help Us Solve This Problem?: How Conversations With Smart Personal Assistant Tutors Increase Task Group Outcomes. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. ACM, New York, NY, USA, Article LBW2311, 6 pages. https://doi.org/10.1145/3290607.3313090

[74] Beste F. Yuksel, Penny Collisson, and Mary Czerwinski. 2017. Brains or Beauty: How to Engender Trust in User-Agent Interactions. *ACM Trans. Internet Technol.* 17, 1, Article 2 (Jan. 2017), 20 pages. https://doi.org/10.1145/2998572

[75] Nima Zargham, Michael Bonfert, Georg Volkmar, Robert Porzel, and Rainer Malaka. 2020. Smells Like Team Spirit: Investigating the Player Experience with Multiple Interlocutors in a VR Game. In *Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY EA '20)*. ACM, New York, NY, USA, 408–412.

[76] Rui Zhao, Kang Wang, Rahul Divekar, Robert Rouhani, Hui Su, and Qiang Ji. 2018. An immersive system with multi-modal human-computer interaction. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, IEEE, New York, NY, USA, 517–524.

Publication 3

Multi-agent Voice Assistants: An Investigation of User Experience

Nima Zargham, Michael Bonfert, Robert Porzel, Tanja Döring, and Rainer Malaka

In Proceedings of the 20th International Conference on Mobile and Ubiquitous Multimedia (MUM 2021). New York, NY, USA, 2021. Association for Computing Machinery.

Personal contribution to this work: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, validation, visualization, and contribution to all parts of the manuscript.

ISBN: 978-1-4503-8643-2/21/05 DOI: 10.1145/3490632.3490662

Multi-Agent Voice Assistants: An Investigation of User Experience

Nima Zargham *
zargham@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

Michael Bonfert *
bonfert@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

Robert Porzel
porzel@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

Rainer Malaka
malaka@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

Tanja Döring
tanja.doering@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

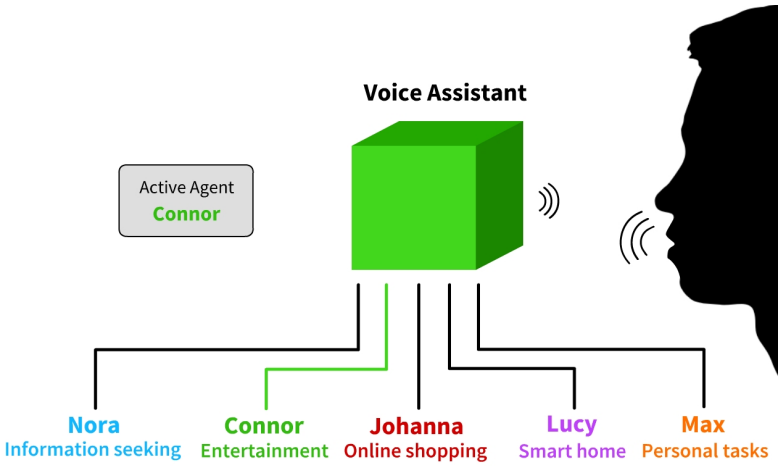


Figure 1: Schematic illustration of a multi-agent voice assistant with five available agents, their specialized task domain, and representing color. Here, the user talks to the currently active agent Conner indicated by the cube-shaped device lighting up green.

ABSTRACT

The use of voice assistants (VAs) is spreading widely. Most common VAs consist of a single, usually female voice that responds to the user's inquiry. We designed a VA system appearing as a group of agents, each with a different voice and a specialized task domain. We conducted a quantitative user study comparing our multi-agent approach with a conventional single-agent assistant in a smart home scenario as virtual reality (VR) simulation. The results show significantly higher user experience ratings for the multi-agent concept. Based on our findings, we discuss the potentials and challenges of designing multi-party VA systems.

CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**; User studies; Virtual reality.

KEYWORDS

Voice Assistants; Multiparty; Virtual Reality; Smart Home; User Experience

ACM Reference Format:
Nima Zargham, Michael Bonfert, Robert Porzel, Rainer Malaka, and Tanja Döring. 2021. Multi-Agent Voice Assistants: An Investigation of User Experience. In *20th International Conference on Mobile and Ubiquitous Multimedia (MUM 2021), December 5-8, 2021, Leuven, Belgium.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3490632.3490662

*Both authors contributed equally to this research.

MUM 2021, December 5-8, 2021, Leuven, Belgium

ACM ISBN 978-1-4503-8643-2/21/05…$15.00
https://doi.org/10.1145/3490632.3490662

1 INTRODUCTION

Today, technological advancements enable humans to use one of the oldest means of communication for interacting with computers: their voice. Voice input is a feature in a variety of devices such as mobile phones, cars, and smart speakers. Early versions of Voice Assistants (VAs) were designed for handling few specialized tasks [40] but current devices have broad capabilities in performing different actions and thus have become more universal, e.g., for smart home control, work management, scheduling, gathering information, navigation, communications, education, or entertainment. VAs such as Apple's Siri, the Google Assistant, Microsoft's Cortana, and Amazon's Alexa are available on smartphones, tablets, cars, and in homes as Apple HomePod, Google Home, or Amazon Echo. The latter is referred to as home assistants. Interaction with most VAs is highly task-oriented and limited to simple question-and-answer dialogues. Except for scripted jokes and humorous responses, the communication is mostly neither conversational nor social [15].

Technical aspects as well as the user experience (UX) of VAs have been the subject of extensive research in recent years (e.g., [6, 10, 12, 14]). Most VAs feature a single human-like and usually female voice as default, which can reinforce the gender stereotypes of women [20]. Using a single voice for VA devices gives users the impression that one agent is assisting them with their tasks.

In our work, we present the concept of a multi-agent voice assistant. The system appears as a team of agents, each using a different voice and responsible for a specific task domain, as shown in Figure 1. Based on the user's input, the most suitable agent is selected to respond to the user. With this, the characters can be perceived as individuals and as competent experts of their responsibilities. Previous research has also suggested that users may benefit from having access to several customizable VA personas [1]. We hypothesize that this concept has the potential to provide an enhanced UX regarding interaction qualities of, e.g., pleasure, fun, or helpfulness compared to conventional single-agent voice assistants.

In some of the current consumer home assistants, users can already use specific third-party features represented by another voice. The default VA agent can be asked to connect to other third-party agents, e.g., local public transportation services. The third-party agent then appears in the device – a form of re-embodiment [29]. After the task is completed, this agent leaves again and is therefore not continuously present. It is not possible to address these third-party agents directly but only through the default agent as an operator. In our proposed concept, in contrast, all agents are directly accessible and readily available in the device.

To evaluate our concept, we developed a prototype and conducted a user study. In our experiment, we simulated an immersive smart home environment in virtual reality (VR) for comparing the UX of our multi-agent system with that of a conventional single-agent VA quantitatively. Conducting the user study in VR enabled a home-like setting without intruding the participants' homes with a wiretap, and facilitated the technical implementation while keeping the interaction authentic. Previous research has established that field studies simulated in VR can be a powerful research tool. Researchers witnessed largely similar behavior between virtual and real settings [2, 16, 31, 33, 38]. They also provided suggestions for conducting studies in VR such as designing scenarios that facilitate natural behavior and free self-discovery of the technology by the user [31], which we applied in our work in order to get ecologically valid results.

The aim of this user study is to provide insights into the user experience with a multiplicity of agents incorporated in a VA and investigate its potentials and challenges. We present our results showing higher ratings for the attractiveness and aspects of hedonic quality in the multi-agent condition. Further, we discuss automatic agent selection and user customization, limits to the number of agents, as well as considerations for applying the concept in scenarios outside VR.

2 RELATED WORK

With a large design space, the development of a VA requires complex and fundamental design decisions, including linguistic aspects and the degree of human likeness as humor, emotions, gender, attractiveness, trust, or politeness. One of this plethora of design parameters is the number of interlocutors in a conversation. All these aspects affect the user experience in its combination. In this section, we summarise related work on voice assistant usage, human characteristics in VAs, and multiparty conversation.

2.1 VA Usage

A large body of literature has investigated how people use their VAs [4, 12, 27]. Research has shown that frequent users of voice assistants use these systems mainly in hands- or eyes-busy situations, and they do not trust the system to do complex tasks such as composing emails or making phone calls [28]. Providing dedicated specialists for certain tasks could highlight unexploited capabilities of the system. In a study on long-term use of home assistants, Bentley et al. [4] found that specific types of commands were made more often at particular times of day and that commands in some domains increased in length over time as participants tried out new ways to interact with their devices, yet exploration of new topics was low. Lopatovska et al. [25] explored user interactions with Amazon Alexa and classified them as casual or leisurely, not exclusively directed at retrieving information. The reports of heavier use over the weekends and satisfaction, even when Alexa did not produce desirable outcomes, suggest that a pleasant UX might be more important to the users than the quality of the outcome. The study by Yuksel et al. [49] also suggests that agent reliability is not more important than agent attractiveness and suggest that the latter may be even more important.

2.2 Human Characteristics in VAs

People find it easier to interact with technology that resembles some of their own characteristics [8]. Consequently, many innovative technologies – as well as VAs – support social interactions and include forms of human-like behavior. Nass, Steuer, and Tauber coined this phenomenon as the *Computers are Social Actors* paradigm [37]. It describes how users apply social rules while interacting with a computer, such as gender stereotypes [51] or politeness [35].

Nass et al. showed that even computers with minimized gender cues in the voice output evoke gender-based stereotypical responses [36]. They examined three gender-based stereotypes with no gender indicators but vocal cues and witnessed stereotypes in all

cases. The researchers claim that the tendency to gender stereotype is extremely powerful, extending even to inanimate machines. In a follow-up study, Nass and Moon showed that users found the praise from a male-voiced computer more compelling than the same comments from a female-voiced computer [34]. A study by Hwang et al. [20] explored how gender stereotypes toward women are reflected in assistants with female voices. They categorized three distinct characteristics: bodily display, subordinate attitude and sexualization. Authors suggested that these stereotypical traits could create a power dynamic between users and female agents. As recently introduced in some commercial VAs, users have the possibility to change the voice of the agent to voices of different genders. However, they still need to decide on either a male or a female voice. Using multiple agents, a balance in VA genders can be achieved.

The conversational nature of VAs has the potential to trigger personification tendencies in users which in turn can translate into consumer loyalty and satisfaction [26]. Personification can be defined as attribution of "human-like properties, characteristics, or mental states to real or imagined nonhuman agents and objects" [19]. A study by Pradhan et al. [43] on device personification in Amazon Echo showed that users who personified the device were more likely to be satisfied with it.

To create a more trustful interaction with digital assistants, researchers have been using social-psychological aspects of human-human interactions and applying them to human-computer interaction [5, 45]. Luger et al. [28] showed in 2016 that user expectations tend to exceed the agent's abilities, which are still limited to simple tasks such as checking the weather or setting reminders. Many of the current VAs fall back on humor in response to complex conversational input and commands that cannot be handled otherwise. Their responses might be seen as sarcastic or entertaining [41] and create the impression of an underlying personality. Morkes et al. [32] found that virtual agents using humor are rated as being more likable, competent, and cooperative. In a qualitative study exploring the experience of infrequent users of VAs, Cowan et al. [15] found that their participants clearly understood that Siri, like most VAs, was designed to be seen as human-like. They imbued Siri with intelligence and personality, with people seeing Siri as being "sassy" and "friendly" and mentioned that its human-like qualities affected how they felt towards it as they did not want to hurt its feelings.

Bonfert et al. [6] investigated how users react when the VA demands to be addressed politely. While all participants yielded to the demand, not everyone was pleased about the agent's attitude or even insulted the agent. For the application of VAs in households with several users, especially with children, the availability of various agents with different behavior and tone settings might be useful, depending on discourtesy, offensive language, or imperiousness, for instance.

2.3 Multiparty Conversation

While conversation with more than two interlocutors have been largely explored among humans [7] and when involving many persons talking with one artificial agent [21, 39], there has been little research on multiple agents conversing with one user. Previous studies have explored empirical models [17], agent embodiment [47], or dialogue management [23], thus the agents' contribution to the dialogue and how it should react to the user. On the other hand, aspects of the users' behavior and their UX when interacting with a multi-agent system are still a young research topic.

A recent study on user reactions towards a text-based multi-agent system was presented by Chaves and Gerosa [11]. In a between-group experiment, the authors analyzed the change of speech and reactions of users to a multiplicity of chatbots compared to a single chatbot. They report no significant effect of the number of agents on conversation structure or content. However, the multi-agent interaction led to more confusion. In a previous study [50], we investigated the interaction with a voice-based multi-agent system in a single-player VR game. We compared a game version where players could talk to a multiplicity of agents using natural language to a version with a single agent. The study showed that the participants preferred conversing with a group of interlocutors, found it more entertaining and felt like being part of a team.

Luria et al. [29] explored user perspectives on different configurations of the social presence of robots and conversational agents. In a study, the researchers tested four conditions: One agent per body, one agent that is present in different bodies, one agent moving from one body to others (called re-embodiment), and one agent joining another within the same body (called co-embodiment). The users reported to feel comfortable with the re-embodiment between different physical entities. The experiment was inconclusive about the co-embodiment scenario as it mostly yielded results on dialogues between the artificial agents observed by the user. The authors propose to integrate individual agents as experts for specialized tasks.

A study by Abdolrahmani et al. [1] suggests that providing simultaneous access to multiple different VA personas would be an effective method for providing suitable support in variable contexts for blind users. Since the appropriateness of the output depends heavily on the context and content of the interaction, authors believe that users may benefit from having access to several customizable personas, such as for cooking or scheduling each. The authors suggest incorporating multiple conversational personas into a single VA device and allow users to flexibly configure the speed, tone, volume, and other characteristics of each persona's interaction style.

To build upon prior work regarding multiple agent personas incorporated into a single VA system, in this study, we co-embody multiple agents into one device, each responsible for one specialized task domain, to determine how this affects the user experience in a smart home setting. To the best of our knowledge, such a multi-agent approach has not yet been scientifically evaluated.

3 USER STUDY

We designed a voice assistant in VR that appears as one device integrating multiple agents, which are each specialized on a specific task domain. When the user initiates a dialogue, the most suitable agent for the respective task is selected automatically by the system for responding with its distinguished voice. An automatic selection is expedient here, as otherwise it might be difficult for the user to learn the names of all agents within the brief time of testing the system for selecting agents manually. Automatic selection avoids confusion and makes the interaction easier.

Figure 2: The virtual smart home environment. The voice assistant device on the right is embodied as a hovering cube. Its orange color represents the currently active agent Max responsible for "personal tasks".

We evaluated the system in a within-subject study in VR, in which the participants (N=20) used both our novel multi-agent VA and, for comparison, an equivalent single-agent VA in a virtual smart home environment that is shown in Figure 2. The experiment was conducted in a lab environment before COVID-19 with the participant and the experimenter being in the same room.

3.1 VR Home Environment Setup

State-of-the-art VR technology allows the creation of interactive high-fidelity simulations. Advantages of virtual environments as a laboratory tool include the high degree of experimental control, the low costs, and the ease of replicating the experiment elsewhere [22]. Using a VR simulation also ensured a convincing apartment environment rather than a lab setting without installing eavesdropping smart speakers in the participants' homes, which would be more difficult in implementation, would limit the control over the structured experiment procedure, and would lead to privacy concerns as audio recordings of the users' daily personal life are transmitted for analysis.

In the virtual apartment, the user found a VA represented by a hovering cube, as shown in Figure 2, and numerous smart appliances to represent a smart home environment. To provide better visual feedback about system activity, we embodied the VA as a hovering cube that rotates around its center when listening or responding. The cube is grey while the device is idle and yellow while listening. It changes its color depending on the responding agent as every agent is represented by an individual color. This is supposed to give the user the impression that the respective agent is present and temporarily occupying the cube.

Moreover, there were some objects in the room, such as a smartphone, that the participants could grab and interact with. The participants were instructed to carry out 12 tasks, which were listed on a black board on the wall of the apartment, depicted in Figure 3. In the multi-agent condition, each agent was responsible for at least 2 of the tasks. A small screen next to the black board displayed the name of the active agent in the agent's corresponding color. In the case that a colorblind user could not attribute the color to an agent, the displayed name and distinctive voice allowed to identify the active agent. After performing a task successfully, the item's font color changed from white to green.

3.2 Prototype Implementation

The prototype was implemented in Unity 3D. We used an HTC Vive Pro with its Wand controllers for the VR setup. Users could move in the environment using teleportation. We used a microphone to capture the participants' voice. A speech recognition system was implemented using the Windows Phrase Recognition[1] for analyzing the users' voice commands. As a fallback solution, we implemented an interface for the experimenter to trigger the responses manually in cases of repetitive failed recognition to avoid user frustration and thus potential biases. The manual trigger was mostly used for non-native speakers, although all participants were fluent in English.

Similar to using conventional home assistants, users had to say a wake word each time before giving a command. They could always interrupt it by saying "stop". When activated with the wake word, the visual representation of the VA changed its color to yellow and started rotating to indicate that the device is ready to receive commands, and otherwise remained grey and static. Only the commands required for the experiment were recognized by the system. All other inquiries were declined with a typical response for unsupported features. We generated all the voices for the VA agents using the Text-To-Speech tool TTSmp3[2].

3.3 Participants

The 20 participants (7 female, 13 male) were between 21 and 34 years of age (M = 27.3, SD = 3.83). 85% of the participants had previous experience with voice assistants (12 rarely, 5 often). A quarter of our participants did not have prior experience in virtual reality. We conducted the experiment in English with international students from the university campus. Thus, the sample consisted of non-native speakers. For the possibility of recognition issues due to a language barrier, the experiment conductor was prepared to trigger the correct response manually.

3.4 Study Design

We conducted a within-subjects user study and compared our approach of a multi-agent VA to a conventional single-agent VA in a VR smart home environment. The participants were given a set of typical tasks to perform with a home assistant triggering typical system reactions: switching the lights on or off, playing music, playing a video, locking the door, asking about the weather, and purchasing a product online are some examples. We avoided tasks that could ambiguously fall into multiple task domains. When a participant made an inquiry, the corresponding agent reacted with a typical system reaction that always included a verbal response. For example, when the user asked for music, the corresponding agent replied "Okay, here's some jazz music" followed by a jazz track. The participants had to perform these tasks using voice commands

[1] https://docs.unity3d.com/ScriptReference/Windows.Speech.PhraseRecognitionSystem.html
[2] https://ttsmp3.com

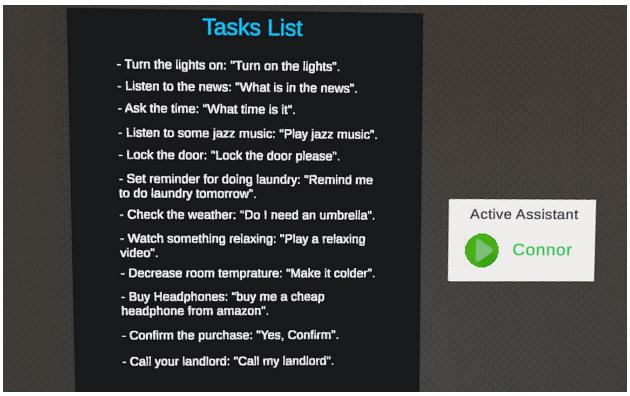


Figure 3: The task list and the indicator panel for the currently speaking agent inside the virtual apartment. Participants could freely choose the order of the tasks. After completion, tasks turned green on the list.

and could freely choose the order. We had two conditions for this experiment. One condition was working with a single-agent VA. Like the industry default, we implemented a female voice. The wake word for this device was either one of "OK Jupiter" or "Hey Jupiter". The color assigned to Jupiter was cyan and it would respond to all of the users' queries.

The other condition was a multi-agent VA which consists of five agents. Each agent was designed to assist the user in a specific task area and each agent was assigned a color to visually represent the agent. Based on previous work regarding main purposes of using VAs [44], we identified the five typical task domains *information seeking*, *entertainment*, *online shopping*, *smart home devices*, and *personal tasks*. Nora (blue, female voice) was designed to respond to the information seeking (news and weather related questions), Connor (green, male) was responsible for Entertainment (music and video related topics), Johanna (red, female) was in charge of the online shopping, Lucy (purple, female) was responsible for the smart home, and Max (orange, male) was responsible for personal tasks (reminders, alarms and shopping list). The agents with their respective genders were assigned arbitrarily to the task domains.

To interact with the device, users had to use the wake word "Super Squad". Because the users could go through the tasks in any preferred order, the sequence of active agents was not predetermined. The environment and mechanics for both conditions were the same, but the two conditions differed in their set of tasks to avoid repetition. For instance, while in one condition the users had to set an alarm, they had to set a reminder in the other condition. We counterbalanced the order of conditions to compensate for habituation effects.

3.5 Procedure

After giving informed consent, we asked all our participants to fill in a questionnaire prior to the experiment indicating the demographics. After this, the participants were given a short tutorial session on how to use the VR system. We explained the functionality of the VA and the interactions needed for the experiment. The experiment conductor explained briefly how to accomplish the tasks that the participant needed to perform. In the beginning of each condition, the VA would give a short explanatory introduction and mention in which ways it could help the user. In the multi-agent condition, all the agents would introduce themselves individually and mention their task domain. They clarified that all agents are part of the same system. Participants would then perform the experiment with one of the two conditions. After finishing the first round, participants filled in the post-exposure questionnaires consisting of the System Usability Scale (SUS) [9] and the User Experience Questionnaire (UEQ) [24], two established measurement tools with high comparability for assessing the aspects of UX and usability separately.

After taking a short break, they would continue with the second condition and its post-exposure questionnaires. The experiment concluded with a short semi-structured interview where the participants were asked to give their overall opinion and elaborate on the potentials and challenges of a multi-agent voice assistant system. Each test session took approximately 30 – 40 minutes with about 15 minutes in VR. The questionnaires were presented on a laptop.

4 RESULTS

Participants rated their experience with the VAs in both conditions separately. In this section, we present our findings in two sections of quantitative and qualitative evaluation.

4.1 Quantitative Evaluation

We identified significant differences in user experience between the two conditions in favor of the multi-agent concept as shown in Figure 4. We witnessed these significant differences in all subscales of hedonic quality but no significant differences in pragmatic quality aspects.

For the comparison of each subscale in the UEQ, we ran a paired-samples t test as suggested by the questionnaire's handbook, with an alpha level of .05. As the visual interpretation of the histograms raised doubts about the normal distribution of some subscales, we double checked the observed effects with a Wilcoxon signed-rank test as non-parametric equivalent. For this reason, we report two test results for comparison.

The analysis shows that the multi-agent condition ($M = 1.44$, $SD = 1.21$) was rated significantly higher in the subscale *stimulation* than the single-agent condition ($M = 0.66$, $SD = 1.04$), $t(19) = 2.105$, $p = .049$, 95% CI $[0.00, 1.55]$ (t test); $Z = -1.99$, $p = .046$ (Wilcoxon test). The effect size of this difference is $d_{Cohen} = 0.47$, which corresponds to a medium effect [13]. In the *novelty* subscale, the multi-agent condition ($M = 1.39$, $SD = 1.13$) was also rated significantly higher compared to the single-agent condition ($M = -0.20$, $SD = 0.96$), $t(19) = 5.137$, $p < .001$, 95% CI $[0.94, 2.23]$ (t test); $Z = -3.464$, $p = .001$ (Wilcoxon test). For novelty, the effect size is $d_{Cohen} = 1.14$, which corresponds to a large effect.

The t test shows a significant difference for the scale *attractiveness* between multiple agents ($M = 1.53$, $SD = 1.16$) and a single agent ($M = 0.75$, $SD = 1.03$), $t(19) = 2.374$, $p = .028$, 95% CI $[0.09, 1.47]$, with a medium effect size of $d_{Cohen} = 0.53$. In contrast to the t-test outcome, the results of the non-parametric Wilcoxon

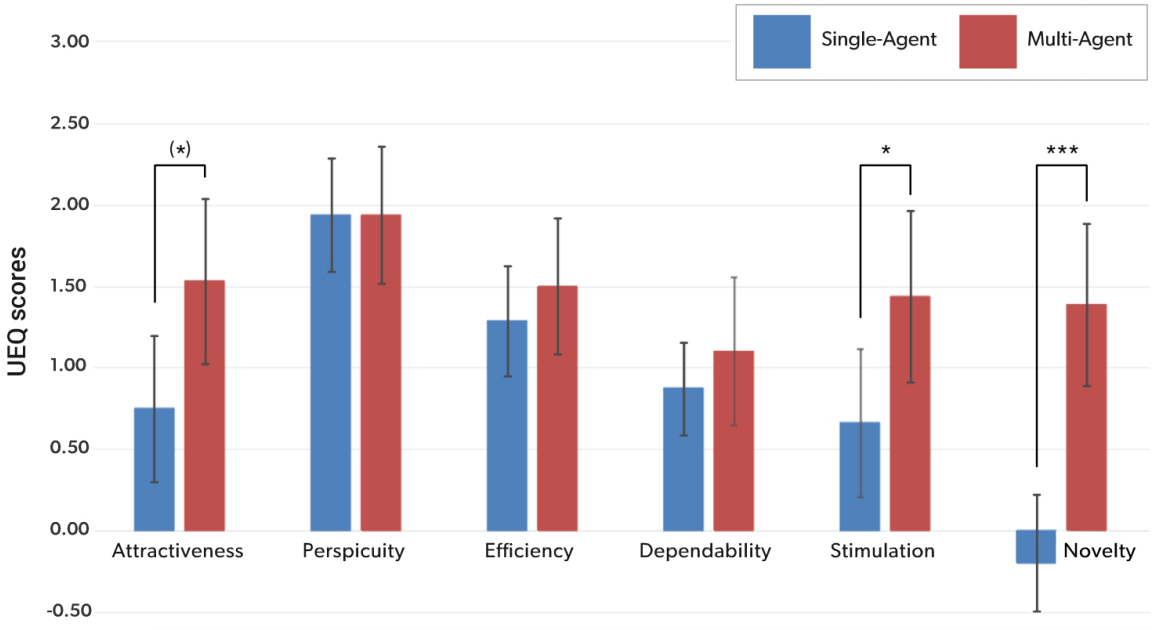


Figure 4: Results of the User Experience Questionnaire (UEQ) indicating means and standard deviations of the system ratings with a single agent (blue) and multi agents (red) on Likert scales from −3 to +3. Significant differences are marked with asterisks. The asterisk in brackets marks inconclusive results of the statistical tests for the subscale attractiveness as explained in section 4

test did not show a significant difference between the groups ($Z = -1.90$, $p = .058$) yielding conflicting analysis outcomes. Considering both test results and the descriptive stats, the data indicate an effect that would presumably become unambiguously verifiable with a larger sample. The other three subscales perspicuity, efficiency and dependability did not show significant differences between the two conditions ($p > .05$).

Regarding usability, the single-agent condition has a mean SUS score of 77.0 ($SD = 12.6$) and the multi-agent condition of 76.8 ($SD = 15.3$). This indicates an above average usability score for both conditions with no significant difference ($Z = -0.228$, $p = .82$).

In a set of multiple-choice questions, we asked the participants to choose their favorite in terms of *entertainment*, *friendliness*, *trust*, and their *overall preference* between the two systems. 70% of the participants rated the multi-agent VA as more entertaining while 20% chose the single-agent variant and 10% had no preference. 60% found the multi agents friendlier, 40% the single agent. 40% of participants trusted the single-agent VA more, 35% had higher trust in the multi-agent system, and 25% had no preference. Overall, 70% of participants preferred the multi-agent VA over the single-agent VA (20%) with 10% having no preference.

4.2 Qualitative Evaluation

In the interviews after the experiment, we learned that the users would have liked to change the agents' task domain based on their voices. For instance, one participant said: "I think Max should do my online shopping. He sounds more trustable." Another participant thought that "it would be funny to hear the news from Connor" instead of Nora. Other participants wanted to make similar changes. Two participants mentioned that "it felt more natural to talk to one person" as currently done by conventional home assistants. Four of our participants pointed out that they found the VAs sounded a bit robotic. Apart from that, the comments in the interviews validated the single-agent system as an adequate equivalent to current standard home assistants. One participant, for example, referred to it as "basically an Alexa clone". Some participants spent more time on getting to know the agents and their voices better by letting them introduce themselves multiple times or by repeating commands to hear the responses again.

5 DISCUSSION

Overall, the ratings and comments by the users of our proposed multi-agent VA were positive and supportive. In the following, we discuss our insights and recommendations derived from this study.

5.1 Usability and User Experience

We evaluated the usability and user experience quantitatively in a virtual smart home environment. In terms of *perspicuity*, *efficiency* and *dependability*, there were no differences found between the two conditions – all of which are pragmatic qualities. This comes as no surprise since pragmatic qualities relate to the perceived usability of the system. There were no differences in the performance or functionality of the two systems. The tasks are performed in the

same way and the multiple agents do not interact with the user concurrently. The results of the SUS validated the comparability of the two systems' usability. No significant differences were found. This suggests that the proposed multi-agent approach is not harder to learn or understand, comparably easy and clear to use, equally fast and practical, and as helpful and predictable as the status quo of a single-agent VA. These findings advocate for multi-agent VAs as a warranted approach for further in-depth investigation. Our findings in terms of conversation properties align with a study on interacting with a multiplicity of chatbots that showed no influence of the number of agents on the structure or content of the conversations [11].

On the other hand, participants rated the user experience of the multi-agent condition significantly better in terms of hedonic qualities, comprising the subscales *novelty* and *stimulation*. As expected, the new concept of being assisted by a team of agents is perceived as more novel, rated along the items *creative*, *inventive*, *leading edge*, and *innovative*. The effect size of $d = 1.14$ indicates that this simple modification of VAs allows prolonged perceived novelty as a central factor to user engagement [3]. Further, the multi-agent approach had significantly better ratings on the subscale *stimulation* comprising the items *valuable*, *exciting*, *interesting*, and *motivating*. An explanation might be the more diversified interactions resulting in less monotony while performing the tasks. Another reason might be the experienced support by a team of agents who are all dedicated to assist the user in a group effort. In accordance with this, the application of a multi-agent system in a game context has been shown to be perceived as more entertaining and produce a feeling of team spirit for the players [50].

We also observed indications for higher *attractiveness* ratings for the multi-agent system. This subscale comprises the items *enjoyable*, *pleasing*, *good*, *pleasant*, *attractive*, and *friendly*. Further data collection is required to determine unequivocally whether the attractiveness of multi-agent VAs is indeed perceived as higher as the applied statistical tests yield ambiguous results. A further supporting indicator is that 70% of the users preferred the multi-agent system overall and that 70% found it more entertaining.

5.2 Agent Selection and Customization

In line with the recommendation by Luria et al. [29], we decided to implement specialized agents in our experiment that are accountable for one dedicated area of expertise. We designed a system selecting the agents automatically based on the task domain to avoid confusion. Another intention was to not overwhelm the participants with memorizing the names and domains of the agents. However, several participants expressed the desire to assign the task domains manually to the individual agents – a form of customization. Future research should look into the possibility of selecting agents for specific task domains. In the context of scientific experiments, we suggest using neutral wake words and agent names to avoid biases between the conditions. In this study, we did not observe or learn about any influences of the wake words "super squad" or "OK Jupiter".

The qualitative feedback showed that participants perceived different voices as different characters with personalities. They believed that specific agents could be more suitable for a certain task domain and convey different character traits based on voice factors such as tone, gender, or accent. This attitude extends to certain expectations toward the character of agents depending on their domain. For instance, users preferred a trustworthy character in charge of online shopping for handling sensitive data. Here, the availability of various agents for different task domains presents a substantial advantage. Heterogeneous responsibilities require individual character traits. While for reasons of comparability we designed all agents as neutral as possible, the results of our study show that a variety of personalities can be beneficial. It is imaginable to have a close relationship with your agent for entertainment and music – somebody who understands you and is funny. The agent responsible for the calendar should be dependable and efficient. Personal bonding with the banking agent is not necessary, but it needs to be trustworthy. Hence, we recommend designing the personality of a voice assistant agent specifically for the assigned expertise.

It is impossible to appeal to all user preferences equally with one implementation. Customization could be of importance for developers and designers to stand out by offering a team tailored to the personal preferences and needs of a user [18]. Two participants reported in our study that speaking to a single agent felt more natural to them. Both participants owned a smart speaker and used them daily. The perception of an unfamiliar multiparty system feeling unnatural may be attributed to a habituation effect.

In current voice assistants, there is a conspicuous bias of predominantly female voices as default. Providing multiple agents allows a balance of different genders and thus counteract stereotypical associations [20]. Although our proposed system does not resolve the problem with gender stereotypes in voice assistant interaction, it provides a new opportunity to decrease the current bias. We encourage designers to use the opportunity of multi-agent systems to foster agent diversity in the voice assistant market.

As an alternative to the assignment of the agents to an area of expertise each, it has been suggested in the literature to provide each user in a household with their individually responsible agent to increase accountability [30]. This feature can already be observed in recent consumer devices. Beyond that, we propose to explore a hybrid-system that allows multiple personalised agents for each resident. This way, the individual role distribution is independent from the system of other household members. This approach could increase the degree of personalization for each user without increasing the system complexity for others. Moreover, in this work we mainly explored single user scenarios where one person interacts with a multi-agent VA. Future work could further investigate multi-user scenarios and how different household members can interact with such systems.

In order to improve the design and customization of agent personality in multi-agent systems more systematically in the future, the application of personality models for speech-based conversational agents as presented by Völkel et al. [48] could be a valuable approach.

5.3 Number of Agents

In our study, we found that participants had difficulties with keeping track of the different characters even though we used indicators

for the active agent. It became apparent that users needed to spend more time with the system to be able to recognize the agents and their respective domain. Our findings align with the results of a study by Chaves and Gerosa [11] on interaction with a multi-agent chatbot interface. As in our experiment, the multiplicity of agents resulted in confusion for some participants. An implication of this is that multi-agent systems should not exceed a certain complexity to prevent overwhelming the user and to not impede establishing a relationship with the individual characters. A long-term study embedded in a real home scenario would help to understand the process of familiarization, the individual bonding, and the dynamic reassignment of the agents' roles. For future work on short-term usage scenarios, we suggest reducing the number of agents to avoid confusion. For investigating long-term usage, we recommend allowing longer usage times or several sessions for the users to familiarize themselves with the system. The complexity of the interaction from the number of artificial and human interlocutors involved should be in a reasonable relation to the intensity and length of the interaction for the respective application, for example a business context (daily and short), customer service (once and short), tutoring (short-term and intense) or at home (long-term and intense).

5.4 Application of the Results to Real-World Settings

VR simulations have become a common research method in a broad range of applications [2, 16, 31, 33, 38, 46]. For example, Ville et al. [31] recently compared the results of a real-world field study to the results of an in-VR study replicating the environment and found that studies conducted in VR can yield ecologically valid results. Among their insights was that users need to be provided with a scenario that facilitates natural behavior and gives the user a chance to discover the technology that is under investigation alone. In our study, we considered these aspects and built a VR environment with a typical living room setting in which participants could explore the technology as they wanted.

Beyond that, we would like to reflect on transferring our results to real-world settings as some aspects of the representation of the agents would need to be adapted. For transferring our design to a real-world setting, it is especially important to consider how available agents are indicated and how the currently speaking agent is embodied. In our prototype, we assigned one representing color to every agent. When processing a command, the device would light up in the color of the active agent. In a smart speaker, for example, this could be adapted by indicating the selected agent through integrated LEDs, as built into Google Home. Alternatively, a translucent casing around the device could be fully illuminated similar to our visualization. The additional indicator in the form of a panel naming the current agent could presumably be omitted, because users realistically work with the system for a longer time on a regular basis and will, therefore, get familiar with their personal team of agents.

Having provided the participants with a fixed set of voice commands could potentially reduce the ecological validity of the experiment. However, the benefits of this approach are a structured and comparable procedure as well as predictable responses by the VA without the need to implement a universally functioning system but merely a prototype with focused capabilities. The participants could phrase the commands flexibly in any suitable way. This approach is an established method for testing dialog systems [42]. In our experiment design, all tasks were distinctively assignable to one task domain. Disambiguation of tasks that might fall into the responsibility of multiple agents should be considered for designing commercial systems.

6 CONCLUSION AND FUTURE WORK

In this paper, we presented a voice assistant appearing as a team of agents – each with a different voice and specialized task domain – and explored its potentials and challenges. We developed a prototype with multiple agents to evaluate our concept in a simulated smart home environment. The results of our experiment, conducted in VR, show encouraging potential for multi-agent VAs and a high user approval. We found significantly better ratings of the user experience for the aspects of hedonic quality, stimulation and novelty. The results further show indications for a higher attractiveness in favor of the multi-agent system.

The findings of this study are a starting point for adapting multi-agent VAs in smart home environments and other contexts. One implication of the results is to involve only a limited number of agents to not strain the user cognitively. User responses in our study indicated the desire to customize some design factors of the voice assistant such as the number of agents, their voices, and their roles – something that could be addressed by future research. In this study, we investigated a smart home scenario with simple tasks to perform that are representative for the domain. The multi-agent concept could be transferred to other use cases and investigate conversationally more complex operations.

During the experiment, we obtained different opinions about the multi-agent approach and its effects on user experience. Investigating the influence of user characteristics, their current mood or their attitude towards assistive technology could, therefore, be an interesting research topic for the future. Beyond that, further studies will be needed to expand on our experiment as long-term field studies to observe the user adaption over several weeks in everyday situations.

The first consumer home assistants started to enable the choice of one preferred voice per user which results in multiple agents in one VA for households with several users. This could be considered a preliminary version of a voice interface with multiple agents, which offers – while still limited to one universally responsible agent per user – a team of agents for the household. This could be extended by user selection of different voices and personalities for different tasks, moods, children, guests, times of the day, rooms, or device types.

ACKNOWLEDGMENTS

We would like to sincerely thank Yvonne Rogers for helping us shaping this work. We also thank the anonymous reviewers whose suggestions helped improve and clarify this work. This work was partially funded by Klaus Tschira Foundation.

REFERENCES

[1] Ali Abdolrahmani, Kevin M. Storer, Antony Rishin Mukkath Roy, Ravi Kuber, and Stacy M. Branham. 2020. Blind Leading the Sighted: Drawing Design Insights from Blind Users towards More Productivity-Oriented Voice Interfaces. *ACM Trans. Access. Comput.* 12, 4, Article 18 (Jan. 2020), 35 pages. https://doi.org/10.1145/3368426

[2] Philipp Agethen, Viswa Subramanian Sekar, Felix Gaisbauer, Thies Pfeiffer, Michael Otto, and Enrico Rukzio. 2018. Behavior Analysis of Human Locomotion in the Real World and Virtual Reality for the Manufacturing Industry. *ACM Trans. Appl. Percept.* 15, 3, Article 20 (July 2018), 19 pages. https://doi.org/10.1145/3230648

[3] Simon Attfield, Gabriella Kazai, Mounia Lalmas, and Benjamin Piwowarski. 2011. Towards a science of user engagement (position paper). In *WSDM workshop on user modelling for Web applications*. ACM, New York, NY, USA, 9–12.

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

[5] Timothy W. Bickmore and Rosalind W. Picard. 2005. Establishing and Maintaining Long-term Human-computer Relationships. *ACM Trans. Comput.-Hum. Interact.* 12, 2 (June 2005), 293–327. https://doi.org/10.1145/1067860.1067867

[6] Michael Bonfert, Maximilian Spliethöver, Roman Arzaroli, Marvin Lange, Martin Hanci, and Robert Porzel. 2018. If You Ask Nicely: A Digital Assistant Rebuking Impolite Voice Commands. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO, USA) *(ICMI '18)*. Association for Computing Machinery, New York, NY, USA, 95–102. https://doi.org/10.1145/3242969.3242995

[7] Holly Branigan. 2006. Perspectives on Multi-party Dialogue. *Research on Language and Computation* 4, 2 (2006), 153–177. https://doi.org/10.1007/s11168-006-9002-2

[8] Cynthia Breazeal. 2003. Toward sociable robots. *Robotics and autonomous systems* 42, 3-4 (2003), 167–175.

[9] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[10] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, IEEE, 4960–4964.

[11] Ana Paula Chaves and Marco Aurelio Gerosa. 2018. Single or Multiple Conversational Agents? An Interactional Coherence Comparison. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173765

[12] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, and et al. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article 475, 12 pages. https://doi.org/10.1145/3290605.3300705

[13] Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Erlbaum, Hillsdale, NJ.

[14] Erica Cooper, Alison Chang, Yocheved Levitan, and Julia Hirschberg. 2016. Data Selection and Adaptation for Naturalness in HMM-Based Speech Synthesis.. In *INTERSPEECH*. ISCA, 357–361.

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

[16] Shuchisnigdha Deb, Daniel W Carruth, Richard Sween, Lesley Strawderman, and Teena M Garrison. 2017. Efficacy of virtual reality in pedestrian safety research. *Applied ergonomics* 65 (2017), 449–460.

[17] Frank Dignum, Michael Rovatsos, Matthias Nickles, and Gerhard Weiss (Eds.). 2004. *An Empirical Model of Communication in Multiagent Systems: Advances in Agent Communication*. Springer Berlin Heidelberg.

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

[19] Nicholas Epley, Adam Waytz, and John T Cacioppo. 2007. On seeing human: a three-factor theory of anthropomorphism. *Psychological review* 114, 4 (2007), 864.

[20] Gilhwan Hwang, Jeewon Lee, Cindy Yoonjung Oh, and Joonhwan Lee. 2019. It Sounds Like A Woman: Exploring Gender Stereotypes in South Korean Voice Assistants. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. ACM, New York, NY, USA, Article LBW2413, 6 pages. https://doi.org/10.1145/3290607.3312915

[21] Martin Johansson, Gabriel Skantze, and Joakim Gustafson. 2014. Comparison of Human-Human and Human-Robot Turn-Taking Behaviour in Multiparty Situated Interaction. In *UM3I 2014*, Samer Al Moubayed, Dan Bohus, Anna Esposito, The NetherlandsHeylenDirk University of Twente, Maria Koutsombogera, Harris Papageorgiou, and Gabriel Skantze (Eds.). ACM Press, New York, New York, USA, 21–26. https://doi.org/10.1145/2666242.2666249

[22] M. Kinateder, E. Ronchi, D. Nilsson, M. Kobes, M. Müller, P. Pauli, and A. Mühlberger. 2014. Virtual reality for fire evacuation research. In *2014 Federated Conference on Computer Science and Information Systems*. IEEE, 313–321. https://doi.org/10.15439/2014F94

[23] Alistair Knott and Peter Vlugter. 2008. Multi-agent human–machine dialogue: issues in dialogue management and referring expression semantics. *Artificial Intelligence* 172, 2 (2008), 69–102. https://doi.org/10.1016/j.artint.2007.06.001

[24] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and Usability Engineering Group*. Springer, Springer Berlin Heidelberg, Berlin, Heidelberg, 63–76.

[25] Irene Lopatovska, Katrina Rink, Ian Knight, Kieran Raines, Kevin Cosenza, Harriet Williams, Perachya Sorsche, David Hirsch, Qi Li, and Adrianna Martinez. 2019. Talk to me: Exploring user interactions with the Amazon Alexa. *Journal of Librarianship and Information Science* 51, 4 (2019), 984–997. https://doi.org/10.1177/0961000618759414 arXiv:https://doi.org/10.1177/0961000618759414

[26] Irene Lopatovska and Harriet Williams. 2018. Personification of the Amazon Alexa: BFF or a Mindless Companion. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (New Brunswick, NJ, USA) *(CHIIR '18)*. ACM, New York, NY, USA, 265–268. https://doi.org/10.1145/3176349.3176868

[27] Silvia Lovato and Anne Marie Piper. 2015. "Siri, is This You?": Understanding Young Children's Interactions with Voice Input Systems. In *Proceedings of the 14th International Conference on Interaction Design and Children* (Boston, Massachusetts) *(IDC '15)*. ACM, New York, NY, USA, 335–338. https://doi.org/10.1145/2771839.2771910

[28] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf Between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[29] Michal Luria, Samantha Reig, Xiang Zhi Tan, Aaron Steinfeld, Jodi Forlizzi, and John Zimmerman. 2019. Re-Embodiment and Co-Embodiment: Exploration of Social Presence for Robots and Conversational Agents. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (San Diego, CA, USA) *(DIS '19)*. Association for Computing Machinery, New York, NY, USA, 633–644. https://doi.org/10.1145/3322276.3322340

[30] 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* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376311

[31] Ville Mäkelä, Rivu Radiah, Saleh Alsherif, Mohamed Khamis, Chong Xiao, Lisa Borchert, Albrecht Schmidt, and Florian Alt. 2020. Virtual Field Studies: Conducting Studies on Public Displays in Virtual Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376796

[32] John Morkes, Hadyn K. Kernal, and Clifford Nass. 1998. Humor in Task-oriented Computer-mediated Communication and Human-computer Interaction. In *CHI 98 Conference Summary on Human Factors in Computing Systems* (Los Angeles, California, USA) *(CHI '98)*. ACM, New York, NY, USA, 215–216. https://doi.org/10.1145/286498.286704

[33] Mehdi Moussaïd, Mubbasir Kapadia, Tyler Thrash, Robert W Sumner, Markus Gross, Dirk Helbing, and Christoph Hölscher. 2016. Crowd behaviour during high-stress evacuations in an immersive virtual environment. *Journal of The Royal Society Interface* 13, 122 (2016), 20160414.

[34] Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues* 56, 1 (2000), 81–103.

[35] Clifford Nass, Youngme Moon, and Paul Carney. 1999. Are People Polite to Computers? Responses to Computer-Based Interviewing Systems1. *Journal of Applied Social Psychology* 29, 5 (1999), 1093–1109. https://doi.org/10.1111/j.1559-1816.1999.tb00142.x

[36] Clifford Nass, Youngme Moon, and Nancy Green. 1997. Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of applied social psychology* 27, 10 (1997), 864–876.

[37] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers Are Social Actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston Massachusetts USA) *(CHI '94)*. Association for Computing

Machinery, New York, NY, USA, 72–78. https://doi.org/10.1145/191666.191703
[38] Viktorija Paneva, Myroslav Bachynskyi, and Jörg Müller. 2020. Levitation Simulator: Prototyping Ultrasonic Levitation Interfaces in Virtual Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376409
[39] Aasish Pappu, Ming Sun, Seshadri Sridharan, and Alex Rudnicky. 2013. Situated Multiparty Interaction between Humans and Agents. In *Human-Computer Interaction. Interaction Modalities and Techniques*, David Hutchison, Takeo Kanade, and Josef Kittler (Eds.). Lecture Notes in Computer Science, Vol. 8007. Springer Berlin Heidelberg, Berlin/Heidelberg, 107–116. https://doi.org/10.1007/978-3-642-39330-3{_}12
[40] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 640, 12 pages. https://doi.org/10.1145/3173574.3174214
[41] Martin Porcheron, Joel E. Fischer, and Sarah Sharples. 2017. "Do Animals Have Accents?": Talking with Agents in Multi-Party Conversation. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. ACM, New York, NY, USA, 207–219. https://doi.org/10.1145/2998181.2998298
[42] Robert Porzel and Manja Baudis. 2004. The Tao of CHI: Towards Effective Human-Computer Interaction. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*. Association for Computational Linguistics, Boston, Massachusetts, USA, 209–216. https://www.aclweb.org/anthology/N04-1027
[43] Alisha Pradhan, Kanika Mehta, and Leah Findlater. 2018. "Accessibility Came by Accident": Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 459, 13 pages. https://doi.org/10.1145/3173574.3174033
[44] Aung Pyae and Tapani N. Joelsson. 2018. Investigating the Usability and User Experiences of Voice User Interface: A Case of Google Home Smart Speaker. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct* (Barcelona, Spain) *(MobileHCI '18)*. ACM, New York, NY, USA, 127–131. https://doi.org/10.1145/3236112.3236130
[45] Byron Reeves and Clifford Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press, New York, NY, USA.
[46] Helmut Schrom-Feiertag, Volker Settgast, and Stefan Seer. 2017. Evaluation of indoor guidance systems using eye tracking in an immersive virtual environment. *Spatial Cognition & Computation* 17, 1-2 (2017), 163–183. https://doi.org/10.1080/13875868.2016.1228654
[47] David Traum and Jeff Rickel. 2002. Embodied agents for multi-party dialogue in immersive virtual worlds. https://doi.org/10.1145/544862.544922
[48] Sarah Theres Völkel, Ramona Schödel, Daniel Buschek, Clemens Stachl, Verena Winterhalter, Markus Bühner, and Heinrich Hussmann. 2020. Developing a Personality Model for Speech-Based Conversational Agents Using the Psycholexical Approach. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376210
[49] Beste F. Yuksel, Penny Collisson, and Mary Czerwinski. 2017. Brains or Beauty: How to Engender Trust in User-Agent Interactions. *ACM Trans. Internet Technol.* 17, 1, Article 2 (Jan. 2017), 20 pages. https://doi.org/10.1145/2998572
[50] Nima Zargham, Michael Bonfert, Georg Volkmar, Robert Porzel, and Rainer Malaka. 2020. Smells Like Team Spirit: Investigating the Player Experience with Multiple Interlocutors in a VR Game. In *Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY EA '20)*. ACM, New York, NY, USA, 408–412.
[51] Sean Zdenek. 2007. "Just roll your mouse over me": Designing virtual women for customer service on the web. *Technical Communication Quarterly* 16, 4 (2007), 397–430.

Publication 4

Smells Like Team Spirit: Investigating the Player Experience with Multiple Interlocutors in a VR Game

Nima Zargham, Michael Bonfert, Georg Volkmar, Robert Porzel, and Rainer Malaka

In Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '20). New York, NY, USA, 2020. Association for Computing Machinery.

Personal contribution to this work: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, validation, visualization, and contribution to all parts of the manuscript.

ISBN: 978-1-4503-7587-0/20/11 DOI: 10.1145/3383668.3419884

Smells Like Team Spirit: Investigating the Player Experience with Multiple Interlocutors in a VR Game

Nima Zargham*
Digital Media Lab
University of Bremen, Germany
zargham@uni-bremen.de

Michael Bonfert*
Digital Media Lab
University of Bremen, Germany
bonfert@uni-bremen.de

Georg Volkmar
Digital Media Lab
University of Bremen, Germany
gvolkmar@uni-bremen.de

Robert Porzel
Digital Media Lab
University of Bremen, Germany
porzel@uni-bremen.de

Rainer Malaka
Digital Media Lab
University of Bremen, Germany
malaka@uni-bremen.de

ABSTRACT

Verbal communication is a central component in collaborative multiplayer gaming and creates a feeling of companionship among the players. In single-player games, this aspect is often missing. Advancements in speech recognition now open new potentials for voice-activated single-player experiences. In this work, we integrated voice interaction to a single-player virtual reality (VR) game. To create a sense of team spirit, we enabled players to talk to a multiplicity of agents using natural language. We hypothesize that conversing with only one agent cannot produce the same level of camaraderie. We conducted a preliminary qualitative user study (N=10) to explore how players experience talking with the in-game characters in the single-agent and the multi-agent condition. Early results suggest that our participants prefer interacting with the group of interlocutors. They perceived the multi-agent condition as more entertaining and liked the feeling of being part of a team.

CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**; Virtual reality; • **Applied computing** → **Computer games**.

KEYWORDS

Multi-agent; Virtual Reality; Voice User Interfaces; Games

ACM Reference Format:
Nima Zargham, Michael Bonfert, Georg Volkmar, Robert Porzel, and Rainer Malaka. 2020. Smells Like Team Spirit: Investigating the Player Experience with Multiple Interlocutors in a VR Game. In *Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '20 EA), November 2–4, 2020, Virtual Event, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3383668.3419884

*Both authors contributed equally to this research.

CHI PLAY '20 EA, November 2–4, 2020, Virtual Event, Canada

ACM ISBN 978-1-4503-7587-0/20/11…$15.00
https://doi.org/10.1145/3383668.3419884

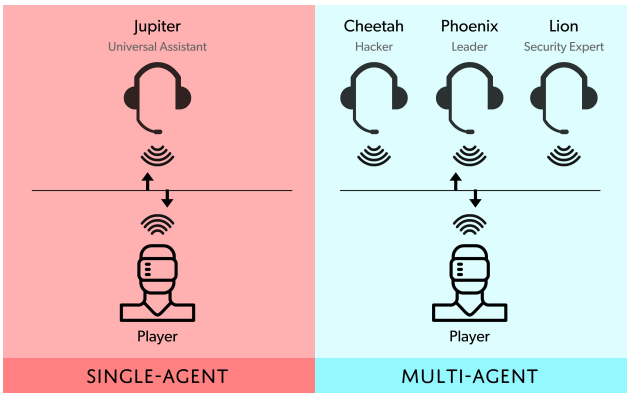


Figure 1: Schematic illustration of the two interaction conditions. (left) In the single-agent version of the game, the player interacts with one universal assistant. (right) In the multi-agent condition, a team of three characters supports the player, each with their unique expertise.

1 INTRODUCTION

Voice interaction receives considerable attention in the entertainment industry, today. Video game companies have integrated voice-activated services and game mechanics in various recent titles. With the rapid technological improvements in speech recognition and the growing availability of microphones in consumer gaming devices, there is an exciting potential for voice user interfaces in gaming [1].

After the release of the Kinect in 2010 with support of voice interaction, Xbox games in a diverse range of genres integrated the new feature, such as *Ryse: Son of Rome* [9], *FIFA 14* [11], *Forza Motorsport 5* [28], and *Dragon Age: Inquisition* [4]. In most cases, however, voice input is an optional feature and not a core element of the game design. For some interactions, such as locomotion or selection, conventional user interfaces, like keyboard or hand-held controllers, provide unrivalled advantages. For social interactions with in-game agents, however, these input techniques seem inexpedient. When agents talk to the player, a more natural interface would enable responding in spoken natural language. Every speech

interaction with a computer automatically evokes an imagined social relationship with the technology and often corresponds with responses that would be given to a human [19]. Verbal exchange is a fundamental component of successful cooperation and cohesion in social groups and key to team performance [18, 25]. We therefore decided upon speech interaction for our investigation on the camaraderie between the player and in-game characters.

In many multiplayer games, players communicate with one another through natural language. This builds a social bond between the players and leads to a feeling of companionship. In single-player games, this aspect is often missing. Voice-activated games attempted to provide natural language input. However, this experience has been described as "uncomfortable" and "awkward" by players [10]. In virtual reality (VR) games, the players' presence is transported to the virtual world allowing them to forget their real environment and being part of the virtual environment [30] as an integrated game character. In contrast to a desktop game, where the player might feel like talking to the computer system rather than in-game agents, we hoped for lower inhibitions in the player to speak freely due to the high presence and involvement in VR. Moreover, the experiment conductors have a high degree of control in VR studies [16] which is especially important for us for stepping in as Wizard of Oz when the automatic recognition failed.

In this paper, we present a VR game that requires bilateral voice interaction with in-game agents for succeeding. We hypothesize that talking with a group of characters has a positive effect on the perceived team spirit, the sense of companionship, and the player experience. Therefore, we developed two versions of the game: The player either talks with one universally assisting character or with a team of specialists, as depicted in Figure 1. The agents are only audible and provide the player with information relevant for accomplishing the level via radio communication. We conducted a preliminary qualitative study to evaluate how our multi-agent approach affects the gaming experience. The experiment is designed around the research question: *What is the impact voice interaction with a multiplicity of interlocutors on the player experience and perceived team spirit in a VR game?* Our results show that the participants preferred the multiplicity of interlocutors as they grew a feeling of group cohesion and found it more entertaining.

2 RELATED WORK

2.1 Voice Interaction in Games

Voice user interfaces are gaining more attention in the recent years due to their intuitive nature – not only in smart homes, phones, or cars, but also in the entertainment industry. Video game companies have been adopting voice input both in games and in game-related services. For instance, Ubisoft recently introduced *Sam*[1] as a virtual assistant which works through voice commands or text chat to provide services for players, such as descriptions of certain items and access to features or games.

Research on voice interaction in digital games dates to the '70s [1]. One of the earliest examples of a game prototype augmented with a speech recognition system is *VoiceChess* supporting standardized chess instructions [24]. Since then, countless video games have embraced the use of voice as input. In a study by Allison et al., the authors surveyed 449 video games and 22 audio games which use players' voice to affect the game state [2]. They observed that academic research has focused on a narrow subset of design patterns, especially pronunciation, and suggest game designers to consider non-verbal forms, which have proven to provide enjoyable game experiences with fast and discrete input possibilities [14, 21, 27, 29].

Researchers argue that the successful integration of voice interaction in digital games is distinct from voice interaction in other contexts, as in games it demands consideration of the identity of the voice. Carter et al. identified that where voice interaction is not related to the virtually embodied experience, it can cause dissonance between the player and their in-game character, resulting in a negative game experience. The authors believe that virtually embodying the player's real voice increases the overlap between the player and character identities [6]. A more recent study shows that in-game voice commands are associated with a sense of taking on a character in the game's world. The researchers believe that voice interactions that are conflicting with the social world can impede the player's engagement with the in-game world [3].

2.2 Multi-agent interaction

A large body of human-computer interaction research investigated the communication between one or more users with one agent [15, 20]. In contrast, research on one user conversing with multiple agents is still a novel topic. In a study by Luria et al., the researchers explored user perspectives on the co-embodiment of multiple voice assistant agents sharing one physical appearance. They encountered reluctance from the users when the agents conversed with each other as this was considered unnecessary in a task-oriented setting [17]. In another study on the user's behavior towards multi-agent systems, Chaves and Gerosa [7] analyzed the change in speech and reactions of users to a multiplicity of chatbots compared to a single chatbot. Their results showed that the multiplicity of chatbots had no significant effect on the conversation structure or its content. The authors report that the multi-agent interaction led to more confusion. Zargham and Bonfert et al. introduced a multi-agent concept for a smart home voice assistant. Five agents supported the user in their specialized task domains, each with an individual voice. Compared to a conventional single-agent system, the user experience was rated significantly higher [31].

2.3 Companionship in Games

In collaborative multiplayer games, players work together as a team and assist one another to reach a common goal. Massively multiplayer online role-playing games (MMORPG), for instance, form a genre that is focused on group cooperation [8]. Studies show that such cooperative games ameliorate the negative effects of violent video game play on cooperative behavior [12]. For many players, the social aspects of playing online games are the most important factor [13]. Single-player games can also have a cooperative component, usually between the player and in-game characters. Some single-player games allow players to control multiple characters, each with a specific purpose or ability, which puts the player in the position of the team leader, e.g., in *Commandos* [23] or *Desperados: Wanted Dead or Alive* [26]. In this genre, however, the player is not

[1] https://club.ubisoft.com/en-US/sam

part of the team but in control of it. In our study, we transferred the concept of being assisted by multiple experts to a collaborative VR game in which the player is part of the team and can speak with the other agents using natural language.

3 EVALUATION

To evaluate our proposed multi-agent concept, we conducted a qualitative within-subject study, in which the participants (N=10) played two versions of a VR game: one with a multi-agent setup and, for comparison, one with a single-agent setup. We counterbalanced the order of the conditions to avoid a bias. The mission, game environment, and game mechanics for both conditions were the same. As the players had to play both conditions one after another, the solutions to the puzzles were different so that the players needed to interact with the agent(s) in both conditions to proceed.

3.1 Prototype Design

The mission for the players in our VR game is to infiltrate a secured bank unnoticed. For stealing the golden statue shown in Figure 2 from the vault, the players solve various puzzles and challenges. They receive instructions and assistance by the agent(s) connected via radio. Different obstacles like security cameras, lasers, and door lock mechanisms introduce challenges that all require consultation with the agents. If the player makes a mistake, e.g., differs from the agents' instructions, the bank's alarm system goes off. In this case, an agent helps and disarm the alarm so that every player is able to finish the mission. The utterances by the agents are triggered by the players' position in the virtual environment, their actions, and their voice commands. For instance, when the player approaches a closed door, the agent(s) will assist in finding the security code to open it. Similarly, the player entering a wrong code or verbally asking for the code triggers the response. The player can ask to repeat instructions or to get additional hints at any time.

For the multi-agent condition, we created a team comprising three characters: Phoenix, the leader of the group; Lion, the security expert; and Cheetah, the hacker. Phoenix and Lion had male voices. Cheetah had a female voice. These agents assisted the player verbally in their specialized area. For instance, if the player needs help with dangerous lasers, Lion responds. For the control condition, we designed a single agent called "Jupiter" with a female voice assisting the player with any kind of challenge. We generated all voices for the agents using an online text-to-speech tool.[2]

The game environment and logic were created with Unity 3D and delivered on an HTC Vive Pro with its native controllers. A speech recognition system was implemented using the Windows Phrase Recognition System[3] and received the audio signal from the built-in microphone of the head-mounted display. As a fallback solution, the experiment conductor could trigger the responses through a graphical interface in case an utterance was not recognized correctly. The responses were limited to those that would help the player accomplishing the mission. Unrelated inquiries have not been answered. It was not required to use a wake word. Players could always interrupt the agents by saying "stop".

[2] https://ttsmp3.com
[3] docs.unity3d.com/ScriptReference/Windows.Speech.PhraseRecognitionSystem.html

Figure 2: Players need to find the depicted gold statue without triggering the alarm by solving a series of puzzles

3.2 Participants

Ten people participated in our experiment (three female, seven male) from 22 to 37 years of age (M = 26.9, SD = 4.06). All our participants were students recruited at university campus. 40% never experienced VR before. We conducted the experiment in English.

3.3 Procedure

After the participants gave informed consent and received an introduction to VR, we demonstrated the game controls and interactions, briefly explained the mission, and how the agent(s) can help throughout the game. The experiment conductor was in the same room as the players for controlling the game and for observations. When the player started the game, the in-game agents introduced themselves shortly. For the multi-agent condition, the characters also mentioned their specialized task domain. After the participants finished both conditions, we conducted semi-structured interviews with the players examining game experience and interactions with the agents. Each test session took approximately 30 – 45 minutes with about 20 minutes in VR.

For the analysis of the qualitative data, we evaluated the comments and reactions collected during game play and in the conversational interview in the end. The data was examined for informative perspectives and concordance between the participants. In the following, we outline notable findings from our exploratory method.

3.4 Results

Overall, the participants liked interacting with the in-game characters vocally and were fond of the implementation of the in-game agents. Although we received contrasting comments about the two conditions, there was large agreement concerning the fun while playing: Nine out of ten participants found the multi-agent condition more entertaining. Participants mentioned that "it was more exciting and motivating". With the single character assisting, the game was perceived as "less fun". The players described that they enjoyed hearing different voices throughout the game. Consistently, two participants mentioned that it became monotone listening to the same voice in the single-agent condition. With the team of

agents supporting the player, one participant said: "It felt like I was a star in a movie". Moreover, we learned that players felt less alone while playing the multi-agent version. One person even stated: "I felt more protected". Working together with a multiplicity of agents was further described as "more professional" and like "teamwork". Participants confirmed the hypothesized sense of companionship and made comments such as: "It felt like we were a real team". Some participants signaled the desire for human-like behavior from the agents, for example that "it would be nicer if they could also talk with each other".

In the interviews, we identified that participants had difficulties in recalling the agents and their expertise. Not only did participants find the single-agent version less confusing, but also more efficient: "I was faster when it was just one agent". Another participant explained: "I do not care who does what, I need to get things done faster". In fact, the utterances were textually identical in both conditions and took the same response time. Moreover, seven out of ten participants mentioned that the single-agent assistant seemed more trustworthy. In the multi-agent version, one participant "felt more pressured" during game. This finding is in line with how "easy" and "reliable" the interaction with the single agent was described. To some, on the other hand, this was perceived as unremarkable or "boring". One player concluded: "So, nothing special".

4 DISCUSSION

The motivation behind the presented study was to provide players of a single-player VR game with a sense of team spirit and dependability familiar from cooperative multiplayer games. The qualitative data from our interviews show that our participants, indeed, perceived conversing with a multiplicity of assisting agents as being part of a team. They described the multi-agent cooperation as teamwork, more exiting and more motivating. We suppose that the supportive nature of the agents and how they actively contributed to succeeding in the game was important for the players to perceive it as teamwork. In contrast, merely commenting and decorative characters that do not influence the game progress might not have been successful in creating a sense of companionship.

Generally, players enjoyed interacting with the characters vocally. Nine out of the ten participants mentioned that they found the multi-agent version more entertaining. Like the support of team members in multiplayer gaming, the assistance from the multiple agents gave players the impression that they have their back and are approachable. The interaction created a feeling of protection for some players and made them feel less lonely. We assess designing the game for VR as expedient as we observed the players to be so immersed in the game environment that they talked freely with the agents, despite the lab setting.

On the other hand, some players preferred the simplicity of contact to a single agent and perceived to succeed faster in the game this way, even though the duration of the responses were the same in both versions. They experienced the multi-agent interaction as too complex and overstraining, which is in line with results from a study on a multi-chatbot system [7]. This seems conclusive considering the brevity of the exposure to the multi-agent system with roughly 10 minutes per session. Longer games allow players to get more familiar with the team members, so that the number of agents could be set even higher for a deliberately increased conversation complexity. Still, talking to only one character was experienced as easier and more reliable by some players. Therefore, we think it is crucial to enable the player to effortlessly predict and understand who the currently active interlocutor is.

In our study, we learned that some participants would like to see more human-like behavior from the agents, for instance, more exchange amongst them or humorous aspects. These findings imply that the participants perceived the agents as individuals with different opinions and personality, not as exchangeable entities of a computer system. This matches with earlier studies indicating that people find it easier to interact with technology that resembles human-like characteristics [5]. In contrast to the users' disapproval of conversations amongst multiple agents in a task-oriented setting [17], players of our game were fond of this idea which we explain with the hedonic purpose of the game.

4.1 Future Work

This was a first exploration on verbal multi-agent interaction in games. While the approach was successful in conveying a high degree of companionship, it is important that it does not compromise other aspects of the game experience. For instance, it is conceivable that the player could feel overpowered by the multitude of characters. The presence of several experts in specialized fields could impair the experienced competence of the player. A thorough follow-up study with standardized questionnaires is needed to quantify these potential effects on the player experience. We suggest that future research further considers the influence of player types, the current emotional and social state of the player, as well as individual preferences. Additionally, a comparison to a condition with no agent interaction at all might be informative. In our study, we learned that in the brief exposure, the participants had difficulties to distinguish the characters and their roles. We therefore recommend longer gaming sessions to foster social bonding. Further, we think it would be interesting to apply the multi-agent concept to multiplayer games as an addition to the team or as a substitution for dropped-out players as suggested by Pfau et al. [22]. Moreover, future work could transfer this concept to applications in other contexts, such as exposure therapy, education or training, and voice assistants in smart homes.

5 CONCLUSION

This study set out to recreate the social atmosphere of companionship from collaborative multiplayer gaming in a single-player game. In our VR game, the players were able to interact verbally with a team of in-game agents in natural language. In a preliminary study, we compared this approach to voice interaction with a single interlocutor. Our findings show that the participants indeed perceived interacting with the multiple agents as playing in a team. Further, players found it more entertaining, felt more motivated, as well as more protected when conversing with a group of characters.

ACKNOWLEDGMENTS

This work was partially funded by Klaus Tschira Foundation.

REFERENCES

[1] Fraser Allison, Marcus Carter, and Martin Gibbs. 2017. Word Play: A History of Voice Interaction in Digital Games. *Games and Culture* 15, 2 (2017), 91 – 113. https://doi.org/10.1177/1555412017746305

[2] Fraser Allison, Marcus Carter, Martin Gibbs, and Wally Smith. 2018. Design Patterns for Voice Interaction in Games. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play* (Melbourne, VIC, Australia) *(CHI PLAY '18)*. Association for Computing Machinery, New York, NY, USA, 5–17. https://doi.org/10.1145/3242671.3242712

[3] Fraser Allison, Joshua Newn, Wally Smith, Marcus Carter, and Martin Gibbs. 2019. Frame Analysis of Voice Interaction Gameplay. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300623

[4] BioWare. 2014. *Dragon Age: Inquisition*. Game [XBox One]. Electronic Arts, Redwood City, California, U.S.

[5] Cynthia Breazeal. 2003. Toward sociable robots. *Robotics and autonomous systems* 42, 3-4 (2003), 167–175.

[6] Marcus Carter, Fraser Allison, John Downs, and Martin Gibbs. 2015. Player Identity Dissonance and Voice Interaction in Games. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play* (London, United Kingdom) *(CHI PLAY '15)*. Association for Computing Machinery, New York, NY, USA, 265–269. https://doi.org/10.1145/2793107.2793144

[7] Ana Paula Chaves and Marco Aurelio Gerosa. 2018. Single or Multiple Conversational Agents? An Interactional Coherence Comparison. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173765

[8] Michael J Coulombe and Jayson Lynch. 2020. Cooperating in video games? impossible! undecidability of team multiplayer games. *Theoretical Computer Science* (2020).

[9] Crytek. 2013. *Ryse: Son of Rome*. Game [XBox One]. Microsoft Studios, Redmond, Washington, U.S.

[10] Steven Dow, Manish Mehta, Ellie Harmon, Blair MacIntyre, and Michael Mateas. 2007. Presence and engagement in an interactive drama. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1475–1484.

[11] EA Sports. 2013. *Fifa 14*. Game [XBox One]. Microsoft Studios, Redwood City, California, U.S.

[12] Tobias Greitemeyer, Eva Traut-Mattausch, and Silvia Osswald. 2012. How to ameliorate negative effects of violent video games on cooperation: Play it cooperatively in a team. *Computers in Human Behavior* 28, 4 (2012), 1465–1470.

[13] Mark D Griffiths, Mark NO Davies, and Darren Chappell. 2004. Demographic factors and playing variables in online computer gaming. *CyberPsychology & behavior* 7, 4 (2004), 479–487.

[14] Susumu Harada, Jacob O Wobbrock, and James A Landay. 2011. Voice games: investigation into the use of non-speech voice input for making computer games more accessible. In *IFIP Conference on Human-Computer Interaction*. Springer, 11–29.

[15] Martin Johansson, Gabriel Skantze, and Joakim Gustafson. 2014. Comparison of Human-Human and Human-Robot Turn-Taking Behaviour in Multiparty Situated Interaction. In *UM3I 2014*, Samer Al Moubayed, Dan Bohus, Anna Esposito, The NetherlandsHeylenDirk University of Twente, Maria Koutsombogera, Harris Papageorgiou, and Gabriel Skantze (Eds.). ACM Press, New York, New York, USA, 21–26. https://doi.org/10.1145/2666242.2666249

[16] Max Kinateder, Enrico Ronchi, Daniel Nilsson, Margrethe Kobes, Mathias Müller, Paul Pauli, and Andreas Mühlberger. 2014. Virtual reality for fire evacuation research. In *2014 Federated Conference on Computer Science and Information Systems*. IEEE, 313–321.

[17] Michal Luria, Samantha Reig, Xiang Zhi Tan, Aaron Steinfeld, Jodi Forlizzi, and John Zimmerman. 2019. Re-Embodiment and Co-Embodiment: Exploration of Social Presence for Robots and Conversational Agents. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (San Diego, CA, USA) *(DIS '19)*. Association for Computing Machinery, New York, NY, USA, 633–644. https://doi.org/10.1145/3322276.3322340

[18] Joseph Edward McGrath. 1984. *Groups: Interaction and performance*. Vol. 14. Prentice-Hall Englewood Cliffs, NJ.

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

[20] Aasish Pappu, Ming Sun, Seshadri Sridharan, and Alex Rudnicky. 2013. Situated Multiparty Interaction between Humans and Agents. In *Human-Computer Interaction. Interaction Modalities and Techniques*, David Hutchison, Takeo Kanade, and Josef Kittler (Eds.). Lecture Notes in Computer Science, Vol. 8007. Springer Berlin Heidelberg, Berlin/Heidelberg, 107–116. https://doi.org/10.1007/978-3-642-39330-3{_}12

[21] Jim R Parker and John Heerema. 2008. Audio interaction in computer mediated games. *International Journal of Computer Games Technology* 2008 (2008).

[22] Johannes Pfau, Jan David Smeddinck, Ioannis Bikas, and Rainer Malaka. 2020. Bot or Not? User Perceptions of Player Substitution with Deep Player Behavior Models *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3313831.3376223

[23] Pyro Studios. 1998. *Commandos: Behind Enemy Lines*. Game [Microsoft Windows]. Eidos Interactive, Southwark, London, England.

[24] D Reddy, Lee Erman, and R Neely. 1973. A model and a system for machine recognition of speech. *IEEE Transactions on Audio and Electroacoustics* 21, 3 (1973), 229–238.

[25] M.E. Shaw, R. Robbin, and J.R. Belser. 1981. *Group Dynamics: The Psychology of Small Group Behavior*. McGraw-Hill.

[26] Spellbound Entertainment. 2001. *Desperados: Wanted Dead or Alive*. Game [Microsoft Windows]. Atari, Paris, France.

[27] Adam J Sporka, Sri H Kurniawan, Murni Mahmud, and Pavel Slavík. 2006. Non-speech input and speech recognition for real-time control of computer games. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*. 213–220.

[28] Turn 10 Studios. 2013. *Forza Motorsport 5*. Game [XBox One]. Microsoft Studios, Redmond, Washington, U.S.

[29] Marco Filipe Ganança Vieira, Hao Fu, Chong Hu, Nayoung Kim, and Sudhanshu Aggarwal. 2014. PowerFall: a voice-controlled collaborative game. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play*. 395–398.

[30] Bob G Witmer and Michael J Singer. 1998. Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence* 7, 3 (1998), 225–240.

[31] Nima Zargham, Michael Bonfert, Tanja Döring, Robert Porzel, and Rainer Malaka. [n.d.]. Multi-Agent Voice Assistants: An Investigation of User Experience [under review].

Publication 5

An Evaluation of Visual Embodiment for Voice Assistants on Smart Displays

Michael Bonfert, Nima Zargham, Florian Saade, Robert Porzel, and Rainer Malaka

In Proceedings of the 3rd Conference on Conversational User Interfaces (CUI '21). New York, NY, USA, 2021. Association for Computing Machinery.

Personal contribution to this work: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, part of software, supervision, validation, visualization, and contribution to all parts of the manuscript.

ISBN: 978-1-4503-8998-3/21/07 DOI: 10.1145/3469595.3469611

An Evaluation of Visual Embodiment for Voice Assistants on Smart Displays

Michael Bonfert*
bonfert@uni-bremen.de
Digital Media Lab
University of Bremen, Germany

Nima Zargham*
zargham@uni-bremen.de
Digital Media Lab
University of Bremen, Germany

Florian Saade
florian.saade@uni-bremen.de
University of Bremen, Germany

Robert Porzel
porzel@uni-bremen.de
Digital Media Lab
University of Bremen, Germany

Rainer Malaka
malaka@uni-bremen.de
Digital Media Lab
University of Bremen, Germany

Figure 1: Our prototype of a smart display with an embodied voice assistant agent performed by an actress

ABSTRACT

Smart displays augment the concept of a smart home speaker with a touchscreen. Although the visual modality is added in this device variant, the virtual agent is still only represented through auditory output and remains invisible in most current products. We present an empirical study on the interaction of users with a smart display on which the agent is embodied with a humanoid representation. Three different conditions are compared in a between-group experiment: no agent embodiment, a digitally rendered character, and a photorealistic representation performed by a human actress. Our quantitative data do not indicate that agent visualization on a smart display affects the user experience significantly. On the other hand, our qualitative findings revealed differentiated perspectives by the users. We discuss potentials and challenges of embodying agents on smart displays, reflect on their continuous on-screen presence, present user considerations on their appearance, and how the visualization influenced the politeness of the users.

CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**; User studies.

KEYWORDS

Voice Assistants, Conversational Agents, Embodiment, Smart Displays

ACM Reference Format:
Michael Bonfert, Nima Zargham, Florian Saade, Robert Porzel, and Rainer Malaka. 2021. An Evaluation of Visual Embodiment for Voice Assistants on Smart Displays. In *3rd Conference on Conversational User Interfaces (CUI '21), July 27–29, 2021, Bilbao (online), Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3469595.3469611

*Both authors contributed equally to this research.

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

ACM ISBN 978-1-4503-8998-3/21/07…$15.00
https://doi.org/10.1145/3469595.3469611

1 INTRODUCTION

The use of voice interaction is spreading widely. Current voice assistants (VA) have broad capabilities in helping the users for different purposes, such as smart home control, work management, scheduling, gathering information, navigation, communication, education, or entertainment. Voice user interfaces are available in mobile phones, personal computers, cars, smart speakers, and other devices to make the interaction easier, more accessible, and more natural. The affordances and accessibility facilities of home assistants – also referred to as smart speakers – differ from those of VAs on smartphones. Interaction with home assistants is possible from a distance and enables to control smart home appliances [46]. With the device's exterior, the system has a physical embodiment within the room, however, not the assisting agent.

Research suggests that emulating human qualities affects how users feel towards VAs [11]. The experiences can be different between users depending on their own personalities [14], but also depends on the assistant's personality. Currently, the personality of a VA is primarily conveyed by its voice, linguistic characteristics of its answers, designated personifications as its name (e.g., Alexa instead of the product name Amazon Echo), and its physical device design [4, 5, 48]. Moreover, research has shown that the identified gender of an agent has an impact on the user experiences [7]. The visual presence of current smart speakers is limited to the device's casing and abstract animations or LEDs that illustrate the assistant's state or audio output. To visually convey personality and human characteristics, the development of virtual assistants could focus further on embodiment. Researchers have studied diverse types of embodiment for conversational agents [1, 24, 31] as well as their visual attractiveness [25]. Embodied virtual agents have become a natural extension of conversational interfaces by enriching the experience visually [2, 8, 43, 56].

A novel opportunity to embody VA agents are smart displays. This new product category of home assistants is equipped with a screen for visual output and touch input. Prominent examples from the consumer market are the Amazon Echo Show and the Google Nest Hub. These devices complement the features of a voice assistant with the possibilities to, for instance, look at pictures, watch videos, browse recipes, or display the smart front door camera.

Moreover, we see the potential in the screen to enhance the visual presence of the virtual agent. Therefore, we conducted a study on the user experience (UX) during the interaction with smart displays featuring an embodied agent. In this research, we pursue the following two research questions:

RQ1: How does the user experience change if a voice assistant agent is visually embodied on a smart display?

RQ2: How does the degree of visual realism of the embodied agent influence the user experience?

Building on prior research on attractiveness, gender, and appearance of embodied agents, we contribute an investigation on displaying an embodied VA agent and, moreover, its visual realism. This is done for the novel use case of a smart display considering social implications of the continuous agent presence in the room. We present an empirical study exploring user interactions when engaging with one of three different smart display prototypes: one with a disembodied agent, one with an artificial, digitally rendered embodiment, and one with a photorealistic embodied agent performed by a human actress. Our quantitative analysis includes two standardized UX questionnaires and the expressed politeness during the interaction. In semi-structured interviews, we collected further impressions, preferences and expectations by the participants.

2 RELATED WORK

In this section, we discuss research on agent embodiment with focus on voice assistants, the Uncanny Valley and gender implications.

2.1 Embodied Agents

Embodied conversational agents are computer-controlled characters that can interact with people using natural language and engage in a dialog [9]. They can use facial expressions, gestures, and eye gaze to enable natural, multimodal human-computer communication. Numerous studies have explored how embodiment and its different forms, as well as a lack of a body, can influence human-machine interaction and users' trust [13, 17, 20, 49] and engagement [26]. One of the most controversial examples of a virtual assistant with a visual embodiment is *Clippy*, an animated paper clip appearing in Microsoft Office 97. It was not well received amongst users and it failed to deliver on the promise of interface agents [57]. Research has found that using humanoid embodiment and voice influences users' perceptions of social presence [3, 47]. This presence of an agent can affect the relationship with a user in many aspects, such as trust and respect [3, 21].

Users treat the system more like a person when an agent has an embodiment [30, 55]. Castillo et al. believe that state-of-the-art embodied conversational agents can change their perceived personality through appearance and behavior [10]. An embodied agent can leverage various means of non-verbal communication to better engage with users beyond speech [56]. Previous work suggests that users' perception of an embodied VA's personality is not just dependent on its visual or audible output. Researchers believe that personality is experienced in a multimodal manner and if designers only focus on either voice or facial characteristics to design personality, they will most probably not succeed [10].

2.2 Voice Assistant Embodiment Across Applications

People feel higher levels of social presence when there is a visual representation available, as the comprehensive review on social presence literature by Oh et al. shows [44]. Hernández-Trapote et al. found that users who interacted with an embodied agent had greater privacy concerns but also perceived the interaction as more pleasant compared to using a voice-only interface. In their study, the authors found no significant difference in user preference [20]. In contrast to avatars depicting a specific person, an embodied agent can be designed in any conceivable way depending on the given context and purpose. Wang et al. conducted a study on interactions with virtual agents in augmented reality. They compared four agent representations: voice-only, non-humanoid, full-size humanoid, and miniature humanoid. The experiment showed that both humanoid and non-humanoid agents were acceptable for users. However, having an agent visualized as a smart speaker strongly impacted users'

conception of the agent not being human – even more than without visualization [56]. In virtual reality (VR) environments, Schmidt et al. showed major benefits for both embodied and thematically related audio-visual agent representations which positively affected the overall user experience in the context of a VR exhibition space. They also found that agent embodiment induces a higher sense of spatial and social presence [50].

With the aim to support information workers to be more productive and focused, Grover et al. designed and compared two productivity agents: a text-based agent, similar to a chatbot, and a virtual agent with a video embodiment. Their results show that users felt more productive and less distracted when being assisted by the embodied agent [16]. These findings are in line with a recent investigations on the effects of VA embodiment in augmented reality (AR). Kim et al. found that users performed better in collaborative decision making when interacting with a VA and reported a significantly lower task load when it was embodied [28]. Kim et al. further observed that users perceived agents in AR as more aware of and able to influence the real world if they are embodied [27]. Similar research has been done in virtual reality environments [50, 51].

The influence of human-like agent behavior was the focus of a study by Mayer et al. who assessed multimedia learning when being taught by on-screen agents. The team measured better performance in learning and recalling information when the agent behaved more like a human in speech and gestures [37]. The attractiveness of virtual agents has also been a topic of research. In a study by Khan and De Angeli, the users formed and maintained a better evaluation of attractive agents independent of the interaction with the agent [25]. It has been demonstrated that an agent's attractiveness may be even more important than its reliability [58].

2.3 Gender Implications

Researchers have extensively expressed their concerns on gendered agents as it can easily reproduce a stereotypical gender script [12, 54, 59]. Most of the common voice assistants available in the market set a female voice as default in most countries, which can amplify gender stereotypes [22]. A study by Nass et al. [42] suggests that even computers with minimized gender cues in the voice output evoke gender-based stereotypical responses. Authors tested three gender-based stereotypes without any gender indicators but vocal cues and witnessed stereotypes in all cases. In another study, Nass and Moon showed that users prefer to hear praises from a male agent rather than the same comments from a female agent [41]. Hwang et al. [22] categorized three distinct characteristics of bodily display, subordinate attitude, and sexualization to investigate the reflections of gender stereotypes toward women in female-voiced VAs. The authors suggest that such stereotypical traits could create a power dynamic between users and female agents. The described studies provide insights into the application of embodied agents across different mediums, use cases and characteristics. Our work extends research on embodied conversational agents to the domain of smart homes by bringing visualizations of a voice assistant to smart displays. Considering the large design space of possible agent visualizations, the question arises how close to a human appearance these should be. Thus, the investigation considers the degree of visual fidelity of the embodiment.

2.4 The Uncanny Valley

The term "uncanny valley" refers to a person's adverse reaction to robots that look and behave almost like a human, but not quite [39]. This effect has furthermore been investigated with any type of human-like entity or object, such as dolls, masks, facial caricatures, movie characters, avatars, and embodied agents [53]. Studies indicate that realistic humanoids can be appealing [18, 35, 38], but to achieve this, a number of aspects need to be considered. The artificial humanoid must attain a certain level of integrated social responsiveness and aesthetic refinement to appeal to the users [18]. Previous research has established that the uncanny valley effect emerges when there are abnormal features, or an insufficient degree of realism [53].

Some studies have explored the uncanny valley hypothesis in terms of human avatars [35, 38]. MacDorman et al. believe that a computer-generated face is not necessarily eeriest when it looks nearly human and argue that even abstract faces can look uncanny [35]. Guidelines for virtual character design by Schwind et al. recommend consistency in realism and deliberate stylization to avoid uncanniness [52]. To avoid uncanny valley effects in our study while still comparing cartoony to highly realistic embodiments of a VA agent, we decided to have an actress perform the agent for the photorealism condition. For most practical applications, this is obviously not an ecologically feasible solution, but provides clearer results in the context of this study.

3 PROTOTYPE DESIGN

We designed three versions of a smart display for the purpose of this experiment: one with a disembodied agent (DEA), one with a digitally rendered, artificial embodied agent (AEA), and one with a prerecorded, photorealistic embodied agent (PEA). All versions had the same functionality and only differed in appearance. We chose a female agent to reflect the predominance of female assistants in current consumer products with the intention to avoid a novelty bias [23]. The VA was called "Joy" and spoke the local official language German.

Disembodiment Agent (DEA) | This version was designed to resemble the current status quo of smart displays with no agent embodiment. The users would only hear the agent's synthetic voice. We generated the voice with the online Text-To-Speech (TTS) tool Natural Readers[1].

Artificial Embodied Agent (AEA) | For this version, we created a digitally rendered, animated visualization to represent the agent on the smart display. It shows a female, about 30-year-old character with blonde hair, light-colored skin and a dark blue dress as can be seen in Figure 2. The appearance reminds of a news anchor in the style of *The Sims*. We compared a variety of available options in an informal pre-study and found this character as best corresponding to the selected voice. To create the renderings, an actor performed in front of a webcam as input for FaceRig[2] to animate the virtual character. The video output was merged with the same TTS voice used for the DEA condition with synchronized lip movements.

[1]https://www.naturalreaders.com
[2]https://facerig.com

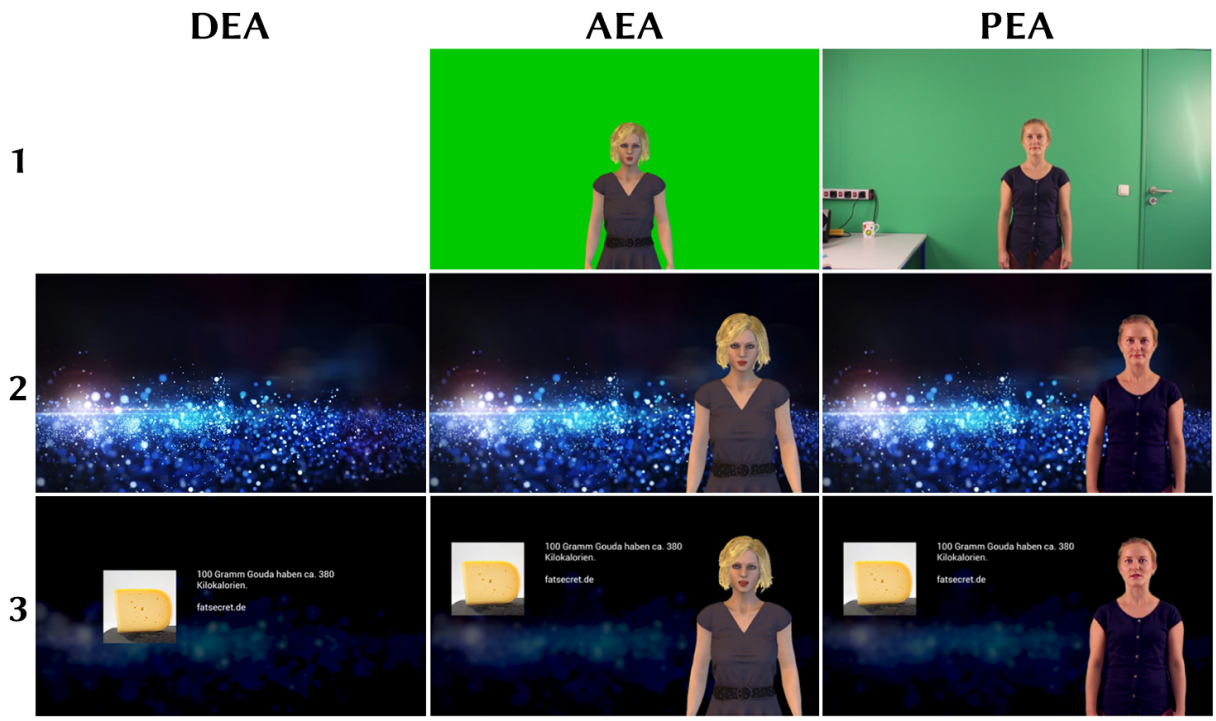


Figure 2: The three stages of the video creation process for the three conditions: 1) Recording or rendering with a green screen, 2) Replacing the background, 3) Augmenting with information cards. The second stage is used as idle loop.

Photorealistic Embodied Agent (PEA) | For this prototype version with a highly realistic embodiment, a theater actress was recorded. She was instructed to perform as similar as possible to the artificial character in terms of intonation, facial expressions, and body language. We refrained from using the TTS audio on account of lip synchronicity and to avoid a mismatch of visual and auditory coherence. The actress and her clothing were selected to resemble the AEA visually. All utterances were recorded in front of a green screen to be used as an overlay for the content.

3.1 Prototype Implementation

For the video and audio output of the smart display, we prepared media snippets of all responses needed for the experiment execution in each condition. Each snippet consisted of a dark, dynamic background, an information card, the audio track, and – where applicable – an agent embodiment, as illustrated in Figure 2. The information cards contained text and images related to the user's commands. They appeared when the assistant initiated the response and faded out when the task was performed. Between tasks, a dynamic idle video was looping. The smoothness of the transitions depended on the timing of the next inquiry, which affected all conditions equally. For the AEA and PEA conditions, the agent embodiment was added as an overlay on the bottom right without overlapping the information cards. The screen layout in the DEA condition was centered to avoid empty space where the agent would be shown in the other versions.

To ensure reliable system operability, we used a Wizard of Oz approach in this study. The Wizard sat in an adjacent room and controlled the smart display. This was disclosed to the participants after the study. The technical setup is illustrated in Figure 3. The prototype was assembled from a Nexus 7 tablet and a Bluetooth speaker. For mounting the components in a way to appear as a smart speaker, three tailored parts were manufactured with 3D printing and laser-cut acrylic glass. The Wizard listened to the user's commands via Skype which was running silently on the tablet in the background. The responses were triggered with the help of a structured playlist on VLC media player to provide an instantaneous responsiveness of the system. Via Splashtop, the video on the Wizard's laptop was streamed to the tablet. Until the user continued with the next inquiry, the system looped an idle sequence that continuously showed the agent. For the analysis of the user's language, an audio device within the room recorded the experiment. For realizing one of the experiment tasks concerning smart home appliances, we used a smart light bulb by Philips Hue activated with a remote control by the Wizard.

4 EXPERIMENT

We evaluated our prototypes in a Wizard-of-Oz experiment with a between-groups design in which the participants ($N = 60$) interacted with one of the prototypes to complete a specified set of tasks. The condition assignment was pseudo-randomized between three equally distributed groups of 20 users each.

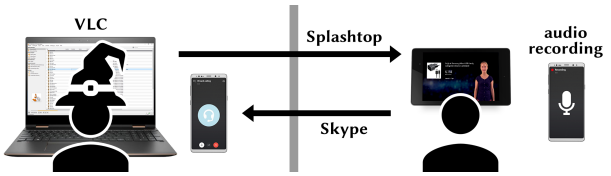


Figure 3: Technical setup of the experiment: The Wizard listens to the user's commands via Skype and triggers the appropriate video snippet in VLC media player which is transferred to the smart display via Splashtop. An audio device records the experiment.

4.1 Participants

We recruited 60 participants for this experiment (47 male, 13 female), between 14 and 38 years of age (M = 24.0, SD = 5.5). About half of the participants were students and 42% had a computer science background. All participants owned a smartphone. The majority of our participants (71.7%) stated that they rarely use a voice assistant on their phone. 11.7% never used a VA once. The other 16.7% use it at least several times a week. Concerning VAs in smart homes, 16.7% indicated using smart speakers regularly. The groups of phone VA users and smart speaker users have an overlap but are not identical. Only one participant had prior experience with a smart display and uses it daily. We conducted the experiment in the local official language German to avoid language barriers.

4.2 Procedure

After giving informed consent, all participants filled in a questionnaire about demographics and prior VA experience. Afterwards, the participants watched a short tutorial video on a separate screen outlining eleven predefined tasks to perform. We provided the participants with a paper list of the tasks to accomplish. Then, the test began with the smart display showing the idle video sequence, including the embodied agent if applicable. The first interaction was initiated by the user.

The activities represent a morning scenario and were designed to include a broad range of everyday commands following an analysis of typical home assistant usage [29]. These included, for example, turning on the light, playing music, retrieving information, setting a timer, or ordering a product online. All tasks are listed in the Appendix A. Sometimes, the participants forgot to use the wake word yielding in no reaction of the VA. When the user asked questions that were not included in the command list, the system explained that it cannot help with this. To ensure comparable interaction experiences and levels of frustration, one simulated failure to comply was included in each session even when a participant followed the task list strictly.

After finishing the tasks, the participants filled in a paper questionnaire comprising the User Experience Questionnaire (UEQ) [32] and the AttrakDiff Short Questionnaire [19]. Both scales are validated and established measurement instruments with a similar underlying theoretical construct to assess the pragmatic and hedonic qualities as well as the attractiveness of a system. The questionnaires provide an authoritative, quantitative measure of the user's subjective experience. In combination, the collected data can be compared to confirm the reliability of the measurements. Finally, the experimenter conducted a brief semi-structured interview covering aspects of reliability, trust, agent appearance, individual preferences, and permanent on-screen presence. At the end, all participants were demonstrated the alternative system versions to allow a comparison, despite the between-groups approach. This was done last in the interviews to not influence any prior assessments. Everyone was shown the same, complete sample snippet from both unfamiliar conditions to ensure comparability. The experiment and interview were recorded acoustically for later analysis. Each test session took 30 – 50 minutes.

4.3 Data Analysis

Two participants gave contradictory answers within three or more scales of the UEQ. As recommended by the handbook, their ratings were excluded from the analysis as it can indicate random or not serious answers [32]. Further, one participant did not fill in the AttrakDiff. For both questionnaires, the visual interpretation of the histograms raised doubts about the normal distribution of the data. This assumption was supported by Shapiro-Wilk tests. Therefore, we applied non-parametric tests. We ran Kruskal-Wallis tests to check for group differences between the three conditions. Due to technical issues, only n_{quant} = 50 audio recordings of the experiment sessions were complete and valid for statistical analysis. The unequal distribution between the conditions (DEA: 19, AEA: 16, PEA: 15) was considered for the statistics. The number of "Thank you" and "Please" utterances per user was compared with Mann-Whitney U tests between the groups. For all statistical tests, we applied an alpha level of .05.

Regarding the qualitative data, three interviews could not be analyzed due to data loss from a defective SD card. The other n_{qual} = 57 interview recordings were systematically examined (DEA: 20, AEA: 18, PEA: 19). Three researchers agreed on a coding system that was generated from a random selection of ten interviews. Then, all recordings were analyzed, coded along this categorization, and summarized. Additionally, we collected insightful and unique statements.

5 RESULTS

We present our findings in three sections: quantitative system evaluation, suggestions for the visual appearance, and considerations regarding the permanent presence of the agent.

5.1 Quantitative System Evaluation

From the standardized questionnaires, we learn that all three conditions, with a disembodied agent (DEA), with an artificial embodied agent (AEA), and with a photorealistic embodied agent (PEA), can result in comparably good user experiences. For all groups, the User Experience Questionnaire, rated from −3 to +3, shows overall high ratings for *attractiveness* ($Mean_{DEA}$ = 1.38 ± Standard Deviation$_{DEA}$ = .61; M_{AEA} = 1.12 ± .92; M_{PEA} = 1.58 ± .81) and the *pragmatic qualities* (M_{DEA} = 1.67. ± .59; M_{AEA} = 1.43 ± .68; M_{PEA} = 1.71±.46). The scores of the *hedonic qualities* (M_{DEA} = .84± .72; M_{AEA} = .81 ± .94; M_{PEA} = .99 ± 1.11) are below average according to the UEQ handbook. The data distribution of the single subscales yielding in these aggregated scores are illustrated in the

Table 1: Statistics on the UEQ and AttrakDiff analysis with Kruskal-Wallis H and asymptotic significance p for the subscales Attractiveness, Pragmatic Qualities, and Hedonic Qualities.

UEQ	**Attr.**	Perspicuity	Efficiency	Dependability	**Prag. Q.**	Stimulation	Novelty	**Hed. Q**
H	3.585	1.721	3.023	1.268	1.200	0.306	2.954	1.358
p	.167	.423	.221	.530	.549	.858	.228	.507

AttrakDiff	**Attr.**	**Pragmatic Qualities**	**Hedonic Qualities**
H	1.131	1.888	0.437
p	.568	.389	.804

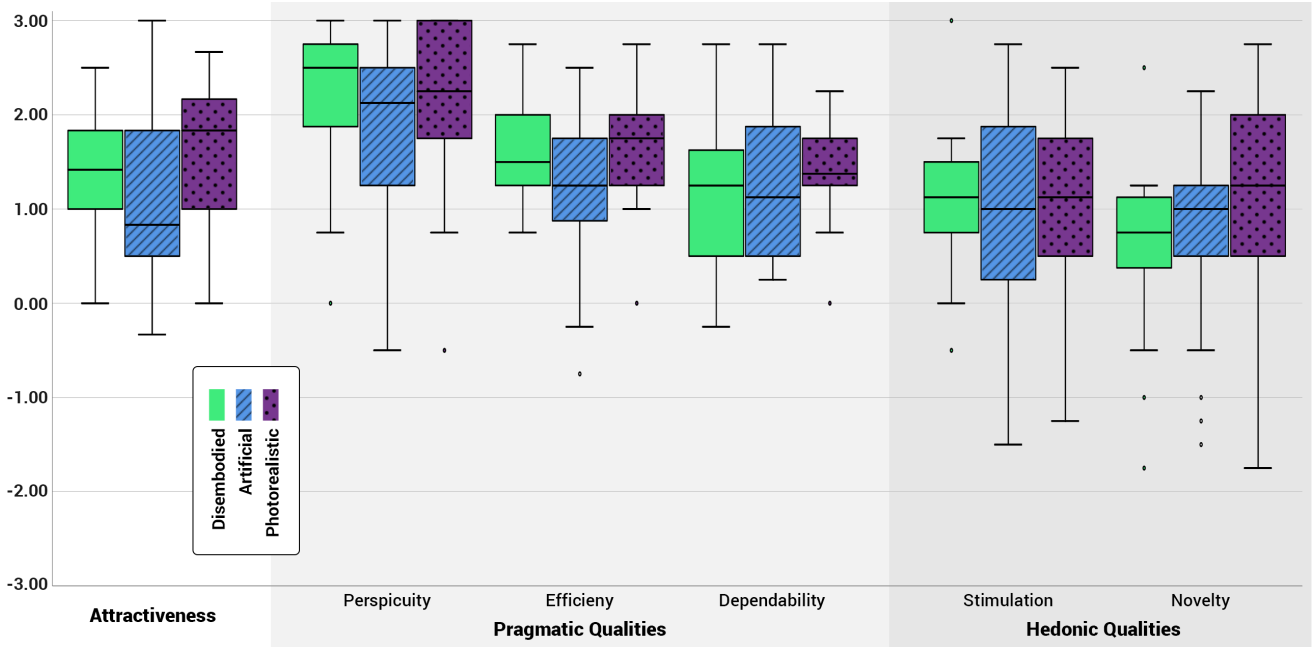


Figure 4: A box plot showing the distribution of ratings along the six subscales of the User Experience Questionnaire (UEQ) comparing the three conditions *Disembodiment Agent* (green), *Artificial Embodied Agent* (blue), and *Photorealistic Embodied Agent* (purple)

box plot in Figure 4. The data show no significant differences between the conditions on any of the subscales or aggregated scores ($p > .05$) as shown in Table 1.

These measurements match the user ratings on the AttrakDiff Short Questionnaire. In line with the UEQ results, no group differences were observed ($p > .05$). This measurement tool classified all our tested systems clearly as "task-oriented" due to high ratings for *pragmatic qualities* ($M_{DEA} = 1.63 \pm .80$; $M_{AEA} = 1.49 \pm .72$; $M_{PEA} = 1.78 \pm .83$) and medium ratings for *hedonic qualities* ($M_{DEA} = .79 \pm .73$; $M_{AEA} = .69 \pm .76$; $M_{PEA} = .75 \pm 1.12$). Like the UEQ, the AttrakDiff evaluates the system's *attractiveness* and yields a similar outcome with medium to high ratings ($M_{DEA} = 1.48 \pm .83$; $M_{AEA} = 1.15 \pm .90$; $M_{PEA} = 1.24 \pm 1.11$).

After the experiment, we showed the participants how the other two system versions look like and asked them to choose their preferred version. More than half decided for the photorealistic agent (51.8%). Only one out of eight users would select the artificial agent (12.5%) and every third person favored the version with a disembodied agent (35.7%). We found a bias in preference for the system version that the user was familiar with from the experiment, especially pronounced for the AEA condition: 71% of the people who favored the artificial embodiment in the interview used it earlier in the study (expected value: 33%). Only two people from another condition preferred the artificial variant. Moreover, 83% of the users who worked with the photorealistic agent embodiment preferred this version. Similarly, 60% of the participants from the DEA condition preferred to have no agent embodiment.

From recordings, we analyzed the users' verbal input in terms of expressed courtesy towards the VA. Overall, 28% of the users said "please" in at least one of their inquiries with no significant differences between the conditions. However, while 47% thanked the disembodied agent and 44% said "thank you" or "thanks" to the

artificial agent, only 13% expressed thankfulness toward the photo-realistic agent. With an average of $M = 0.13$ thankful utterances per session, participants in the PEA condition used significantly less thankfulness indicators than the DEA users with $M = 0.95$ ($U = 91.0$, $p = .030$, $\eta^2 = .094$). The difference in comparison to the AEA users ($M = 0.75$) is borderline but not statistically significant ($U = 80.5$, $p = .050$, $\eta^2 = .079$).

5.2 Agent Appearance

When we asked our participants about the ideal appearance of the agent, we observed attitudes that can be roughly categorized as pragmatic, personal, and playful. Firstly, the pragmatic group argued for omitting any agent visualization as there is no functional purpose of it in voice interaction. It was described as an unnecessary distraction taking up space, that could be used to display important content. If the agent was supposed to be embodied, users in this group would typically prefer a non-humanoid appearance. As abstract representations, they proposed eyes, an emoji, or minimalist animations such as waves, dots, or a point cloud. A user argued that it should be visually clear that the interlocutor is a machine and not a human. For this, a robot was suggested.

Secondly, the group with a preference for a more personal interaction were in favor of a human-like embodiment. A typical reason for this was that it is perceived as more trustworthy and more natural. However, three users were concerned about the authenticity of a digitally rendered visualization. It was perceived as "creepy" as it looked "not human enough" (P23), for example due to the lack of gestures. Seven participants were in support of a cartoon style. Three participants would like to be assisted by "an attractive woman" and one even specified the preferred hair color. For almost half of our participants (45%), the gender of the agent does not matter. Most of the rest would rather have a female (42%) than a male agent (13%). Several users (22 of 57=39%) explained that the agent should ideally be of similar age as themselves. It should not seem "too young, so it is reliable" (P34) and knowledgeable. Others were concerned about the agent being much older than themselves because it might make them feel parented or patronized.

Thirdly, the playful users proposed creative and fun ideas for the agent embodiment. These included animals, dinosaurs, and fantasy creatures. 13 participants requested celebrities, such as musicians, actors, or athletes, but also fictional characters from pop culture, such as Spiderman, Darth Vader, Pokémon, Hermione, Dobby, Rick and Morty, Yoshi, or – as "someone who fits into this role" (P35) – Batman's butler Alfred. Even a modern adaptation of Microsoft Office's *Clippy* was proposed. One user suggested to show the user's self-avatar as the agent. Moreover, someone proposed changing characters for specialized task areas, e.g., a depiction of a grandmother for recipes. One person advocated gender-neutral solutions to not further increase the bias in the perception of children, that the typical assistant should be female.

5.3 Permanent Presence of the Agent

Several users (12 of 57=21%) appreciated that the agent was always visible – also in the idle state. It was perceived as steady availability of the system: while the agent is present it can obviously be addressed at any time. In contrast, the majority (63%) of the users in our sample expressed that they would like the assistant to disappear from the screen after a task was performed and only reappear when called upon. Most often, this was explained with the awkward feeling of being watched by the agent. One user described the impression that "the device is alive" when there is an agent starring at him (P45). Nine users found it unsettling that the assistant seems to be waiting for them: "it feels like [the agent] expects something" (P33). Another participant was concerned that "when there is a human [agent] idling around, it would be very creepy" (P38). Six participants would appreciate the transition as an indicator for the successful recognition of the activation word. By some users, the agent's unchanged presence was misinterpreted as a permanent responsiveness. This led to misunderstandings in which the users continually omitted the activation word and were frustrated by the lack of feedback.

Five users (9%), who prefer the agent to disappear when idle, speculated about the design of the transition. For P12, it is important to avoid a sudden disappearance because in reality, people do not suddenly vanish. Similarly, one participant proposed a reality-inspired design in which the agent would walk in and out of the frame as needed. User P37 suggested a humorous adaption of this idea. She would like if the agent occasionally walked through the screen as when passing by, or read a newspaper while not needed.

6 DISCUSSION

In this study, we set out to understand the UX during the interaction with different visual representations of a smart display agent (RQ2) and compared it to a system with a disembodied agent (RQ1). With the two standardized UX questionnaires (UEQ and AttrakDiff Short), no significant differences between the conditions were found in terms of pragmatic qualities, hedonic qualities, or attractiveness of the systems.

Considering the qualitative findings, however, it is evident that an embodied agent does influence the interaction in various ways, beyond the measurements of the standardized instruments that we applied. The discrepancy between quantitative and qualitative results might be due to the broad range of UX aspects that the universally applicable questionnaires cover – which were found to be similar in all conditions during the short-term usage in our lab experiment – while the insights from the interviews mostly concerned social context, the imagined usage in a home environment, as well as design speculations specific to smart displays. These findings could hardly be brought to light with standardized scales but provide exciting avenues for future research. In the following, we will discuss what aspects are promising for a future, more targeted quantitative examination.

6.1 Embodied vs. Disembodied Agent

A third of our sample (35.7%) would prefer to use the status quo of a smart display with no depiction of the agent. Reasons for not showing a visualization were mostly of pragmatic nature. For a voice user interface, it was regarded as unnecessary, distracting, and blocking space that could be used for more relevant content. However, the pragmatic qualities of all systems yielded similar ratings and did not reveal advantages of not displaying an agent

regarding how efficient, clear, fast, or predictable the system was perceived.

Every second participant (51.8%) preferred the version with a photorealistic agent. Although we observed a tendency towards preferring the system familiar from the experiment, especially noticeable in the AEA condition, only one out of eight (12.5%) users favored the artificial visualization. This is in line with results by Hernández-Trapote et al. [20] who assume a "balance of likeability and rejection factors" causing ambivalent preferences concerning agent embodiment. Similarly, our results demonstrate that embodiment is not beneficial, in principle, but depends on its implementation.

One of the key advantages of conversing with an embodied agent was explained with its higher subjective trustworthiness, supporting the findings of previous works [17, 49]. Similarly, in accordance with existing research [30, 55], the interaction was perceived as more natural with an embodiment. The possibility to see the agent seems to make it more approachable and dependable compared to only hearing the volatile voice. While for some users it was only important to visually focus on the interlocutor independent of its appearance, others had clear ideas of how it should look like. A group of users explained that it should be obvious that they are talking to a machine, so they are not led to believe that they are speaking with a human. For this, abstract and non-humanoid representations were suggested. Further, several playful and fun concepts were shared by the participants, including fictional characters, animals, or mythical creatures.

Overall, we learned a wide variety of conflicting preferences and reasons and can therefore assume that no universal solution will satisfy the expectations of all users equally. Consequently, we recommend enabling smart display users to determine whether an embodied agent is displayed and to customize its appearance to their liking.

6.2 Humanoid Appearance of the Agent

Among the participants who like the idea of seeing the agent, only a fifth preferred the prototype version with artificial, cartoon-style rendering. The users were skeptical about the artificial embodied agent as it was described as "not human enough" indicating an uncanny valley effect [40]. Indeed, the visualization in our AEA condition was technically not sophisticated, for example, due to the lack of gestures or detailed micromotions which influence the perceived humanness [37]. On the other hand, some users liked the deliberate cartoony realization, because the humanoid shape conveys a human-like conversation style, while the style maintains the obvious artificiality of the interlocutor.

80% of the users, who were in favor of displaying the agent, liked the photorealistic embodied agent the most. This is a notable outcome considering the mismatch of visual realism and behavioral artificiality. As a meta-analysis shows, the literature describes various differences in the perceived social influence of human-controlled avatars compared to computer-controlled agents independent of their degree of visual realism [15]. A factor that might have effected the participants' preference for a specific version might have been the different voices used in the conditions. We used a computer-generated voice using a TTS tool for the DEA and AEA conditions, and recorded a human voice for the PEA condition. This design decision was made to keep the system variants as coherent as possible in terms of audio-visual match and synchronicity. This consistently conveys to the user that one version is entirely artificial and the other as realistic as possible at the cost of using different voices.

Of course, recording actors to embody agents that are meant for universal application is not feasible outside a Wizard of Oz experiment. However, our results clearly indicate a preference of life-like realism over uncanny renderings or cartoony stylization. We, therefore, advocate for sophisticated, photorealistic renderings or alternatively fully abstract visualizations, depending on the requirements, target group, and objectives of the system. Outside of professional context, entertaining approaches with funny and fictional characters can be considered as an additional option for consumer products. Offering famous characters from pop culture could appeal to fans and create a fun experience.

Nearly half of our participants reported that the agent's gender is not important for them. Three quarters of users, who stated a preference, prefer to talk to a female agent which supports previous literature on gender stereotypes with conversational agents [7, 22]. This could be explained with habituation as participants mentioned that they are used to female VAs, which is still the default setting in most popular consumer products. Another explanation could be the sample skew toward male participants (78.3%). We also observed a sexualized component in the relationship to the agent, as a few participants wanted to have an attractive agent with customizable appearance, e.g., preferred hair color. These findings align with previous research by Khan and De Angeli regarding the attractiveness of embodied agents [25].

With an already pronounced gender bias in the VA market and a clear status imbalance in the interaction, we advocate for gender-aware solutions. While gender-ambiguous voices could be an apparent solution, Sutton argues that also other factors than voice can lead to binary assumptions on an agent's gender making the voice ambiguity redundant [54]. This becomes especially evident for embodied VAs and requires careful considerations for gender-sensitive interfaces. An alternative approach could be a balanced team of agents with various genders co-embodying the smart display [34].

Our findings showed that the assumed age of the agent could affect the user's assessment of reliability, which corresponds with findings by Marin et al. [36]. According to the participants, the agent should appear experienced, hence should not be too young. On the other hand, younger users explained that they do not want the agent to look much older than themselves since they could feel patronized or mothered. Participants described the ideal age as similar to their own.

6.3 Social Aspects and Awkward Presence

We observed that the participants expressed significantly less thankfulness towards the PEA compared to the DEA and AEA. This finding might seem counter-intuitive as we could expect more courtesy in interactions with a more realistic assistant. Contrasting views among users have been observed whether a voice assistant is entitled to politeness [6]. Another explanation could be the perceived real-time processing of the system. The reactions by the actress must have been recorded before the interaction; hence, the

agent cannot rejoice in the expressed thankfulness. The user might assume a predefined emotional state that cannot be influenced. Whereas, the artificial and disembodied agents are experienced as "live" and capable to adapt to the user's politeness during runtime.

The continuous display of the AEA and PEA agents was interpreted differently by our participants. For some, it was an indicator that the system is online and ready. Others assumed that the device would be listening to commands non-stop. Consequently, they omitted the wake word and were irritated by being ignored from the attentive-looking agent. Moreover, users expressed discomfort of the agent starring at them during idle time. We noticed that the permanent presence of the agent leads to a feeling of constantly being observed by it. Some even felt like the agent would be waiting impatiently for the user until assistance is needed again.

We recommend hiding the agent between tasks to avoid social awkwardness and domestic intrusion. The reappearance can serve as feedback to indicate that the system recognized the wake word and is listening to commands. As another advantage of temporarily hiding the embodiment, other agents have the opportunity to re-embody the device, for example, for handling different task domains or assisting different users [34]. The transition could be implemented either as a fading effect or, for instance, with the agent walking in and out. In a playful context or in situations that require a continuous indicator of availability, such as for interactive public displays, the agent could alternatively be always visible but suggestive of being distracted. One participant proposed the agent being occupied reading a newspaper. As this might suggest to the user that the agent is busy and unavailable, we recommend trying more subtle deflections, for example letting the agent's gaze wander to the sides of the screen.

7 LIMITATIONS AND FUTURE WORK

The lower popularity of the prototype with the AEA is not necessarily due to the nature of digital rendering but might stem from our specific implementation. Users criticized the "creepy" appearance of the artificial agent and its lack of gestures. We assume that a higher technical sophistication could have improved its humanness. For this study, however, it was a deliberate design decision to compare different levels of realism. To exclude uncanniness effects, it would be insightful to replicate the presented experiment with a highly realistic rendered agent compared to a live-recorded human agent for distinguishing between effects from visualization and agency. Although the user experience might have benefited if the PEA had more human-like behaviour in terms of intonation, gestures, or facial expressions, we decided to match the artificial agent closely for higher comparability. For this study, we aimed for consistent agent realization, hence, we gave the AEA a computer-generated voice and the PEA a real human voice. Future work could investigate the impact of the AEA having a human voice and compare it to the PEA condition with a human voice.

We provided our participants with a predefined list of tasks. Some tested the system capabilities by asking additional questions. The VA responded that it cannot assist with that inquiry. Potentially, this could reduce the ecological validity. On the other hand, advantages of this established method for dialog system testing [45] are a structured procedure with high comparability and a feasible response preparation, as the questions are predictable. Further, the experiment covered mostly simple commands. More complex interactions, such as multi-step conversations, could also be investigated in future studies.

The experiment sample was skewed toward young ($M = 24.0$) men (78%) with a computer science background (42%). We cannot exclude an influence of this bias on the results concerning the agent's gender and age, or the participants' affinity towards technological innovations. Prior research suggested an effect of technical knowledge on the interactions with VAs [33]. Moreover, we decided to follow the conventional industry default of a female agent to avoid gender novelty as confounding factor. Since the focus of this research was on realism and the resulting artificiality of the agent, and not on gender comparisons, we did not further investigate this aspect with additional conditions. Therefore, we suggest future research to look at users' preferences for agents of different gender.

As the present study investigated short-term effects in a lab environment, it would be interesting to compare the results to a long-term exploration of different agent embodiments in a home setting – especially in terms of social presence, privacy concerns, and emotional bonding with the agent.

8 CONCLUSION

In this paper, we set out to understand the potentials and challenges of introducing agent embodiment to smart displays and compared different degrees of visual realism. Our contribution builds on a between-groups study with 60 participants. Using a Wizard of Oz method, we compared three conditions: no agent embodiment, artificial embodiment and photorealistic embodiment. In the quantitative system evaluation, we found similar user experience ratings across all conditions. Yet, the users had clear preferences and provided valuable insights on their views about the visualization and the permanent agent visibility. Moreover, we unexpectedly observed that the users were less polite towards the agent with photorealistic appearance. Our work identifies critical design considerations on how to embody voice assistant agents on smart displays to achieve a higher user satisfaction. The findings also provide orientation for researchers to quantitatively examine embodied smart display agents with targeted measurements.

ACKNOWLEDGMENTS

This work was partially funded by Klaus Tschira Foundation, by the FET-Open Project 951846 "MUHAI – Meaning and Understanding for Human-centric AI" funded by the EU program Horizon 2020, as well as the German Research Foundation DFG as part of Collaborative Research Center (Sonderforschungsbereich) 1320 "EASE – Everyday Activity Science and Engineering", University of Bremen (http://www.ease-crc.org/) conducted in subproject H02.

REFERENCES

[1] Elisabeth André. 2011. Design and evaluation of embodied conversational agents for educational and advisory software. In *Gaming and Simulations: Concepts, Methodologies, Tools and Applications*. IGI Global, Hershey, Pennsylvania, USA, 668–686.

[2] Sean Andrist, Michael Gleicher, and Bilge Mutlu. 2017. Looking Coordinated: Bidirectional Gaze Mechanisms for Collaborative Interaction with Virtual Characters. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*

(Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2571–2582. https://doi.org/10.1145/3025453.3026033

[3] Wilma A Bainbridge, Justin Hart, Elizabeth S Kim, and Brian Scassellati. 2008. The effect of presence on human-robot interaction. In *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, IEEE Press, NJ, USA, 701–706.

[4] Erin Beneteau, Olivia K. Richards, Mingrui Zhang, Julie A. Kientz, Jason Yip, and Alexis Hiniker. 2019. Communication Breakdowns Between Families and Alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article 243, 13 pages. https://doi.org/10.1145/3290605.3300473

[5] Timothy W. Bickmore and Rosalind W. Picard. 2005. Establishing and Maintaining Long-term Human-computer Relationships. *ACM Trans. Comput.-Hum. Interact.* 12, 2 (June 2005), 293–327. https://doi.org/10.1145/1067860.1067867

[6] Michael Bonfert, Maximilian Spliethöver, Roman Arzaroli, Marvin Lange, Martin Hanci, and Robert Porzel. 2018. If you ask nicely: A digital assistant rebuking impolite voice commands. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction : ICMI'18 : Boulder, CO, USA, October 16 - 20, 2018*. ACM, New York, NY, 95–102. https://doi.org/10.1145/3242969.3242995

[7] Sheryl Brahnam and Antonella De Angeli. 2012. Gender affordances of conversational agents. *Interacting with Computers* 24, 3 (2012), 139–153.

[8] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. 1999. Embodiment in Conversational Interfaces: Rea. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 520–527. https://doi.org/10.1145/302979.303150

[9] Justine Cassell, Joseph Sullivan, Elizabeth Churchill, and Scott Prevost. 2000. *Embodied conversational agents*. MIT press, MA, USA.

[10] Susana Castillo, Philipp Hahn, Katharina Legde, and Douglas W. Cunningham. 2018. Personality Analysis of Embodied Conversational Agents. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (Sydney, NSW, Australia) *(IVA '18)*. Association for Computing Machinery, New York, NY, USA, 227–232. https://doi.org/10.1145/3267851.3267853

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

[12] Andreea Danielescu. 2020. Eschewing Gender Stereotypes in Voice Assistants to Promote Inclusion. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (Bilbao, Spain) *(CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 46, 3 pages. https://doi.org/10.1145/3405755.3406151

[13] Munjal Desai, Kristen Stubbs, Aaron Steinfeld, and Holly Yanco. 2009. Creating trustworthy robots: Lessons and inspirations from automated systems. In *Proceedings of AISB Convention: New Frontiers in Human-Robot Interaction*. Society for the Study of Artificial Intelligence and the Simulation of Behaviour, Bath, UK, 49–56.

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

[15] Jesse Fox, Sun Joo (Grace) Ahn, Joris H. Janssen, Leo Yeykelis, Kathryn Y. Segovia, and Jeremy N. Bailenson. 2015. Avatars Versus Agents: A Meta-Analysis Quantifying the Effect of Agency on Social Influence. *Human–Computer Interaction* 30, 5 (2015), 401–432. https://doi.org/10.1080/07370024.2014.921494

[16] Ted Grover, Kael Rowan, Jina Suh, Daniel McDuff, and Mary Czerwinski. 2020. Design and Evaluation of Intelligent Agent Prototypes for Assistance with Focus and Productivity at Work. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 390–400. https://doi.org/10.1145/3377325.3377507

[17] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* 53, 5 (2011), 517–527.

[18] David Hanson, Andrew Olney, Steve Prilliman, Eric Mathews, Marge Zielke, Derek Hammons, Raul Fernandez, and Harry Stephanou. 2005. Upending the uncanny valley. In *AAAI*, Vol. 5. ACM, New York, NY, USA, 1728–1729.

[19] Marc Hassenzahl, Michael Burmester, and Franz Koller. 2003. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität [AttracDiff: A questionnaire to measure perceived hedonic and pragmatic quality]. In *Mensch & Computer 2003*, Gerd Szwillus and Jürgen Ziegler (Eds.). B. G. Teubner, Stuttgart, 187–196.

[20] Álvaro Hernández-Trapote, Beatriz López-Mencía, David Díaz, Rubén Fernández-Pozo, and Javier Caminero. 2008. Embodied Conversational Agents for Voice-Biometric Interfaces. In *Proceedings of the 10th International Conference on Multimodal Interfaces* (Chania, Crete, Greece) *(ICMI '08)*. Association for Computing Machinery, New York, NY, USA, 305–312. https://doi.org/10.1145/1452392.1452454

[21] Guy Hoffman, Jodi Forlizzi, Shahar Ayal, Aaron Steinfeld, John Antanitis, Guy Hochman, Eric Hochendoner, and Justin Finkenaur. 2015. Robot Presence and Human Honesty: Experimental Evidence. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (Portland, Oregon, USA) *(HRI '15)*. Association for Computing Machinery, New York, NY, USA, 181–188. https://doi.org/10.1145/2696454.2696487

[22] Gilhwan Hwang, Jeewon Lee, Cindy Yoonjung Oh, and Joonhwan Lee. 2019. It Sounds Like A Woman: Exploring Gender Stereotypes in South Korean Voice Assistants. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. ACM, New York, NY, USA, Article LBW2413, 6 pages. https://doi.org/10.1145/3290607.3312915

[23] Gilhwan Hwang, Jeewon Lee, Cindy Yoonjung Oh, and Joonhwan Lee. 2019. It Sounds Like A Woman: Exploring Gender Stereotypes in South Korean Voice Assistants. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312915

[24] Katherine Isbister and Patrick Doyle. 2002. Design and evaluation of embodied conversational agents: A proposed taxonomy. In *The first international joint conference on autonomous agents & multi-agent systems*. ACM, NY, USA.

[25] Rabia Khan and Antonella De Angeli. 2009. The attractiveness stereotype in the evaluation of embodied conversational agents. In *IFIP Conference on Human-Computer Interaction*. Springer, Springer, Heidelberg, Germany, 85–97.

[26] Sara Kiesler, Aaron Powers, Susan R Fussell, and Cristen Torrey. 2008. Anthropomorphic interactions with a robot and robot–like agent. *Social Cognition* 26, 2 (2008), 169–181.

[27] K. Kim, L. Boelling, S. Haesler, J. Bailenson, G. Bruder, and G. F. Welch. 2018. Does a Digital Assistant Need a Body? The Influence of Visual Embodiment and Social Behavior on the Perception of Intelligent Virtual Agents in AR. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE Press, NJ, USA, 105–114. https://doi.org/10.1109/ISMAR.2018.00039

[28] Kangsoo Kim, Celso M de Melo, Nahal Norouzi, Gerd Bruder, and Gregory F Welch. 2020. Reducing Task Load with an Embodied Intelligent Virtual Assistant for Improved Performance in Collaborative Decision Making. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, IEEE Press, NJ, USA, 529–538.

[29] Bret Kinsella and Ava Mutchler. 2019. U.S. Smart Speaker Consumer Adoption Report 2019. https://voicebot.ai/smart-speaker-consumer-adoption-report-2019/

[30] Tomoko Koda and Pattie Maes. 1996. Agents with faces: The effect of personification. In *Proceedings 5th IEEE International Workshop on Robot and Human Communication. RO-MAN'96 TSUKUBA*. IEEE, IEEE, NJ, USA, 189–194.

[31] Michael Lankes, Regina Bernhaupt, and Manfred Tscheligi. 2007. An Experimental Setting to Measure Contextual Perception of Embodied Conversational Agents. In *Proceedings of the International Conference on Advances in Computer Entertainment Technology* (Salzburg, Austria) *(ACE '07)*. Association for Computing Machinery, New York, NY, USA, 56–59. https://doi.org/10.1145/1255047.1255058

[32] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and Usability Engineering Group*. Springer, Springer Berlin Heidelberg, Berlin, Heidelberg, 63–76.

[33] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf Between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[34] Michal Luria, Samantha Reig, Xiang Zhi Tan, Aaron Steinfeld, Jodi Forlizzi, and John Zimmerman. 2019. Re-Embodiment and Co-Embodiment: Exploration of Social Presence for Robots and Conversational Agents. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (San Diego, CA, USA) *(DIS '19)*. Association for Computing Machinery, New York, NY, USA, 633–644. https://doi.org/10.1145/3322276.3322340

[35] Karl F MacDorman, Robert D Green, Chin-Chang Ho, and Clinton T Koch. 2009. Too real for comfort? Uncanny responses to computer generated faces. *Computers in human behavior* 25, 3 (2009), 695–710.

[36] Angie Lorena Marin Mejia, Doori Jo, and Sukhan Lee. 2013. Designing Robotic Avatars: Are User's Impression Affected by Avatar's Age?. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction* (Tokyo, Japan) *(HRI '13)*. IEEE Press, NJ, USA, 195–196.

[37] Richard E Mayer and C Scott DaPra. 2012. An embodiment effect in computer-based learning with animated pedagogical agents. *Journal of Experimental Psychology: Applied* 18, 3 (2012), 239.

[38] Rachel McDonnell, Martin Breidt, and Heinrich H Bülthoff. 2012. Render me real? Investigating the effect of render style on the perception of animated virtual humans. *ACM Transactions on Graphics (TOG)* 31, 4 (2012), 1–11.
[39] Masahiro Mori et al. 1970. The uncanny valley. *Energy* 7, 4 (1970), 33–35.
[40] Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine* 19, 2 (2012), 98–100.
[41] Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues* 56, 1 (2000), 81–103.
[42] Clifford Nass, Youngme Moon, and Nancy Green. 1997. Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of applied social psychology* 27, 10 (1997), 864–876.
[43] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers Are Social Actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston Massachusetts USA) *(CHI '94)*. Association for Computing Machinery, New York, NY, USA, 72–78. https://doi.org/10.1145/191666.191703
[44] Catherine S Oh, Jeremy N Bailenson, and Gregory F Welch. 2018. A systematic review of social presence: Definition, antecedents, and implications. *Frontiers in Robotics and AI* 5 (2018), 114.
[45] Robert Porzel and Manja Baudis. 2004. The Tao of CHI: Towards Effective Human-Computer Interaction. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*. Association for Computational Linguistics, Boston, Massachusetts, USA, 209–216. https://www.aclweb.org/anthology/N04-1027
[46] Alisha Pradhan, Kanika Mehta, and Leah Findlater. 2018. "Accessibility Came by Accident": Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 459, 13 pages. https://doi.org/10.1145/3173574.3174033
[47] Lingyun Qiu and Izak Benbasat. 2009. Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems. *Journal of management information systems* 25, 4 (2009), 145–182.
[48] Byron Reeves and Clifford Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press, New York, NY, USA.
[49] Kristin E. Schaefer, Jessie Y. C. Chen, James L. Szalma, and P. A. Hancock. 2016. A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. *Human Factors* 58, 3 (2016), 377–400. https://doi.org/10.1177/0018720816634228 arXiv:https://doi.org/10.1177/0018720816634228 PMID: 27005902.
[50] Susanne Schmidt, Gerd Bruder, and Frank Steinicke. 2018. Effects of Embodiment on Generic and Content-Specific Intelligent Virtual Agents as Exhibition Guides.. In *ICAT-EGVE*. The Eurographics Association, Geneve, Switzerland, 13–20.
[51] Susanne Schmidt, Gerd Bruder, and Frank Steinicke. 2019. Effects of virtual agent and object representation on experiencing exhibited artifacts. *Computers & Graphics* 83 (2019), 1–10.
[52] Valentin Schwind, Katrin Wolf, and Niels Henze. 2018. Avoiding the Uncanny Valley in Virtual Character Design. *Interactions* 25, 5 (Aug. 2018), 45–49. https://doi.org/10.1145/3236673
[53] Jun'ichiro Seyama and Ruth S Nagayama. 2007. The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence: Teleoperators and virtual environments* 16, 4 (2007), 337–351.
[54] Selina Jeanne Sutton. 2020. Gender Ambiguous, Not Genderless: Designing Gender in Voice User Interfaces (VUIs) with Sensitivity. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (Bilbao, Spain) *(CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 11, 8 pages. https://doi.org/10.1145/3405755.3406123
[55] Akikazu Takeuchi and Taketo Naito. 1995. Situated Facial Displays: Towards Social Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '95)*. ACM Press/Addison-Wesley Publishing Co., USA, 450–455. https://doi.org/10.1145/223904.223965
[56] Isaac Wang, Jesse Smith, and Jaime Ruiz. 2019. Exploring Virtual Agents for Augmented Reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article 281, 12 pages. https://doi.org/10.1145/3290605.3300511
[57] Jun Xiao, John Stasko, and Richard Catrambone. 2004. An Empirical Study of the Effect of Agent Competence on User Performance and Perception. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 1* (New York, New York) *(AAMAS '04)*. IEEE Computer Society, USA, 178–185.
[58] Beste F. Yuksel, Penny Collisson, and Mary Czerwinski. 2017. Brains or Beauty: How to Engender Trust in User-Agent Interactions. *ACM Trans. Internet Technol.* 17, 1, Article 2 (Jan. 2017), 20 pages. https://doi.org/10.1145/2998572
[59] Sean Zdenek. 2007. "Just roll your mouse over me": Designing virtual women for customer service on the web. *Technical Communication Quarterly* 16, 4 (2007), 397–430.

A APPENDIX: EXPERIMENT TASK LIST

The participants were provided with the following list as a printout and asked to take care of the tasks with the help of the voice assistant:

- Switch on the lights
- Turn on music (new "Fettes Brot" single)
- How long do eggs have to cook (they should be wax soft and are size M)
- Start a timer for the eggs
- Find out how many calories Gouda has
- Find out how many calories cashew cheese has
- Listen to science news
- Learn how the women's world cup final ended
- Find out if it's gonna rain today
- Check if there are appointments for today in the calendar
- Order a new micro USB charger (under 8€ and from Samsung)

The items were written out in German and English.

Publication 6

"Let's Face It": Investigating User Preferences for Virtual Humanoid Home Assistants

Nima Zargham, Dmitry Alexandrovsky, Thomas Mildner, Robert Porzel, Rainer Malaka

In Proceedings of the 11th International Conference on Human-Agent Interaction (HAI '23). New York, NY, USA, 2023. Association for Computing Machinery.

Personal contribution to this work: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, validation, visualization, and contribution to all parts of the manuscript.

ISBN: 979-8-4007-0824-4/23/12 DOI: 10.1145/3623809.3623821

"Let's Face It": Investigating User Preferences for Virtual Humanoid Home Assistants

Nima Zargham
zargham@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

Dmitry Alexandrovsky
dmitry.alexandrovsky@kit.edu
HCI and Accessibility
KIT
Germany

Thomas Mildner
mildner@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

Robert Porzel
porzel@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

Rainer Malaka
malaka@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

ABSTRACT

While a growing number of households contain home assistants, they mainly remain voice-only devices where the virtual agent is not represented visually. The visual representation of the agent is limited to the device's housing and abstract light animations that signify the assistant's state to its users. However, the audio channel is limited in conveying information beyond semantic content. Embodied virtual assistants can enhance interaction with conversational interfaces by adding a visual layer to further convey personality and human characteristics. In this work, we conducted an online survey ($N = 78$) to explore people's preferences for visualizing humanoid assistants. Our findings suggest that participants prefer an agent who appears mature, healthy, competent, and attractive. Furthermore, demographic similarities between the users and agents are wished for the agent to look more relatable. We discuss these findings and their implications for the design of virtual humanoid home assistants.

CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**; User studies.

KEYWORDS

Home Assistants, Conversational Agents, Embodied Agents, Smart Displays, Humanoids, Virtual Agents

ACM Reference Format:
Nima Zargham, Dmitry Alexandrovsky, Thomas Mildner, Robert Porzel, and Rainer Malaka. 2023. "Let's Face It": Investigating User Preferences for Virtual Humanoid Home Assistants. In *International Conference on Human-Agent Interaction (HAI '23), December 04–07, 2023, Gothenburg, Sweden.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3623809.3623821

HAI '23, December 04–07, 2023, Gothenburg, Sweden

ACM ISBN 979-8-4007-0824-4/23/12…$15.00
https://doi.org/10.1145/3623809.3623821

1 INTRODUCTION

Home assistants, such as Google Home, Amazon Echo, or Apple HomePod, are common devices in many households nowadays. With an increasing number of these systems finding their way into homes, Voice Assistants (VAs) play a more significant role as everyday digital assistants [97]. These systems support users with everyday tasks such as smart home control, weather forecast, playing music, and scheduling appointments, among others. The affordances and accessibility facilities of home assistants differ from VAs on smartphones [12]. Besides enabling users to control smart home appliances, these systems are mainly stationary and non-portable, and interaction with them is possible from a distance [71, 73]. Moreover, home assistants are often shared devices in a household, which might be used by multiple individuals, such as family members or roommates.

Despite the technological advancements and the benefits that such systems provide, it is still highly challenging to design a satisfying experience with VAs, as many users still find interaction with such systems unsatisfactory, disappointing, or embarrassing [11, 18, 56, 72, 96]. Research suggests that this is partly because such systems do not fulfill the user's expectations as an interlocutor [32, 45, 56, 63, 99].

It has been established that imitating human qualities affects how users feel about digital assistants [12, 28, 32]. Communication between people is more than just verbal exchange. Face expressions and body movements are also essential factors for transmitting information. Through non-verbal communication, such as visual cues, people can communicate information other than just the semantic content of the message, such as their emotions and current mood [21]. Incorporating a visual dimension in communication holds additional value by enhancing accessibility for individuals with hearing impairments. Furthermore, these non-verbal factors also convey part of the person's personality. With regards to home assistants, due to their voice-only characteristics, the agents' personality is predominantly conveyed through their voice, pre-configured personifications such as their names (e.g., Amazon's Alexa, Google Home, and Apple's Siri), and the design of the physical device [8, 10, 75]. However, research has shown that the digital systems' visual characteristics further impact human-machine interaction, users' trust, engagement, and perception of

the agent's personality [31, 37, 41, 48, 78]. Today, the visual representation of home assistants is usually limited to their casing and abstract light animations that function as signifiers to communicate the assistant's states to its users. To further convey personality and human characteristics, virtual agents can be augmented with an embodiment. Embodiment here refers to the representation of these agents in a physical or virtual form. It involves giving the virtual agent a visual form, such as a virtual body or avatar, and enabling it to exhibit realistic movements, gestures, and expressions. Throughout this paper, we refer to the agent's embodiment as the virtual embodiment, not its physical body.

Embodied VAs are a natural extension of conversational interfaces as they enhance the interaction by supplementing the experience visually [3, 19, 67, 90]. The more recent product category of home assistants is equipped with a screen to display visual output, enabling the virtual agent to be embodied. Such devices are referred to as *smart displays* [12, 82]. Human-Computer Interaction (HCI) researchers have been highlighting the expedience of embodied virtual agents and how the interaction with these systems is generally perceived as more natural compared to agents without embodiment [3, 12, 19, 20, 51, 85, 90]. Moreover, the literature recommends that having a visualized assistant can make it more approachable and dependable compared to a voice-only assistant [12]. The experience with home assistants depends on how users perceive an agent's personality [12, 75], and literature suggests that the virtual embodiment of agents could change their perceived personality through appearance and behavior [21]. Altering this perception can affect users' trust and engagement with such devices [12, 17, 102]. It is important to note that users' preferences for an embodied home assistant can differ strongly [41]. For instance, while some might prefer a humanoid assistant with specific human characteristics such as gender, age, and ethnicity, others might favor a fictional character with ambiguous characteristics. Due to these individual differences, the systematic adaptation of agents to users is very challenging [88], which makes a singular, universal design that satisfies all users' expectations impractical. Moreover, previous research indicates that users find it easier to interact with technologies that, to some extent, resemble their own characteristics [15]. For example, people with high traits in agreeableness would like their home assistant to be highly agreeable [96].

While recent work has established an early understanding of how embodiment in VAs positively affects users' experience, the community still lacks a clear understanding of how users' personal traits relate to the expectations they pose toward their VAs. Closing this gap, in this work, we explore people's preferences for a humanoid visualization of home assistants with respect to users' perceptions of their characteristics. In this research, we focus on the representation of an agent in humanoid form. Past studies have consistently emphasized the benefits of utilizing humanoid designs in human-computer interaction as they can enhance human-likeness, likability, and the perception of shared reality [77], foster a sense of familiarity, credibility, and trust during interactions [69, 75, 81], and facilitate a form of attachment [94]. Moreover, research by Bonfert et al. [12] indicates that users commonly prefer humanoid forms for their home assistants over abstract shapes or fictional characters. Given these findings, focusing on a humanoid design aligns with the established literature.

We conducted an online survey asking people to indicate their preferences for a desired virtual home assistant. More specifically, our work addresses the following research questions:

RQ1: How do users imagine the visualization of their desired humanoid home assistant?

RQ2: What is the relation between users' own characteristics and their preferences for virtual assistants?

We contribute an empirical study exploring users' preferences for a desirable virtually embodied agent with respect to the users' characteristics. Specifically, the study examines the agent's demographics, looks, and personality, taking into account the user's attributes based on their self-perception. Our results indicate that users prefer agents that look healthy, attractive, and mature as they can convey proficiency. Additionally, participants preferred an agent with similar demographics as themselves, as this facilitates a stronger sense of relatedness with the agent. We discuss these findings on users' preferences for virtual home assistants and their implications for future research avenues. The insights of this work can be particularly valuable for designers and developers seeking to create VAs that tailor the users' individual needs, preferences, and expectations. While HCI researchers have extensively studied the topic of embodied agents, our research contributes a new dimension to this field by focusing on the relation between users' attributes and their preferred virtually embodied agent. This approach provides valuable insights into the design of human-centered voice assistants.

2 RELATED WORK

HCI researchers have been investigating different forms of visual representation for virtual assistants for decades. This section discusses research on embodied virtual agents, agents' anthropomorphism, and home assistants' personalization.

2.1 Embodied Virtual Agents

Facial expressions are an essential part of communication and social interactions. They can express an interlocutor's involvement in a conversation, their emotional state, responsiveness, and understanding [36]. As the perception of an agent's personality is experienced multimodally, it is challenging for developers to convey the personality if they only focus on either voice or visual characteristics [21]. A considerable body of research has been investigating embodied virtual agents and their impact on user experience [12, 31, 37, 41, 48, 78]. Embodied agents can use non-verbal communication techniques on top of speech interaction to further convey information and emotions and to better connect with the users [49, 90].

The use of humanoid embodiment for an agent impacts the users' perception of social presence, which can affect their trust and engagement with the system [5, 42, 74]. Grover et al. [35] compared a text-based agent with a virtual agent with a video embodiment to support information workers to be more productive and focused. Their findings suggest that working with the embodied agent felt more productive and less distractive. Kim et al. [49] suggest that supplementing voice assistants with a virtual embodiment increases users' trust and confidence in the agent's awareness of real-world events and its ability to influence them. Similarly, Nowak and Rauh

[69] suggest that more anthropomorphic embodied virtual agents are perceived as more credible and trustworthy. Generally, human likeness has shown to be effective in increasing trust, loyalty, and engagement [15, 41, 48]. Studies have shown that people treat a system more like a person when it entails embodiment components [51, 85].

Beyond the visual characteristics of the agents, it is not well understood how the embodied assistants should act within periods of interaction in daily life. For instance, Bonfert et al. [12] showed ambiguous results regarding people's preferences on the display of embodied agents when they are not actively assisting. Some prefer to have the agent continuously visible on the screen to indicate the agent's presence and readiness to support. In contrast, others may feel irritated and observed if the agent is constantly visible [12]. In a study by Kim et al. [49] where they used a humanoid visualization for an agent in augmented reality, the agent would walk out of the room to give the user more privacy when the user requested it, which led to higher perceived privacy among their users. A study by Zargham et al. [96] also points to a similar result. The authors recommend hiding the agent between tasks to avoid social awkwardness. To provide further insights on these mixed results, our survey includes questions targeting if and how VAs should appear to the users.

2.2 Anthropomorphism of Agents

Anthropomorphism refers to people's tendency to attribute human characteristics to non-human objects to make the entity's actions comprehensible [33]. These characteristics entail cognitive capabilities, personality, and physical appearance [91, 103]. Researchers distinguish anthropomorphism into two categories of implicit and explicit anthropomorphism [104]. Implicit anthropomorphism happens intuitively and unconsciously. In contrast, explicit anthropomorphism occurs consciously in a reflective process that might moderate the initial judgment. Regarding virtual agents, implicit anthropomorphism occurs within the initial interactions, and in time, explicit anthropomorphism occurs through questioning and further exchanges [89]. Researchers have looked into various human qualities for agents to enhance HCI and improve user experience, including realistic voices, embodiment, and agent personalities [12, 88, 98]. Utilizing anthropomorphism in products has been shown to be beneficial [39]. A study by Yuan and Dennis [94] suggests that visual anthropomorphization can result in higher purchase rates by creating a form of attachment to the product. Likewise, Seymour and Van Kleek [81] observed a strong correlation between trust and anthropomorphism shown by users towards their voice assistants.

People generally urge to anthropomorphize agents by assigning specific human characteristics such as age, gender, and ethnicity to them [75]. By doing so, certain human stereotypes are also attributed to agents as a consequence. One of the most common types of such stereotypes is gender [30, 84, 101]. Studies have shown that the gender characteristics of an agent can potentially impact the user experience [7, 12, 13, 44]. To avoid gender stereotypes, researchers have recommended using androgynous or gender-ambiguous agents [65]. Others recommended using a team of agents with various genders [98]. With the change in the notions of gender stereotypical traits and roles in modern society, recent studies also highlight that gender stereotypes are not as effective as previously assumed for virtual agents [65].

The attractiveness of virtual agents has also been highlighted as an influential element in human-agent communication. Researchers argue that attractive and more elaborate agents have more successful social interactions [6]. It has been argued that the attractiveness of agents is sometimes more important than their reliability [95]. A study by Bonfert et al. [12] has shown that users favor an attractive agent with a customizable appearance, such as preferred hair color. Khan and De Angeli [46] claim that users maintained a better evaluation of attractive agents regardless of their interaction. In a study by Zargham et al. [96], authors found that users prefer home assistants to look fit and healthy. They also witnessed that some users would like such agents to occasionally change their outfits to exhibit that they also have a routine life. Another study by Khan and Sutcliffe [47] demonstrated that an embodied agent's attractiveness significantly impacts users' perceptions and behavior. Their results further highlight that attractive agents are more persuasive in influencing the users' decision-making than unattractive agents. In our study, we explore the factor of attractiveness in relation to virtually embodied home assistants.

Although anthropomorphization is an expected human behavior when interacting with agents, if the human-like characteristics are close to real humans but not quite, this can create an adverse reaction referred to as the "uncanny valley" [62]. This effect emerges when the degree of realism is inadequate, or people observe anomalous features [80]. Although realistic humanoids can appeal to users [38, 57, 59], a certain level of social responsiveness and aesthetic refinement should be incorporated to achieve this [38]. Researchers argue that even abstract faces can be eerie when it comes to computer-generated faces [57]. To avoid this uncanny effect, Schwind et al. [79] believe that designers should be consistent in realism while deliberate in stylization. Systems that heavily depend on human-like paradigms can form unrealistic expectations that the system may not be able to meet [34]. Anthropomorphism can also be increased by designing agent personalities [50], which can be conveyed through multiple channels, including the agent's voice characteristics, use of language, and appearance [21].

2.3 Personalization and Customization of Agents

The adaptation of services and features of a system to user-specific preferences has been a common practice in HCI. Previous studies on voice assistants highlight that users prefer systems that can adapt to their preferences [14, 26, 27, 29, 53, 54]. By implementing user models which contain information about individual users and their preferences, systems can adapt to users' needs [4, 87]. This process is referred to as personalization, which has been shown to benefit product owners and customers by enabling higher efficiency [40, 70]. For instance, Braun et al. [14] found that users prefer and trust a personalized car agent more than a non-personalized one, especially if their personalities match. Similarly, customization is an approach to enhance user experience and performance [23, 61, 64]. Customization means that users can explicitly select certain features between specific options, whereas personalized features are

automatically adjusted by machines based on users' individual needs [68]. An extensive body of research has highlighted the value of customization for agents, specifically for people with special needs [1, 2, 61, 64, 93]. Giving users more agency and control could enhance their interaction with computers [76, 93, 100]. Generally, users wish to customize different design factors of home assistants, including voice characteristics, the number of agents, and their roles and personalities [98]. Paay et al. [71] argue that, even as voice assistants improve their ability to adapt to users' personality preferences and expectations in different situations, individuals still desire control over the agent's responses to their interactions. Regarding agent visualization, previous research recommends that having an embodiment for a home assistant should be a customizable feature of these devices as users' preferences often vary, and no universal solution could meet the expectations of all users equally [96].

One principle that has been transferred from human-human interaction to human-agent interaction is the similarity-attraction principle. This principle states that people are attracted to others when they perceive them to be similar to themselves [9]. In human-human interaction, it has been shown that initial interpersonal attraction positively correlates with the number of similar attitudes that two people hold [16]. The simplicity of this effect makes it ideal for human-agent interaction, as it could provide a basic and dependable way of influencing interactions. Nass and Lee [66] suggests that the similarity-attraction principle is an influential tool designers can employ to enhance product satisfaction and increase positive impressions towards the company producing the product. Research on embodied agents has shown that people prefer interacting with an assistive agent whose personality matches their own [86, 96]. Additionally, a study by Bernier and Scassellati [9] demonstrated that people rated a social robot more favorably when it displayed preferences similar to their own.

Although the similarity-attraction effect has been observed in various settings and interpersonal situations, previous literature has primarily focused on the dimensions of behavior and personality [9, 22, 25, 96]. However, there is a gap in the literature regarding the application of this principle to preferences for the visual embodiment of an agent. This study aims to address this gap by examining the similarity-attraction principle in relation to preferences for the visual appearance of a home assistant. Moreover, while previous research has established the importance of customization and personalization for improving user experience, there is still a gap in understanding how users prefer to see such agent features and which anthropomorphic characteristics of the agents are relevant to the user. Our work addresses this gap by exploring people's preferences for virtual humanoid agents in a domestic setting. By examining different elements of the agent's embodiment, we aim to identify common trends in user preferences that can guide future design choices.

3 STUDY DESIGN

In order to explore users' preferences on home assistant embodiment with regard to users' perceptions of their own characteristics, we designed an online survey where we collected the preferences of 78 anonymized participants. After a welcome text and a short introduction, they gave informed consent. Participants then answered demographic questions, including questions about their age, gender, ethnicity, nationality, native language, and accent or dialect. This was followed by questions about participants' appearance, including their hair color and body shape. Furthermore, we collected information about their usage of voice assistants, interest in such devices, and enthusiasm for technology. Next, participants were asked to indicate how they would like their home assistant to look regarding the agent's gender, age, ethnicity, body shape, hairstyle and hair color, and outfit. Participants also had to indicate the most important part of the agent's face and rate how much the attractiveness of the agent matters to them. Afterward, we asked participants about the embodied agent's idle state. Next followed a series of questions about the agent's degree of realism and rendering on seven-point Likert scales. We asked participants to indicate if they would like the agent to look more abstract or photo-realistic, robotic or human-like, and 2-dimensional or 3-dimensional. Participants also answered questions about the assistant's personality. We asked participants about the agent's emotional expressions and which emotions they found inappropriate for the agent to express. Lastly, we asked participants to rate the importance of the assistant's looks, voice, and personality individually on seven-point Likert scales and collected concluding comments. The survey contained 48 questions and took approximately 20-30 minutes to complete. The full list of questions can be found in the supplementary material.

3.1 Participants

Participants were recruited for this study through various channels, including internet forums, mailing lists, social networks, and word-of-mouth. Participation in the survey was voluntary and uncompensated. Initially, a total of 85 individuals completed the survey. Seven participants were excluded from the analysis as they were identified to be below 18, which is required for giving consent in the legal sense. As a result, the final evaluation included data from 78 participants (42 male, 32 female, three androgynous, one preferred not to say) aged from 18 to 59 years ($M = 28.21$, $SD = 7.28$). We had participants of 22 nationalities, mostly from Germany (20.5%), the United States (18%), Iran (14.1%), Italy (9%), and Great Britain (9%), primarily residing in Germany (41%), the US (18%), Canada (9%), or Italy (9%). Our sample consisted of 19 native languages, with English (34.6%), German(20.5%), and Farsi (14.1%) being the most common. The majority of our participants were Caucasians (80.7%), and others were Asian (12.8%), Latin American (2.5%), Arab/North African (1.3%), or had a mixed ethnicity (2.5%). Most participants indicated they had brown (dark 42.3%, light 12.8%) or black (28.2%) hair, with the rest having blonde (10.2%), red (2.5%) or grey (2.5%) hair colors. On a seven-point Likert scale, participants indicated their body shape (1 - Extremely Skinny to 7 - Extremely Overweight), with most participants reporting an average body shape ($Mean = 3.79$, $SD = 1.12$). 43.6% of the participants did not use voice assistants before, while 34.6% were regular users, and the rest (21.8%) rarely used such systems. 38.4% of the participants owned a home assistant. From those who owned a home assistant, 80% used it in English, 12% German, and 8% Italian.

4 RESULTS

In this section, we present our findings in three parts: Agent's demographic features, agent's visualization, and agent's personality.

To determine if the participants rated the importance of the agent's personality, looks, or voice differently, we conducted a one-way ANOVA with types of ratings as the independent variable and the score as the dependent variable. The analysis showed a significant effect $F(2, 252) = 28.54, p < 0.01, \eta_p^2 = 0.185$ (see Figure 1). Post-hoc comparisons confirmed that the participants rated the importance of the agent's personality ($M = 5.63, SD = 1.35$) higher than the agent's visual appearance (M =4.22, SD = 1.71, $t(84) = 5.905$, $p < .0001$), displaying a large effect ($d_{Cohen} = 0.906$) [24]. Similarly, the agent's voice (M = 5.8, SD = 1.33) was rated more important than the agent's visual appearance ($t(84) = 6.646, p < .0001$), with a large effect ($d_{Cohen} = 1.020$). No significant difference was observed between the agent's voice and personality.

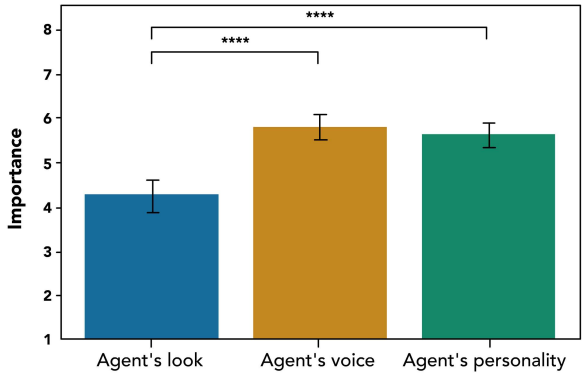


Figure 1: Means and standard deviation of ratings on the importance of the agent's look, voice, and personality. Significant differences are highlighted with astricts.

We further investigated whether different preferences of the avatar's personality, physiology, or looks differed based on gender. Therefore, we conducted a series of independent sample t-tests between males (n = 42) and females (n = 32) on the preference responses. Due to a small number of non-binary (n = 3) or not-to-disclose (n = 1) participants, they were excluded from the analysis. The analysis showed no significant differences between users' preferences in this case. Likewise, there were no significant correlations between participants' interest in technology and their preferences.

4.1 Agent's Demographic Features

We analyzed participants' preferences regarding the agent's demographic features with respect to their own demographics. We observed that the majority of the participants (59%) chose their native language for the assistant. Regarding the agent's spoken language, even though only 34.6% of our participants were native English speakers, 69.2% preferred the agent to speak in English.

Concerning the agent's dialect and accent, most participants (89.7%) had a pragmatic preference and preferred accents or dialects that are more understandable for them: "Australian accent, as I am from Australia" (P65). Others (10.3%) had more of a hedonic approach to this. One user mentioned: "British accent because it sounds more official" (P39). We also observed a sense of playfulness among some (6.4%) participants. For instance, one participant said: "I prefer the agent to have a southern accent because it would be funny" (P37).

Regarding the agent's gender, 38.4% of our participants preferred an androgynous agent, 33.3% preferred a female agent, 20.5% male agent, and 7.7% had no preference. There was a slight trend that men preferred women as virtual agents. However, a Chi-Square test could not confirm a dependence of avatar gender preference on the participants' gender ($\chi^2(9) = 9.28$, ($p = 0.41$).

Participants' mean preferred age for the agent was around 30 years old ($Mean$ = 31.298, SD = 11.263). We witnessed that most participants wanted the agent to be around their own age. One participant mentioned: "It is more trustable around my age" (P66). Another believed the agent is more "relatable" (P64). We observed a positive correlation between our participants' ages and their preference for the agent's age, with a Pearson's coefficient of r = .265 (p = 0.018). Users generally wanted the agent to be young. Different reasons were mentioned for this preference. Some believed the agent would look more attractive in their 20s or 30s. One person mentioned that the agent looks more competent: "between 25 to 50 feels smarter and fresh-minded to me." (P5) On the other hand, one participant wanted the agent to look older as they believed "the assistant is someone designed to help and give guidance, so it should be someone who is old and experienced in life." (P40)

82% of our participants stated a preference regarding the ethnicity of the agent. Among those, 45.3% (n = 29) stated a preference selected their own ethnicity, making their own ethnicity as a preference for the agent the top choice ($\chi^2(24) = 38.85$, $p = 0.028$).

4.2 Agent's Visualization

Around half of our participants (46.2%) preferred the agent to be 3-dimensional, while 21.8% preferred 2-dimensional visualization. One-third of our sample (32%) did not have a preference on this matter. Also, in terms of the realism of the rendering, participants wanted the agent to be somewhat realistic, with a mean rating of 4.35 (SD = 1.76). Regarding agents' human-likeness, people rated above average ($Mean$ = 4.11, SD = 1.86). We found a strong correlation between human-likeness ratings and how realistic it is visualized with a Pearson's r = .582 ($p < 0.001$), suggesting that participants generally consider the agent's human-likeness and realistic visualization as closely related (see Figure 2).

50% preferred the agent to look just as or more attractive than them. For 50% of our participants, it did not matter if the agent looked more or less attractive than them, and no participant indicated a preference for a less attractive agent. Furthermore, in our Likert-scale question about the agent's attractiveness, participants rated above average scores ($Mean$ = 4.02, SD = 1.88). One-third of the participants (34.6%) did not want the agent's appearance to match theirs, while 14.1% thought otherwise. Half of the participants (51.3%) did not have a preference for this.

Regarding the agent's outfit, 57.7% of people had a preference for how the agent should be dressed. Similar to attractiveness,

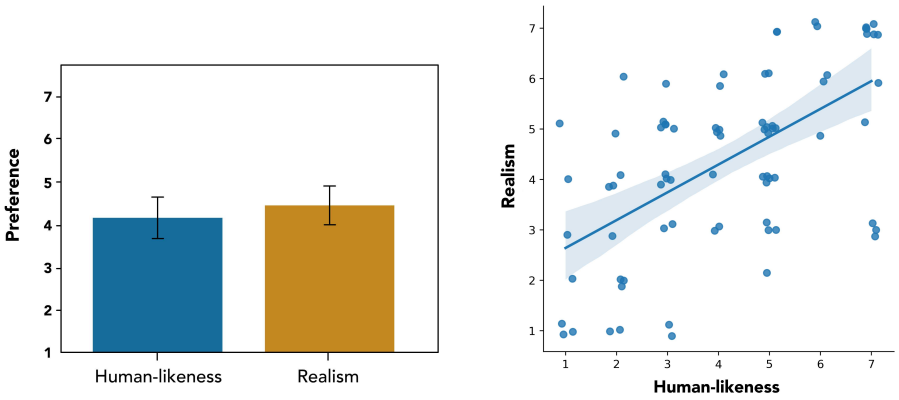


Figure 2: Mean scores for the agent's human-likeness and realism of the rendering (left), as well as the correlation plot of the two scales (right).

Table 1: Mean values and standard deviations of the Likert scale questions about agent characteristics. All questions were on a scale of 1 to 7.

	Mean	Standard Deviation
Realism of the rendering	4.35	1.76
Human-likeness	4.11	1.86
Body shape	3.73	0.76
Attractiveness	4.02	1.88
Importance of outfit	4.08	1.89
Expressiveness of emotions	4.24	1.58

participants rated the importance of the agent's outfit slightly above average (*Mean* = 4.08, *SD* = 1.89). One-third of the participants (20.5%) did not want the agent's outfit to match theirs, while 23.1% found it interesting to have matching outfits. More than half (56.4%) did not have a preference for this. 20.5% of participants wanted the agent to dress casually. One participant said: "A simple t-shirt and jeans, so the assistant does not look or feel too out of place" (P44). Another group of participants (10.25%) wanted the agent to dress formally, as one user mentioned: "[The agent] should dress professionally or business casual" (P44). Others (14.1%) had specific wishes: "[The agent] should be wearing Yoga pants with a regular long-sleeved shirt" (P62). 12.8% of the people wanted the agent to change outfits depending on the weather, time of the day, the topic of the interaction, or ongoing events. One person said: "early in the morning [The agent] can wear pajamas. If it is hot outside, the agent can wear light clothing. It could also incorporate the schedule of the owner. So if a business meeting is ahead, the agent wears formal clothing" (P35). Four participants (two males and two females) specifically mentioned that the agents' clothing should not be revealing: "The outfit should not be designed to be overly revealing, but does not have to be extremely modest either" (P75). On the other hand, we also witnessed that two participants (both males) wanted the agent to have a revealing or sexualized outfit.

Regarding the agent's body shape, we observed a correlation between participant perception of their own body shape (*Mean* = 3.79, *SD* = 1.12) and the agent' (*Mean* = 3.73, *SD* = 0.76) with a Pearson's $r = .221$ ($p = 0.05$). Participants indicated eyes as the most important part of the agent's face (64.1%), followed by mouth (20.5%), and hair (6.4%). 56.4% of participants did not have any preferences regarding the agent's hairstyle. However, most participants (85.9%) had a preference for the agents' hair color. Among those who stated a preference, 41.7% ($n = 28$) selected their own hair color. The majority reported they prefer natural hair colors with the top-ranked colors *brown* (26×) and *black* (13×). Unnatural colors such as *blue*, *green*, or *purple* were stated 6×, and three participants suggested interchangeable hair color. When asked if the assistant should look like someone they would know where participants could give multiple answers, 47.8% noted a desire for fictional characters, 17.9% preferred celebrities they like, and 3% wanted it to look like someone they know in their personal life. 32.8% of the participants did not want the agent to look familiar.

Over three-quarters of the people (80.7%) did not want the agent to be constantly present on the screen. 33.3% recommended that the agent simply fade in when called and fade out after an action is performed. Others mentioned that the agent should perform a random act before and after executing specific actions, such as walking in and out of the screen (P37), taking a nap and waking up when called (P70), acting as a genie by going back to the lamp after the task is performed (P10), or look away until the next inquiry by the user (P26).

4.3 Agent's Personality

Among participants, 82% favored agents' who can show emotions, and the expressiveness of emotions was rated above average (*Mean* = 4.28, *SD* = 1.58). We asked participants to indicate which emotions would be inappropriate for the agent to show, where participants could mention multiple options. Participants found it inappropriate for the agent to express disgust (33.3%), anger (33.3%), happiness (33.3%), surprise (28.2%), excitement (28.2%), fear (27%), and sadness

(27%). 34.6% found all emotions to be appropriate for the agent. One participant requested to be able to turn certain emotions on or off.

5 DISCUSSION

In this study, we surveyed user preferences for their desired humanoid home assistant (RQ1) and how they relate to their perception of their own personal characteristics (RQ2). Our survey results suggest that, even though the pragmatic aspects of home assistants are still more important to users, many highly value the hedonic facets of the agent. By hedonic qualities, we refer to the subjective and emotional aspects of the user experience, relating to the enjoyment, pleasure, and overall affective response that a user derives from using a product [52]. On the other hand, by pragmatic qualities, we are referring more to the utilitarian aspects of a product, focusing on the practical and functional aspects that contribute to achieving users' goals and meeting their specific needs.

Participants considered all three aspects of the agent's voice, personality, and appearance relevant and important. However, they placed a higher priority on the agent's voice and personality compared to their appearance, highlighting that understandability, approachability, and the agent's character are more influential in shaping the participants' perception of the agent than their visual appearance. In the following, we discuss our insights and implications derived from this study.

5.1 Linguistic Features

This study revealed that most participants preferred clear and understandable language, accent, and dialect in their virtual agent. This finding highlights a pragmatic approach by most participants who prioritize effective communication over other aspects. However, a small percentage of participants desired specific languages, dialects, and accents for entertainment purposes. Given these contrasting preferences, we recommend that virtual agent designers provide a default setting of clear and understandable language, accent, and dialect while also offering the option to customize the linguistic features of the agent for less serious interactions. This approach balances the need for effective communication with the desire for playful and entertaining linguistic features. Our findings further suggest that linguistic nuances can be an effective method to communicate different contexts to the user. For instance, in a utilitarian context such as work, the agent could use formal language, and while communicating personal information, the agents could change their speech to a more relaxed or intimate language use. However, one should keep in mind that playfulness with languages, dialects, and accents can be a sensitive topic. While it can be seen as a way to create a friendly and fun atmosphere, it also carries the risk of being perceived as insensitive or even offensive.

5.2 Agent's Appearance

A significant proportion of participants stated a preference for virtual agents that share similar demographics as themselves. This preference was evident in the correlations between the participants' age, language, accent, and dialect and the respective characteristics they preferred in the virtual agent. Additionally, a majority of participants who had a preference for the virtual agent's ethnicity chose their own ethnicity for the agent. This trend was also observed with regard to hair color preferences. These results suggest that users prefer virtual agents that are perceived as more relatable. This aligns with previous work, suggesting people prefer technologies that resemble their own characteristics [15]. Previous research further indicates that people tend to have more trust in systems that are similar to them [43]. Our study results also indicate that participants had a strong preference for an agent that has a young and mature appearance. They found elderly-looking agents to seem unfit and very young-looking to be immature. We argue that the assumed age of the agent could potentially impact users' perception of the agent's capabilities as well as its reliability. Moreover, people preferred the agent to have an average body shape. This is because people might consider overweight and underweight less fit. Additionally, the preference for an average body shape can also be influenced by societal attractiveness standards. Taken together, in line with previous research [12, 58, 96] on the appearance of VAs, these suggest that participants prefer an agent that looks healthy, mature, and competent. Moreover, although almost half of the people indicated that the attractiveness of the agent is not that important, we witnessed that no participant wanted the agent to look less attractive than themselves. Participants also rated the importance of this factor above average. Furthermore, participants' specific preferences for the agent's hair color and outfit highlight the value of the agent's attractiveness to the users. We also observed that users wanted the agent to look more human-like and the rendering to be more realistic than abstract. These findings align with previous work on agent visualization (cf. [6, 46, 95]) and indicate that agents should have a minimum level of visual appeal. Attractiveness is often associated with positive qualities such as popularity, competence, and desirability [55]. Previous research has shown that attractive communicators reach greater opinion agreement [60, 92]. Having an equally or more attractive virtual agent may create a perception of higher social status for the user. This can lead to a sense of prestige and satisfaction in the interaction, as individuals may feel they are associating themselves with a highly regarded entity. Previous research has highlighted that individuals associated with an attractive individual are evaluated more favorably by others [83]. Moreover, Society places significant emphasis on physical attractiveness and its perceived benefits. Societal norms and expectations may influence people's preferences for attractive virtual agents. They may believe that an attractive virtual agent reflects better design, higher quality, or more advanced technology and thus prefer it based on these perceived societal standards.

Regarding the agent's gender, we witnessed that participants mainly preferred an androgynous agent. This contrasts with some of the previous findings about gender stereotypes of agents [7, 12, 13, 44]. Moreover, we did not identify significant differences between the preferences of our male and female participants. In line with Nag and Yalçın [65], in our study, we observed that gender stereotypes were not as effective as previously assumed for virtual agents. We did witness the sexualization of the agent by a few participants, for instance, regarding the agent's outfit. Nevertheless, a more significant portion of our participants wished for the agent's outfit to be standard and modest. Some participants even explicitly stated that they did not want the agent's outfit to be revealing or sexualized. More than half of the participants had specific outfit preferences for the agent. Although this aspect of the

agent's appearance was mainly perceived as a playful feature, some suggested it could also be used to convey important information such as calendar events or weather forecasts. A theme observed from the participant responses was the desire for customizability and changeability in the virtual agent's visual appearance, voice characteristics, and behavior. People noted that they would like to be able to change the agent entirely (6.4%) to "not get tired of it". Others wanted to have the possibility to modify specific features such as the agent's outfit (12.8%), hairstyle and color (3.8%), ethnicity (2.5%), gender (2.5%), and voice (2.5%). In line with previous work [2, 27, 64, 93], these findings suggest that people value the ability to customize and personalize the virtual agent to their preferences and needs.

5.3 Character Building

Some participants expressed a desire for the agent to have a dynamic behavior and exhibit autonomous actions, indicating a preference for a more developed and unique character with a life of its own. For instance, participants desired the agent to change their hairstyle, color, or outfit without user interventions. This is also pointed out by our participants' comments about the agent's presence on the screen. Many recommended that the agent should perform random actions, such as reading a book or taking a nap, to simulate a routine life, indicating that it has something to do. However, this might also be because participants do not want to feel observed [12]. Consistent with prior research by Zargham et al. [96], our findings show participants' interest in developing a character for the agent with high capabilities and a unique personality.

We witnessed that the majority of our participants wanted the agent to express emotions. Their ratings suggest that participants wanted the agent to be expressive but not excessively so. Moreover, people found the agent's eyes and mouth to be as most important aspects. This might be due to the fact that these parts convey emotions and expressiveness. These are important aspects for increasing anthropomorphism and creating a sense of personality and emotional connection. Over one-third of our participants desired the agent to have the freedom to express all kinds of emotions. However, some participants felt that certain emotions were inappropriate for a virtual agent. This highlights the complex nature of emotional expression in virtual agents, as people may have different perceptions and views on agents and their roles. Some might view these devices as computer programs that do not need to fake emotions, while others anthropomorphize these systems and shape an emotional bond. Based on these findings, we conclude that virtual agents should have the capability to express emotions but that the emotional expression should be balanced and in line with users' expectations and cultural norms. Companies designing virtual agents may need to consider providing options for users to customize the emotional expressions of the agent to meet their individual preferences.

Eventually, we interpreted the results of this study to provide answers to the following research questions:

RQ1: How do users imagine the visualization of their desired humanoid home assistant?
RQ2: What is the relation between users' own characteristics and their preferences for virtual assistants?

Regarding **RQ1**, analyzing users' preferences for humanoid home assistants reveals a prioritization of pragmatic features, while hedonic aspects remain significant and valued. In terms of visual appearance, we observed that users would like the agent to seem mature, healthy, relatable, and attractive. Considering this and the pragmatically accurate demands for the agent, users' preferences resemble somewhat of a superhuman.

In response to **RQ2**, we witnessed that people do not necessarily want the agent to look like them but rather have some demographic similarities to be more relatable. Participants preferred virtual agents with similarities in age, language, accent, dialect, and ethnicity. This was also witnessed regarding hair color and body shape.

The varying preferences for different aspects of the virtual agent's appearance underscore the reality that a one-size-fits-all solution cannot adequately fulfill the diverse expectations of all users. Our findings repeatedly traced back to the inherent subjectivity in the preferences of individual participants. It became evident that granting the ability to modify various agent characteristics can enhance user satisfaction and engagement with such systems. Participants also demonstrated a keen interest in defining specific personality traits for the agent that could adapt to their individual interests. As a result, we recommend designers of VAs should shift their focus toward providing a greater array of customized and personalized features. By doing so, they can better address the diverse range of users' requirements and preferences.

6 LIMITATIONS & FUTURE WORK

The findings of this work present important considerations for designing home assistants. However, our research has several limitations that need to be acknowledged. Firstly, the results of this study must be viewed in the context of the specific group studied, despite having a diverse sample of various ages and backgrounds. In our study, we conducted a survey involving 78 participants from 22 different nationalities. While our sample included individuals with various cultural backgrounds, it's important to note that the sample size remains relatively small and might not offer a fully representative picture. Future research should validate our findings by exploring wider and distinct populations. More specifically, future work should explore more diverse group of participants, including those who identify as non-binary or who have less typical body shapes. Also, marginalized groups should be considered in future research as they may introduce unique but important characteristics that require tailored designs. Also, this study relies on self-reported participant data, which may be subject to participant bias, such as social desirability bias. The results have to be interpreted keeping such possible effects in mind. While we touched upon the ethical implications that may arise from the visualization of virtual agents, such as reinforcing gender or ethnic stereotypes, it is important that future studies place a greater emphasis on addressing these concerns. Specifically, issues related to privacy, trust, and potential social impacts should be thoroughly investigated and taken into consideration. This will help ensure that the development and use of virtual agents align with scientific and ethical standards. Moreover, we explored people's preferences for a humanoid agent in

an online survey, meaning the participants did not interact with the agents. Participants' feelings and perceptions might differ once they interact with such agents in real life. Hence in future work, we aim to investigate the user experience when people can actually interact with their designed agents. Lastly, studies should also explore the long-term usage of virtual humanoid home assistants, specifically to further investigate the impact of agent's presence in social settings, uncanny effect, usability, and privacy concerns.

7 CONCLUSION

In this exploratory study, we conducted a survey with 78 participants to collect people's preferences for an embodied humanoid home assistant with respect to their perception of their own characteristics. Our results suggest that participants prefer an agent who looks mature, healthy, and capable. Furthermore, several demographic similarities between the users and agents were requested in order for the agent to look more relatable. We also witnessed a wish for the agent to be an autonomous character with a unique personality, while customization and personalization of the agent's visual features were highly demanded. Overall, our findings highlight that the primary preference of most users lies in the pragmatic aspects of their home assistants. Nevertheless, by utilizing our findings to consider further the hedonic aspects of home assistants, designers and developers can potentially enhance the user experience with home assistants and expand the design space.

ACKNOWLEDGMENTS

This work was partially funded by the FET-Open Project 951846 "MUHAI – Meaning and Understanding for Human-centric AI" funded by the EU program Horizon 2020, as well as the German Research Foundation DFG as part of Collaborative Research Center (Sonderforschungsbereich) 1320 "EASE – Everyday Activity Science and Engineering", University of Bremen (http://www.ease-crc.org/) conducted in subproject H02.

REFERENCES

[1] Ali Abdolrahmani, Ravi Kuber, and Stacy M Branham. 2018. " Siri Talks at You" An Empirical Investigation of Voice-Activated Personal Assistant (VAPA) Usage by Individuals Who Are Blind. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. Association for Computing Machinery, New York, NY, USA, 249–258.

[2] Ali Abdolrahmani, Kevin M. Storer, Antony Rishin Mukkath Roy, Ravi Kuber, and Stacy M. Branham. 2020. Blind Leading the Sighted: Drawing Design Insights from Blind Users towards More Productivity-Oriented Voice Interfaces. *ACM Trans. Access. Comput.* 12, 4, Article 18 (Jan. 2020), 35 pages. https://doi.org/10.1145/3368426

[3] Sean Andrist, Michael Gleicher, and Bilge Mutlu. 2017. Looking Coordinated: Bidirectional Gaze Mechanisms for Collaborative Interaction with Virtual Characters. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2571–2582. https://doi.org/10.1145/3025453.3026033

[4] Muhammad Asif and John Krogstie. 2012. Taxonomy of personalization in mobile services. In *Proceedings of 10th IADIS International Conference e-Society*. IADIS – International Association for the Development of the Information Society, -, 343–350.

[5] Wilma A Bainbridge, Justin Hart, Elizabeth S Kim, and Brian Scassellati. 2008. The effect of presence on human-robot interaction. In *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, IEEE Press, NJ, USA, 701–706.

[6] Domna Banakou, Konstantinos Chorianopoulos, and Kostas Anagnostou. 2009. Avatars' appearance and social behavior in online virtual worlds. In *2009 13th Panhellenic Conference on Informatics*. IEEE, IEEE, NJ, USA, 207–211.

[7] Amy L. Baylor and Yanghee Kim. 2004. Pedagogical Agent Design: The Impact of Agent Realism, Gender, Ethnicity, and Instructional Role. In *Intelligent Tutoring Systems*, James C. Lester, Rosa Maria Vicari, and Fábio Paraguaçu (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 592–603.

[8] Erin Beneteau, Olivia K. Richards, Mingrui Zhang, Julie A. Kientz, Jason Yip, and Alexis Hiniker. 2019. Communication Breakdowns Between Families and Alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article 243, 13 pages. https://doi.org/10.1145/3290605.3300473

[9] Emily P Bernier and Brian Scassellati. 2010. The similarity-attraction effect in human-robot interaction. In *2010 IEEE 9th international conference on development and learning*. IEEE, IEEE, Manhattan, New York City, 286–290.

[10] Timothy W. Bickmore and Rosalind W. Picard. 2005. Establishing and Maintaining Long-term Human-computer Relationships. *ACM Trans. Comput.-Hum. Interact.* 12, 2 (June 2005), 293–327. https://doi.org/10.1145/1067860.1067867

[11] Michael Bonfert, Maximilian Spliethöver, Roman Arzaroli, Marvin Lange, Martin Hanci, and Robert Porzel. 2018. If you ask nicely: A digital assistant rebuking impolite voice commands. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction : ICMI'18 : Boulder, CO, USA, October 16 - 20, 2018*. ACM, New York, NY, 95–102. https://doi.org/10.1145/3242969.3242995

[12] Michael Bonfert, Nima Zargham, Florian Saade, Robert Porzel, and Rainer Malaka. 2021. An Evaluation of Visual Embodiment for Voice Assistants on Smart Displays. In *CUI 2021-3rd Conference on Conversational User Interfaces*. ACM, New York, NY, USA, 1–11.

[13] Sheryl Brahnam and Antonella De Angeli. 2012. Gender affordances of conversational agents. *Interacting with Computers* 24, 3 (2012), 139–153.

[14] Michael Braun, Anja Mainz, Ronee Chadowitz, Bastian Pfleging, and Florian Alt. 2019. At your service: Designing voice assistant personalities to improve automotive user interfaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–11.

[15] Cynthia Breazeal. 2003. Toward sociable robots. *Robotics and autonomous systems* 42, 3-4 (2003), 167–175.

[16] Donn Byrne and Don Nelson. 1965. Attraction as a linear function of proportion of positive reinforcements. *Journal of personality and social psychology* 1, 6 (1965), 659.

[17] Angelo Cafaro, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2016. First impressions in human–agent virtual encounters. *ACM Transactions on Computer-Human Interaction (TOCHI)* 23, 4 (2016), 1–40.

[18] Marcus Carter, Fraser Allison, John Downs, and Martin Gibbs. 2015. Player Identity Dissonance and Voice Interaction in Games. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play* (London, United Kingdom) *(CHI PLAY '15)*. Association for Computing Machinery, New York, NY, USA, 265–269. https://doi.org/10.1145/2793107.2793144

[19] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. 1999. Embodiment in Conversational Interfaces: Rea. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 520–527. https://doi.org/10.1145/302979.303150

[20] Justine Cassell and Kristinn R Thorisson. 1999. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence* 13, 4-5 (1999), 519–538.

[21] Susana Castillo, Philipp Hahn, Katharina Legde, and Douglas W. Cunningham. 2018. Personality Analysis of Embodied Conversational Agents. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (Sydney, NSW, Australia) *(IVA '18)*. Association for Computing Machinery, New York, NY, USA, 227–232. https://doi.org/10.1145/3267851.3267853

[22] Fang Fang Chen and Douglas T Kenrick. 2002. Repulsion or attraction? Group membership and assumed attitude similarity. *Journal of personality and social psychology* 83, 1 (2002), 111.

[23] Dasom Choi, Daehyun Kwak, Minji Cho, and Sangsu Lee. 2020. *"Nobody Speaks That Fast!" An Empirical Study of Speech Rate in Conversational Agents for People with Vision Impairments*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376569

[24] Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Erlbaum, Hillsdale, NJ.

[25] John W Condon and William D Crano. 1988. Inferred evaluation and the relation between attitude similarity and interpersonal attraction. *Journal of personality and social psychology* 54, 5 (1988), 789.

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

[27] Benjamin R Cowan, Derek Gannon, Jenny Walsh, Justin Kinneen, Eanna O'Keefe, and Linxin Xie. 2016. Towards Understanding How Speech Output Affects Navigation System Credibility. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, USA, 2805–2812.

[28] Benjamin R. Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "What Can I Help You with?": Infrequent Users' Experiences of Intelligent Personal Assistants. In

Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services (Vienna, Austria) *(MobileHCI '17)*. ACM, New York, NY, USA, Article 43, 12 pages. https://doi.org/10.1145/3098279.3098539

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

[30] Andreea Danielescu. 2020. Eschewing Gender Stereotypes in Voice Assistants to Promote Inclusion. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (Bilbao, Spain) *(CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 46, 3 pages. https://doi.org/10.1145/3405755.3406151

[31] Munjal Desai, Kristen Stubbs, Aaron Steinfeld, and Holly Yanco. 2009. Creating trustworthy robots: Lessons and inspirations from automated systems. In *Proceedings of AISB Convention: New Frontiers in Human-Robot Interaction*. Society for the Study of Artificial Intelligence and the Simulation of Behaviour, Bath, UK, 49–56.

[32] Philip R. Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R. Cowan. 2019. Mapping Perceptions of Humanness in Intelligent Personal Assistant Interaction. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services* (Taipei, Taiwan) *(MobileHCI '19)*. Association for Computing Machinery, New York, NY, USA, Article 5, 12 pages. https://doi.org/10.1145/3338286.3340116

[33] Brian R Duffy. 2002. Anthropomorphism and robotics.

[34] Leonard Foner. 1993. *What's an agent, anyway? a sociological case study*. Technical Report. Agents Memo 93.

[35] Ted Grover, Kael Rowan, Jina Suh, Daniel McDuff, and Mary Czerwinski. 2020. Design and Evaluation of Intelligent Agent Prototypes for Assistance with Focus and Productivity at Work. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 390–400. https://doi.org/10.1145/3377325.3377507

[36] Rosanna E Guadagno, Kimberly R Swinth, and Jim Blascovich. 2011. Social evaluations of embodied agents and avatars. *Computers in Human Behavior* 27, 6 (2011), 2380–2385.

[37] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* 53, 5 (2011), 517–527.

[38] David Hanson, Andrew Olney, Steve Prilliman, Eric Mathews, Marge Zielke, Derek Hammons, Raul Fernandez, and Harry Stephanou. 2005. Upending the uncanny valley. In *AAAI*, Vol. 5. ACM, New York, NY, USA, 1728–1729.

[39] Phillip M Hart, Shawn R Jones, and Marla B Royne. 2013. The human lens: How anthropomorphic reasoning varies by product complexity and enhances personal value. *Journal of Marketing Management* 29, 1-2 (2013), 105–121.

[40] Martina Hasenjäger and Heiko Wersing. 2017. Personalization in advanced driver assistance systems and autonomous vehicles: A review. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE Computer Society, USA, 1–7. https://doi.org/10.1109/ITSC.2017.8317803

[41] Álvaro Hernández-Trapote, Beatriz López-Mencía, David Díaz, Rubén Fernández-Pozo, and Javier Caminero. 2008. Embodied Conversational Agents for Voice-Biometric Interfaces. In *Proceedings of the 10th International Conference on Multimodal Interfaces* (Chania, Crete, Greece) *(ICMI '08)*. Association for Computing Machinery, New York, NY, USA, 305–312. https://doi.org/10.1145/1452392.1452454

[42] Guy Hoffman, Jodi Forlizzi, Shahar Ayal, Aaron Steinfeld, John Antanitis, Guy Hochman, Eric Hochendoner, and Justin Finkenaur. 2015. Robot Presence and Human Honesty: Experimental Evidence. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (Portland, Oregon, USA) *(HRI '15)*. Association for Computing Machinery, New York, NY, USA, 181–188. https://doi.org/10.1145/2696454.2696487

[43] Hsiao-Ying Huang, Michael Twidale, and Masooda Bashir. 2020. 'If you agree with me, do I trust you?': An examination of human-agent trust from a psychological perspective. In *Intelligent Systems and Applications: Proceedings of the 2019 Intelligent Systems Conference (IntelliSys) Volume 2*. Springer, Springer International Publishing, Cham, 994–1013.

[44] Gilhwan Hwang, Jeewon Lee, Cindy Yoonjung Oh, and Joonhwan Lee. 2019. It Sounds Like A Woman: Exploring Gender Stereotypes in South Korean Voice Assistants. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. ACM, New York, NY, USA, Article LBW2413, 6 pages. https://doi.org/10.1145/3290607.3312915

[45] Martin Jentsch, Maresa Biermann, and Evelyn Schweiger. 2019. Talking to Stupid?!? Improving Voice User Interfaces. In *Mensch und Computer 2019 - Usability Professionals*. Gesellschaft für Informatik e.V. Und German UPA e.V., Bonn.

[46] Rabia Khan and Antonella De Angeli. 2009. The attractiveness stereotype in the evaluation of embodied conversational agents. In *IFIP Conference on Human-Computer Interaction*. Springer, Springer, Heidelberg, Germany, 85–97.

[47] Rabia Fatima Khan and Alistair Sutcliffe. 2014. Attractive agents are more persuasive. *International Journal of Human-Computer Interaction* 30, 2 (2014), 142–150.

[48] Sara Kiesler, Aaron Powers, Susan R Fussell, and Cristen Torrey. 2008. Anthropomorphic interactions with a robot and robot–like agent. *Social Cognition* 26, 2 (2008), 169–181.

[49] K. Kim, L. Boelling, S. Haesler, J. Bailenson, G. Bruder, and G. F. Welch. 2018. Does a Digital Assistant Need a Body? The Influence of Visual Embodiment and Social Behavior on the Perception of Intelligent Virtual Agents in AR. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE Press, NJ, USA, 105–114. https://doi.org/10.1109/ISMAR.2018.00039

[50] Youjeong Kim and S Shyam Sundar. 2012. Anthropomorphism of computers: Is it mindful or mindless? *Computers in Human Behavior* 28, 1 (2012), 241–250.

[51] Tomoko Koda and Pattie Maes. 1996. Agents with faces: The effect of personification. In *Proceedings 5th IEEE International Workshop on Robot and Human Communication. RO-MAN'96 TSUKUBA*. IEEE, IEEE, NJ, USA, 189–194.

[52] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *HCI and Usability for Education and Work: 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008, Graz, Austria, November 20-21, 2008. Proceedings 4*. Springer, Springer International Publishing, Cham, 63–76.

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

[54] Irene Lopatovska. 2020. Personality Dimensions of Intelligent Personal Assistants. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval* (Vancouver BC, Canada) *(CHIIR '20)*. Association for Computing Machinery, New York, NY, USA, 333–337. https://doi.org/10.1145/3343413.3377993

[55] Genevieve L Lorenzo, Jeremy C Biesanz, and Lauren J Human. 2010. What is beautiful is good and more accurately understood: Physical attractiveness and accuracy in first impressions of personality. *Psychological science* 21, 12 (2010), 1777–1782.

[56] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf Between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[57] Karl F MacDorman, Robert D Green, Chin-Chang Ho, and Clinton T Koch. 2009. Too real for comfort? Uncanny responses to computer generated faces. *Computers in human behavior* 25, 3 (2009), 695–710.

[58] Angie Lorena Marin Mejia, Doori Jo, and Sukhan Lee. 2013. Designing Robotic Avatars: Are User's Impression Affected by Avatar's Age?. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction* (Tokyo, Japan) *(HRI '13)*. IEEE Press, NJ, USA, 195–196.

[59] Rachel McDonnell, Martin Breidt, and Heinrich H Bülthoff. 2012. Render me real? Investigating the effect of render style on the perception of animated virtual humans. *ACM Transactions on Graphics (TOG)* 31, 4 (2012), 1–11.

[60] Arthur G Miller. 1970. Social perception of internal-external control. *Perceptual and Motor Skills* 30, 1 (1970), 103–109.

[61] Kathleen K Molnar and Marilyn G Kletke. 1996. The impacts on user performance and satisfaction of a voice-based front-end interface for a standard software tool. *International Journal of Human-Computer Studies* 45, 3 (1996), 287–303.

[62] Masahiro Mori et al. 1970. The uncanny valley. *Energy* 7, 4 (1970), 33–35.

[63] Christine Murad and Cosmin Munteanu. 2019. "I Don't Know What You're Talking about, HALexa": The Case for Voice User Interface Guidelines. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (Dublin, Ireland) *(CUI '19)*. Association for Computing Machinery, New York, NY, USA, Article 9, 3 pages. https://doi.org/10.1145/3342775.3342795

[64] Christine Murad, Cosmin Munteanu, Leigh Clark, and Benjamin R. Cowan. 2018. Design Guidelines for Hands-Free Speech Interaction. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct* (Barcelona, Spain) *(MobileHCI '18)*. Association for Computing Machinery, New York, NY, USA, 269–276. https://doi.org/10.1145/3236112.3236149

[65] Procheta Nag and Özge Nilay Yalçın. 2020. Gender Stereotypes in Virtual Agents. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents* (Virtual Event, Scotland, UK) *(IVA '20)*. Association for Computing Machinery, New York, NY, USA, Article 41, 8 pages. https://doi.org/10.1145/3383652.3423876

[66] Clifford Nass and Kwan Min Lee. 2001. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of experimental psychology: applied* 7, 3 (2001), 171.

[67] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers Are Social Actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing*

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

[68] Jakob Nielsen. 1998. Personalization is over-rated. *Jakob Nielsen's Alertbox for October* 4 (1998), 1998.

[69] Kristine L Nowak and Christian Rauh. 2008. Choose your "buddy icon" carefully: The influence of avatar androgyny, anthropomorphism and credibility in online interactions. *Computers in Human Behavior* 24, 4 (2008), 1473–1493.

[70] Dennis Orth, Nadja Schömig, Christian Mark, Monika Jagiellowicz-Kaufmann, Dorothea Kolossa, and Martin Heckmann. 2017. Benefits of Personalization in the Context of a Speech-Based Left-Turn Assistant. In *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Oldenburg, Germany) *(AutomotiveUI '17)*. Association for Computing Machinery, New York, NY, USA, 193–201. https://doi.org/10.1145/3122986.3123004

[71] Jeni Paay, Jesper Kjeldskov, Kathrine Maja Hansen, Tobias Jørgensen, and Katrine Leth Overgaard. 2022. Digital ethnography of home use of digital personal assistants. *Behaviour & Information Technology* 41, 4 (2022), 740–758.

[72] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 640, 12 pages. https://doi.org/10.1145/3173574.3174214

[73] Alisha Pradhan, Kanika Mehta, and Leah Findlater. 2018. "Accessibility Came by Accident": Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 459, 13 pages. https://doi.org/10.1145/3173574.3174033

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

[75] Byron Reeves and Clifford Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press, New York, NY, USA.

[76] Leon Reicherts, Nima Zargham, Michael Bonfert, Yvonne Rogers, and Rainer Malaka. 2021. May I Interrupt? Diverging Opinions On Proactive Smart Speakers. In *Proceedings of the 3rd Conference on Conversational User Interfaces* (Bilbao (online), Spain) *(CUI '21)*. Association for Computing Machinery, New York, NY, USA, Article 34, 10 pages. https://doi.org/10.1145/3469595.3469629

[77] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2013. To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics* 5 (2013), 313–323.

[78] Kristin E. Schaefer, Jessie Y. C. Chen, James L. Szalma, and P. A. Hancock. 2016. A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. *Human Factors* 58, 3 (2016), 377–400. https://doi.org/10.1177/0018720816634228 arXiv:https://doi.org/10.1177/0018720816634228 PMID: 27005902.

[79] Valentin Schwind, Katrin Wolf, and Niels Henze. 2018. Avoiding the Uncanny Valley in Virtual Character Design. *Interactions* 25, 5 (Aug. 2018), 45–49. https://doi.org/10.1145/3236673

[80] Jun'ichiro Seyama and Ruth S Nagayama. 2007. The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence: Teleoperators and virtual environments* 16, 4 (2007), 337–351.

[81] William Seymour and Max Van Kleek. 2021. Exploring Interactions Between Trust, Anthropomorphism, and Relationship Development in Voice Assistants. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 371 (oct 2021), 16 pages. https://doi.org/10.1145/3479515

[82] Shradha Shalini, Trevor Levins, Erin L Robinson, Kari Lane, Geunhye Park, and Marjorie Skubic. 2019. Development and comparison of customized voice-assistant systems for independent living older adults. In *Human Aspects of IT for the Aged Population. Social Media, Games and Assistive Environments: 5th International Conference, ITAP 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26-31, 2019, Proceedings, Part II 21*. Springer, Springer International Publishing, Cham, 464–479.

[83] Harold Sigall and David Landy. 1973. Radiating beauty: Effects of having a physically attractive partner on person perception. *Journal of Personality and Social Psychology* 28, 2 (1973), 218.

[84] Selina Jeanne Sutton. 2020. Gender Ambiguous, Not Genderless: Designing Gender in Voice User Interfaces (VUIs) with Sensitivity. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (Bilbao, Spain) *(CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 11, 8 pages. https://doi.org/10.1145/3405755.3406123

[85] Akikazu Takeuchi and Taketo Naito. 1995. Situated Facial Displays: Towards Social Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '95)*. ACM Press/Addison-Wesley Publishing Co., USA, 450–455. https://doi.org/10.1145/223904.223965

[86] Adriana Tapus, Cristian Ţăpuş, and Maja J Matarić. 2008. User—robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy. *Intelligent Service Robotics* 1 (2008), 169–183.

[87] TM Van der Geest, JAGM van Dijk, WJ Pieterson, WE Ebbers, BM Fennis, NR Loorbach, MF Steehouder, E Taal, and PW de Vries. 2005. *Alter ego: State of the art on user profiling: An overview of the most relevant organisational and behavioural aspects regarding User Profiling*. Telematica Instituut, Netherlands.

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

[89] Katja Wagner and Hanna Schramm-Klein. 2019. Alexa, are you human? Investigating anthropomorphism of digital voice assistants–a qualitative approach.

[90] Isaac Wang, Jesse Smith, and Jaime Ruiz. 2019. Exploring Virtual Agents for Augmented Reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article 281, 12 pages. https://doi.org/10.1145/3290605.3300511

[91] Adam Waytz, Nicholas Epley, and John T Cacioppo. 2010. Social cognition unbound: Insights into anthropomorphism and dehumanization. *Current Directions in Psychological Science* 19, 1 (2010), 58–62.

[92] Klaus-Peter Wiedmann and Walter Von Mettenheim. 2020. Attractiveness, trustworthiness and expertise–social influencers' winning formula? *Journal of Product & Brand Management* 30, 5 (2020), 707–725.

[93] Maria Wolters, Kallirroi Georgila, Johanna D Moore, Robert H Logie, Sarah E MacPherson, and Matthew Watson. 2009. Reducing working memory load in spoken dialogue systems. *Interacting with Computers* 21, 4 (2009), 276–287.

[94] Lingyao Yuan and Alan R Dennis. 2019. Acting like humans? Anthropomorphism and consumer's willingness to pay in electronic commerce. *Journal of Management Information Systems* 36, 2 (2019), 450–477.

[95] Beste F. Yuksel, Penny Collisson, and Mary Czerwinski. 2017. Brains or Beauty: How to Engender Trust in User-Agent Interactions. *ACM Trans. Internet Technol.* 17, 1, Article 2 (Jan. 2017), 20 pages. https://doi.org/10.1145/2998572

[96] Nima Zargham, Dmitry Alexandrovsky, Jan Erich, Nina Wenig, and Rainer Malaka. 2022. "I Want It That Way": Exploring Users' Customization and Personalization Preferences for Home Assistants. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 270, 8 pages. https://doi.org/10.1145/3491101.3519843

[97] Nima Zargham, Vino Avanesi, Leon Reicherts, Ava Elizabeth Scott, Yvonne Rogers, and Rainer Malaka. 2023. "Funny How?" A Serious Look at Humor in Conversational Agents. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (Eindhoven, Netherlands) *(CUI '23)*. Association for Computing Machinery, New York, NY, USA, Article 27, 7 pages. https://doi.org/10.1145/3571884.3603761

[98] Nima Zargham, Michael Bonfert, Robert Porzel, Tanja Doring, and Rainer Malaka. 2021. Multi-Agent Voice Assistants: An Investigation Of User Experience. In *20th International Conference on Mobile and Ubiquitous Multimedia* (Leuven, Belgium) *(MUM 2021)*. Association for Computing Machinery, New York, NY, USA, 98–107. https://doi.org/10.1145/3490632.3490662

[99] Nima Zargham, Johannes Pfau, Tobias Schnackenberg, and Rainer Malaka. 2022. "I Didn't Catch That, But I'll Try My Best": Anticipatory Error Handling in a Voice Controlled Game. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 153, 13 pages. https://doi.org/10.1145/3491102.3502115

[100] Nima Zargham, Leon Reicherts, Michael Bonfert, Sarah Theres Voelkel, Johannes Schoening, Rainer Malaka, and Yvonne Rogers. 2022. Understanding Circumstances for Desirable Proactive Behaviour of Voice Assistants: The Proactivity Dilemma. In *Proceedings of the 4th Conference on Conversational User Interfaces* (Glasgow, United Kingdom) *(CUI '22)*. Association for Computing Machinery, New York, NY, USA, Article 3, 14 pages. https://doi.org/10.1145/3543829.3543834

[101] Sean Zdenek. 2007. "Just roll your mouse over me": Designing virtual women for customer service on the web. *Technical Communication Quarterly* 16, 4 (2007), 397–430.

[102] Michelle X Zhou, Gloria Mark, Jingyi Li, and Huahai Yang. 2019. Trusting virtual agents: The effect of personality. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 9, 2-3 (2019), 1–36.

[103] Jakub Zlotowski and Christoph Bartneck. 2013. The inversion effect in HRI: Are robots perceived more like humans or objects?. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, IEEE, Manhattan, New York City, 365–372.

[104] Jakub Złotowski, Hidenobu Sumioka, Friederike Eyssel, Shuichi Nishio, Christoph Bartneck, and Hiroshi Ishiguro. 2018. Model of dual anthropomorphism: the relationship between the media equation effect and implicit anthropomorphism. *International Journal of Social Robotics* 10, 5 (2018), 701–714.

Publication 7

"I Want It That Way": Exploring Users' Customization and Personalization Preferences for Home Assistants

Nima Zargham, Dmitry Alexandrovsky, Jan Erich, Nina Wenig, and Rainer Malaka

In Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22). New York, NY, USA, 2022. Association for Computing Machinery.

Personal contribution to this work: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, validation, visualization, and contribution to all parts of the manuscript.

ISBN: 978-1-4503-9156-6/22/04 DOI: 10.1145/3491101.3519843

"I Want It That Way": Exploring Users' Customization and Personalization Preferences for Home Assistants

Nima Zargham
zargham@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

Dmitry Alexandrovsky
dimi@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

Jan Erich
janerich@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

Nina Wenig
nwenig@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

Rainer Malaka
malaka@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

ABSTRACT

Home assistants are becoming a widespread product, but they mostly come as a compact device and offer very few customization and personalization features, which often leads to dissatisfaction. With the technological advancements, these systems are becoming more adaptable to the users' needs and can better imitate a human personality. To achieve that efficiently, understanding how different users envision their desired assistant is crucial. To identify people's customization and personalization preferences and their desired personality for a home assistant, we designed a set of storyboards depicting a variety of possible features in a domestic setting and conducted a user study ($N = 15$), including a series of semi-structured interviews. Our quantitative results suggest that users prefer an agent which is highly agreeable and has higher conscientiousness and emotional stability. Furthermore, we discuss users' customization and personalization preferences for a home assistant, which could be considered when designing the future generation of home assistants.

CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**; *User studies.*

KEYWORDS

Voice Assistants, Home Assistants, Personality, Big Five, Customization, Personalization

ACM Reference Format:
Nima Zargham, Dmitry Alexandrovsky, Jan Erich, Nina Wenig, and Rainer Malaka. 2022. "I Want It That Way": Exploring Users' Customization and Personalization Preferences for Home Assistants. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '22 Extended Abstracts), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3491101.3519843

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

ACM ISBN 978-1-4503-9156-6/22/04…$15.00
https://doi.org/10.1145/3491101.3519843

1 INTRODUCTION

The use of voice assistants (VAs) is spreading widely and home assistants such as Apple HomePod, Google Home, or Amazon Echo are becoming increasingly common [10]. VAs are designed to provide assistance whenever they are called upon for various different tasks such as smart home control [26, 38, 46], scheduling [50, 59], navigation [29, 51], education [5, 57], and entertainment [4, 62].

Despite all the benefits that such systems provide, it still remains a difficult task to design a fulfilling experience for users as the interaction with voice-controlled systems is often characterized as troublesome [46], disappointing [37], or embarrassing [16]. This is partially due to limitations in speech recognition and partially because these systems do not fulfill the user's expectations [32, 37, 41, 63]. Due to the individual differences in preferences for a desired home assistant, systematic adaptation of these systems to the user is challenging [55]. Therefore, commonly available devices tend to take a one-size-fits-all approach, which ignores the possible advantages of adapting to user preferences. Moreover, previous research has shown that users unconsciously assign personalities to conversational agents [48], and purposefully manipulating this perception can affect users' trust and engagement with such devices [14, 64].

Customization and personalization features have been shown to enhance user satisfaction and improve performance [18, 20, 21, 40, 42]. Customized features are those the user can explicitly select between specific options, whereas personalized features are driven by computers based on users' individual needs [44].

In this work, we present a human-centered approach [11] intending to identify customization and personalization preferences of home assistants, as well as desired personality types of the agent. More specifically, this work is guided by the following research questions:

- RQ1: How do users imagine the personality of their desired home assistant?
- RQ2: In which ways do users want to customize and personalize their home assistants?

We conducted semi-structured interviews using storyboards that contained scenarios of everyday domestic situations. Our results

show that users prefer an agent with high *agreeableness*, *conscientiousness*, and *emotional stability*. We also discuss users' customization and personalization preferences for a home assistant and their implications for future research avenues of personalized assistants.

2 RELATED WORK

There has been extensive research regarding home assistants and how people interact with such devices [8, 19, 36]. In this section, we discuss research on voice assistant customization, personalization, and the importance of personality in human-agent interaction.

2.1 VA Customization

Customization describes the extent to which technology or service can be modified to comply with user preferences [31, 49]. In contrast, *personalization* refers to automatic adaptation to users' needs based on observed behaviors [31]. Providing users with control over interaction can improve performance and user satisfaction [58]. A large body of research has highlighted the importance of customization in VAs, particularly for people with special needs [1, 2, 40, 42, 58]. Molnar and Kletke [40] emphasize that the lack of flexibility reduces productivity and satisfaction. Murad et al. [42] identified lack of control as a common cause of user frustration and highlighted the need for user control and freedom in speech interfaces. Therefore, to enhance the experience, VAs should allow for the configuration of speed, tone, and volume along with other characteristics of the virtual agents [2]. Furthermore, Zargham et al. [61] observed users' desire to customize additional design factors of home assistants beyond voice characteristics, such as the number of agents and their roles and personalities. Moreover, allowing users to customize the agent's appearance to their liking could potentially improve the interaction with such systems [10].

2.2 VA Personalization

Asif and Krogstie [7] define personalization as "a controlled process of adaptation of a service to achieve a particular goal by utilizing the user model and the context of use." The user model refers to the recorded data, which includes user information with the aim to adapt systems to the individual needs [52]. User models can be created with the user's own entries or systems automatically adapting to user behavior, which is also referred to as direct and indirect user input [39]. However, while direct user inputs might be very advantageous in the first place, an inappropriately set assistant may be less accepted by a user than its default configuration [12]. Therefore, many systems additionally apply indirect adaptation of user behavior [39, 66]. Personalization yields a number of benefits to business owners and customers since it allows for higher efficiency and user acceptance for a product or service [30, 45]. Studies have shown that users prefer VAs that can adapt to their preferences and background [12, 20–22, 34]. People tend to find it easier to interact with technologies that partially resemble their own characteristics [13]. A study by Braun et al. [12] has shown that users prefer and trust a personalized in-car VA more than the default version, especially when the agent's personality matches the user's.

2.3 Personality and VAs

Many innovative technologies, including home assistants, tend to be seen as social actors in general [43], which describes how users apply social rules and assign them personalities [48] while interacting with such systems. An extensive body of work has examined the role of personality in human-computer interaction [9, 24, 25, 27, 47, 48]. The Five-Factor Model (FFM), also known as the big five [33] is the most widely accepted personality theory in scientific research for modeling human personality [24, 25, 53] which boils down human personality to five core factors of *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism*.

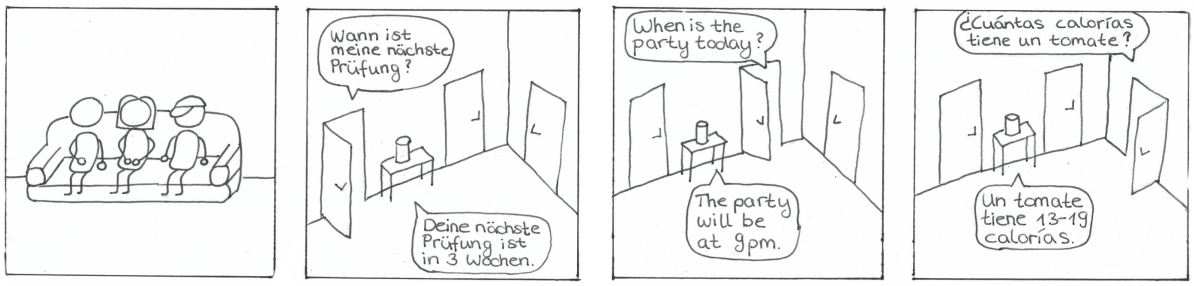
Personality has been described as an important aspect for designing and adapting voice assistants [47]. While the personality of the user has an impact on their desired agent, designing the agent's personality is also crucial. Research has shown that distinctive agent personalities could be successful in promoting trust and user acceptance [6, 9, 27, 60]. For instance, Chen et al. [17] suggest that users prefer an extroverted agent over an introverted one and are more talkative towards it. Moreover, the authors found that agents' personality traits influence users' preferences, dialogues, and behavior. Andrist et al. [6] showed that matching the agent's personality to the user's has a positive effect on users' motivation and willingness to interact with the agent. For chatbots, Völkel and Kaya [54] found that users with higher agreeableness prefer an agreeable agent. Understanding user perceptions and preferences for the personality of the agent can improve user experience [35].

Eventually, customization and personalization features of voice user interfaces (VUIs) have been shown to enhance the user experience with these devices. Moreover, designing specific personalities for agents could also impact the experience. Nevertheless, there is still a gap in terms of understanding users' preferences for such agent features.

To build upon the prior work, we engage users to think about possible features of home assistants to shed further light on how to build an ideal agent. We specifically focus on users' customization and personalization preferences for home assistants and how they envision the personality of their desired agent with respect to their own personality.

3 STUDY DESIGN

We conducted a user study to explore users' customization and personalization preferences, as well as their desired personalities for home assistants. For this, we used an approach inspired by scenario-based design methods [15] and vignette experiments [3], where we present a series of hypothetical situations to the participants and ask them to reflect on the scenarios. This approach is similar to the so-called Speed Dating method [23, 65] and allows investigating technologies despite current technical limitations. Using this method, designers can identify possible problems as well as opportunities regarding specific technologies, and create more appropriate and innovative solutions [23]. Moreover, a variety of use cases can be considered which do not necessarily apply to the participant but are relevant for other users, e.g., impairments, medical conditions, or social contacts. We used graphical storyboards to visualize the situational context and spatial relation between the characters and the assistant.



A German, an English, and a Spanish person are living together in an apartment. Every one of them has set their own profile on the home assistant.

The German person asks a question from the assistant and gets a response in German.

The American person asks a question from the assistant and gets a response in English.

The Spanish person asks a question from the assistant and gets a response in Spanish.

Figure 1: Exemplary storyboard that was used in the interview in which the home assistant responds differently to each household member.

Scenarios and Storyboards. Iteratively and in discussions, we developed scenarios within the context of the home environment involving either a single person or multiple people. First, in brainstorming sessions between two researchers, 24 scenarios were conceived covering different aspects in a domestic setting. The scenarios were based on what researchers imagined being a compelling customization/personalization feature of home assistants in everyday situations which they were familiar with themselves, or they knew from others. After three iterations, we selected ten scenarios from the initial list covering different features, made final adjustments, and created the storyboards for the chosen scenarios. We classified the scenarios based on the type of assistance and the number of people involved. For instance, in the scenario shown in Figure 1, the agent's type of assistance is fact-checking, and the people involved are the three people who live in the same apartment, making it a multi-party scenario. Using these storyboards, we ran a pilot study with two participants to find possible flaws within the study design and to see if the scenarios inspired participants to reflect on. We worked with different styles of storyboards in an iterative process with the aim to minimize the cultural and ethnic cues so that the participants can put themselves in the shoes of the characters. The appearance of the home assistants was similar to a common cylinder shaped smart speaker available on the market. Furthermore, the characters within the storyboards did not have face expressions in an attempt to avoid any influence from the reactions of the characters. An exemplary scenario with the final design is depicted in Figure 1. The complete set of storyboards designed and used for this research can be found in the supplementary materials. In the following, we briefly describe the scenarios:

- *S1 Multilingual*: The assistant can understand multiple languages simultaneously and can respond accordingly.
- *S2 Pace Customization*: The user adjusts the speaking pace of the assistant to ease the use for an elderly user.
- *S3 Speaking Adaptation*: The assistant picks up the speaking characteristics of the user and responds similarly.
- *S4 Limited Access*: The user limits the interaction with the assistant in a way that it only responds to the owner.
- *S5 Interest Adaptation*: The assistant learns about the users' interests and uses that when responding to inquiries.
- *S6 Mood Adaptation*: The assistant senses the users' current mood and adjusts its responses based on that.
- *S7 Multi-Persona*: The assistant is set with multiple characters with distinct personalities and areas of responsibility.
- *S8 Self Embodiment*: The user sets their embodied agent to look like them.
- *S9 Shared Assistant*: Two siblings use the same personalized agent that is only capable of having one user profile.
- *S10 Private Information*: The user sets the assistant to share certain information only with the owner.

Procedure. The interview sessions were held remotely via video calls, where the experimenter recorded verbal statements and observations while providing assistance in cases of issues. The screen of the interviewer was shared to display the storyboards. In the beginning of the session, the participants were briefed about the procedure of the experiment. After giving informed consent, all participants filled out a demographics questionnaire and stated their prior experience with home assistants as well as their typical usage of VAs. Then, the participants stated their personality traits using the Ten Item Personality Inventory (TIPI) [28]. TIPI is a 10-item questionnaire that assesses personality according to the Five-Factor Model. Each item is rated on a 7-point Likert scale ranging from 1 (disagree strongly) to 7 (agree strongly). After this, the semi-structured interviews started where in random order, the participants were presented with all ten scenarios one by one and commented on each story. First, participants were asked to give their impressions about the individual scenarios, and point out the positive and negative aspects of the interaction with the home assistant. Afterwards, they provided overall suggestions and recommendations in terms of customization and personalization, as well as the desired personality of a home assistant. Finally, the participants filled in a second TIPI where they stated the desired personality of the voice agent. For further, qualitative analysis, we recorded the audio of each session. A single session took between 50-75 minutes.

Participants. 15 participants at the age between 22 and 41 ($M = 26.6$, $SD = 4.29$) volunteered for our study (4 self-identified as female, 11

male). All participants had previous experience with home assistants (4 rarely, 11 often), 12 owned a smart speaker. We conducted the experiment in the local official language to avoid language barriers.

4 RESULTS

We performed a thematic content analysis to evaluate patterns within participants' responses. The transcripts of the interviews were independently coded by two researchers using inductive coding, where a single quote could be assigned to multiple codes. The researchers organized, reviewed, and discussed the codes, resolved disagreements, and derived themes which were categorized into (i) *agent's speech characteristics*, (ii) *agent's visualization*, (iii) *agent's personality*, and (iv) *Privacy and Security*. In this section, we first present the analysis of the personality questionnaires and then discuss the interview findings.

Personality. To investigate the self-personality differs significantly from the desired personality of the home assistant, we performed paired-sample t-tests on each subscale of the TIPI between both responses. The descriptive statistics and the results of the t-tests are shown in Table 1. The analysis identified significant differences for agreeableness, conscientiousness, and emotional stability between the self-reported personality and the desired personality.

To determine if the desired personality of the home assistant is in line with the self-personality of the users, we performed correlation analysis between the two ratings on all subscales of TIPI. The analysis showed several significant correlations. A positive correlation was found between the agent's *agreeableness* and the users' *agreeableness* with a Pearson's coefficient of $r = .550$ ($p = .033$). Further, the agent's *emotional stability* negatively correlated with users' *openness to experiences* with $r = -.556$ ($p = .031$).

Interview Analysis. Overall, participants demanded more customization and personalization features for a home assistant. All but one participant mentioned that the personalization features of the current home assistants are not sufficient. 13 participants would be willing to pay extra for a system that can be more personalized. 12 people mentioned that they would use a better customized home assistant more often. The qualitative analysis of the interviews identified four overarching themes: *agent's speech characteristics*, *agent's visualization*, *agent's personality*, and *Privacy and Security*. We report the results of the interview based on the categories.

Agent's Speech Characteristics. Participants found it highly useful to be able to modify the speed, tone, and volume of the agent's voice, and participants demanded to have the ability to adjust the voice themselves. However, users had concerns regarding the difficulty of making such adjustments ("is it adjustable by my grandmother?"). Eight participants said that the agent's voice should sound more human-like and were not satisfied with their current agent's voice. Eleven participants preferred a specific gender for their desired agent's voice. Participants also said that their desired agent's voice should sometimes convey certain emotions such as excitement, sarcasm, friendliness, and calmness. On the other hand, two participants demanded a neutral voice which doesn't convey any specific emotions. Six Participants mentioned that celebrity voices could be used for the agent as an entertaining feature. One user found this as "momentary entertainment", while one participant mentioned "I would not like to use a voice that I know from anywhere, it makes me uncomfortable". All participants mentioned that the assistant has to understand different languages to be able to assist more people, and three participants suggested it could encourage users to learn foreign languages. Five users mentioned changing the language from the settings is not sufficient and this should be done automatically.

Agent's Visualization. Six participants preferred the agent to have a digital embodiment, where the other nine did not find it helpful. From those who preferred an embodied agent, all wanted a humanoid representation, and two participants mentioned that the visual representation should match the voice. Two participants mentioned that the agent should look healthy and fit. Three participants specifically asked for a female embodiment for the agent. For most users, the humanoid clothing wasn't that important. However, two users preferred the assistant to occasionally change its outfit. Regarding the agent's presence on the screen, six participants preferred that the agent is constantly present. Users said comments such as "If I see the assistant, it makes me want to talk to it" or "I think it's a good reminder that the assistant is there, sometimes I may forget I have one if it is not speaking". However, the other nine participants preferred the assistant to appear only when asked. One user said, "I would feel observed if it's constantly there".

Personality Adaptation. 12 participants were in favor of agent's adapting to the user's personality. One participant said, "it should adjust the information content based on my desires, not necessarily behave like me". Another suggested "people change, so the agent should be able to change and keep up with the changes in my interest". Eight participants were in favor of mood adjustments by the agent, where the others were resistant against this feature. Participants in favor mentioned that "it makes the device seem more human". On the other hand, users also said "I don't need a piece of software to show me empathy, I know it's programmed". One user had doubts regarding the VA's judgment "If it misjudges my mood, it would work horribly wrong", while another participant feared that "this might make people rely on such devices and have less interaction with real humans". Ten participants wanted the agent to use humor, while two suggested that the agent should be serious and prevent using humor.

Privacy and Security. Eleven people mentioned that they would trust the VA more if it provides more security and privacy related features. All participants were concerned regarding the agent sharing personal information in front of others. They all agreeably mentioned that an access control is a helpful feature that home assistants could provide, meaning that the home assistant identifies the role of the users and responds accordingly. A participant mentioned that this could be very helpful for families with small children. One user stated that "the system should be able to respond appropriately to the users that are not the owner". Another said, "It should talk to friends and family the same as to me, but talk more seriously to my colleagues". One user suggested that "The agent should distinguish between age groups, talk differently to kids than to adults". Participants had different suggestions regarding sharing

Table 1: Descriptive statistics and results of paired-sample t-tests (df=14) of TIPI ratings for self-personality and desired personality of the VA.

	Self-Personality M (SD)	Agent-Personality M (SD)	$t(14)$	$p-value$	Cohen's d
Extroversion	4.13 (1.18)	4.63 (1.07)	1.3693	0.1925	0.443
Agreeableness	4.70 (1.42)	5.56 (0.96)	2.792	0.0144	0.712
Conscientiousness	5.20 (1.39)	6.63 (0.71)	3.930	< 0.01	1.285
Emotional Stability	4.80 (1.08)	6.20 (0.64)	3.7936	< 0.01	1.568
Openness to Experiences	5.40 (1.22)	5.40 (1.27)	0.0	1.0	0

private information. One believed that the system should include standard and private features. Another said that "private information should require a fingerprint/password". One user proposed that "the system should have specific modes where the microphone is deactivated for a specific time or in specific rooms".

5 DISCUSSION

The results from our scenario-based study suggest that participants are not satisfied with the customization and personalization aspects of current home assistants and demand more of such features.

Regarding the personality of the agent (RQ1), users rated equally or higher in all the five scales of the TIPI for their desired agent's personality in comparison to their own personality ratings. We found significantly higher ratings in terms of *agreeableness*. Furthermore, for this subscale, a strong positive correlation between the users' and their desired agent's was observed. This finding further supports previous work by Völkel and Kaya [54] where they found that users with higher agreeableness prefer an agreeable agent. Participants also rated significantly higher in subscales of *conscientiousness* and *emotional stability* for the agent in comparison to self-personality. These subscales constitute reliability, which is an important factor in human-agent communication. It is not surprising that users expect such systems to behave reliably, since they have to trust the agent with their tasks. This personality trait was even more crucial for users who rated themselves lower on *openness to experiences*. This finding is supported by the negative correlation between the agent's *emotional stability* and the users' *openness to experiences*.

Concerning **RQ2**, the qualitative analysis revealed interesting tendencies towards different customization and personalization features. In line with previous work [1, 18], being able to modify an agent's voice characteristics such as gender, speed, tone, and volume was highly favored by the participants. Most users preferred to control and adjust such characteristics themselves, rather than it being a personalization feature. However, some raised concerns about the difficulty of such adjustments and demanded a simple user interface. Commonly, the customization features of the current home assistants, require users to use a graphical interface such as their smartphone to configure the VA. On many occasions, users have difficulties to find the dedicated section in the app for specific features. Further, in homes with multiple household members, if the person whose smartphone is connected to the device is not available, such modifications would not be possible. These obstacles could frustrate users and make the customization process rather complex. A number of our participants who owned a smart speaker requested features that are already available in their systems. This underlines that users are often not aware of the existing customization and personalization features.

In line with Andrist et al. [6], participants were mostly in favor of the agent matching the user's personality and adapting to their interests, and users would highly benefit from such personalized systems. However, users expected the agent to evolve with any changes in the users' interests. Although most participants wanted the assistant to show emotions, be more humorous, sense users' mood and adjust the responses to it, and overall behave more human-like, some were still rather hesitant towards such features. The main reason against the human-like behavior of the agent was the mistrust in the technology. Users believed if the agent falsely recognizes users' mood and responds inappropriately, it would affect the users extremely negatively. Another concern raised was that some users did not like to see human-like features in an agent and considered such interactions as fake. These users wanted to see a clear distinction between a human and a software. Such agent capabilities could be attractive personalization features to consider when designing the future home assistants. Nevertheless, since such social skills are complex abilities that are still very difficult for computer systems to master, we recommend providing customizable features in which the users are in charge of setting when and how the agent should use such skills.

Features which could potentially improve accessibility, such as supporting multiple languages simultaneously, were considered to be highly valuable by the participants. A number of users proposed creative and fun ideas such as having the agent sound or look like a celebrity such as musicians, actors, or athletes. Six participants also showed interest in having a digital embodiment for the agent, all of which preferred a humanoid. Modifying further details regarding the agent's representation such as hairstyle, clothing and body size was also requested by our participants. Moreover, some users wanted the agent to occasionally change its visualization, suggesting the agent has a daily life. However, similar to Bonfert et al. [10], participants raised concerns regarding the constant presence of an embodied agent. Consequently, since we learned that users' preferences regarding agent embodiment often vary and no universal solution could satisfy the expectations of all users equally. Therefore, we recommend having the digital embodiment as a customizable feature.

Furthermore, we observed that privacy concerns are still one of the main challenges of home assistants. Users want to have more

privacy related features in order to have greater trust and to avoid inconvenient situations. Features such as having individual user profiles, user roles, and access control were suggested to better protect users' privacy. Beyond that, the participants wanted to have different modes for particular circumstances where the agent varies the language depending on the persons in the room to avoid inconvenient or inappropriate situations. Future home assistants would presumably require collecting more personal data from users in order to provide better services. Therefore, we encourage developers to consider such privacy features in the design of future generations of these systems.

5.1 Limitations and Future Work

Although the findings of this work present important considerations for designing home assistants, Our research and the findings are still limited in several ways which need to be addressed. Firstly, in this study, we used a limited set of scenarios covering only a specific range of possible features. As a result, our participants may not fully understand the range of services or possibilities for customization and personalization features of a home assistant. In order to scale up the insights of our work, future research can expand these scenarios by investigating other possible features. Moreover, since many of the features used in the storyboards are not available in the market, we explored people's perceptions of such features by presenting hypothetical scenarios, meaning that the participants did not actually experience the situation. One can assume that participants' feelings and perceptions about these features could differ once they interact with them in real life. Therefore, future work should investigate the user experience in the scenarios while interacting with the home assistant. The experiment sample mostly consisted of male users (11 out of 15). An influence of such bias on the results cannot be excluded. Hence, future studies should validate our findings by investigating a wider population. We will expand our experiment design and investigate closer how different demographic factors such as gender and cultural background corroborate our results. Lastly, a recent study suggests that the big five model may not be applicable to describe the personality of a conversational agent, since this model was derived from human language use in order to describe human personality [56]. However, in our work, we consider a concept of future VA technology that aims to adapt human-like behavior. Therefore, while agreeing with Völkel et al. that the FFM may not be suitable to assess the current state-of-the art VAs, we argue that it is the most eligible method to describe human-like behavior of VAs. In our future studies, we plan to develop a home assistant which conveys the desired agent personality based on the present results and evaluate it in terms of usability and UX.

6 CONCLUSION

In this work, we explored users' customization and personalization preferences for home assistants as well as the desired personality of the agent. We created ten scenarios which were depicted in storyboards that demonstrate how VAs might act in different domestic situations and interviewed 15 participants. The results show that users prefer an agent which is highly agreeable, and has high conscientiousness and emotional stability. Furthermore, the interviews yielded four categories of features users wish for home assistants (*agent's speech characteristics*, *agent's visualization*, *agent's personality*, and *Privacy and Security*). These findings highlight that current assistants contain several limitations and provide important suggestions and considerations which could guide the design of future generation home assistants in order to achieve a higher user satisfaction.

ACKNOWLEDGMENTS

This work was partially funded by the FET-Open Project 951846 "MUHAI – Meaning and Understanding for Human-centric AI" funded by the EU program Horizon 2020, as well as the German Research Foundation DFG as part of Collaborative Research Center (Sonderforschungsbereich) 1320 "EASE – Everyday Activity Science and Engineering", University of Bremen (http://www.ease-crc.org/) conducted in subproject H02.

REFERENCES

[1] Ali Abdolrahmani, Ravi Kuber, and Stacy M Branham. 2018. " Siri Talks at You" An Empirical Investigation of Voice-Activated Personal Assistant (VAPA) Usage by Individuals Who Are Blind. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. Association for Computing Machinery, New York, NY, USA, 249–258.

[2] Ali Abdolrahmani, Kevin M. Storer, Antony Rishin Mukkath Roy, Ravi Kuber, and Stacy M. Branham. 2020. Blind Leading the Sighted: Drawing Design Insights from Blind Users towards More Productivity-Oriented Voice Interfaces. *ACM Trans. Access. Comput.* 12, 4, Article 18 (Jan. 2020), 35 pages. https://doi.org/10.1145/3368426

[3] Herman Aguinis and Kyle J. Bradley. 2014. Best Practice Recommendations for Designing and Implementing Experimental Vignette Methodology Studies. *Organizational Research Methods* 17, 4 (2014), 351–371. https://doi.org/10.1177/1094428114547952 arXiv:https://doi.org/10.1177/1094428114547952

[4] Fraser Allison, Joshua Newn, Wally Smith, Marcus Carter, and Martin Gibbs. 2019. Frame Analysis of Voice Interaction Gameplay. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300623

[5] James N Anderson, Nancie Davidson, Hazel Morton, and Mervyn A Jack. 2008. Language learning with interactive virtual agent scenarios and speech recognition: Lessons learned. *Computer Animation and Virtual Worlds* 19, 5 (2008), 605–619.

[6] 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*. Association for Computing Machinery, New York, NY, USA, 3603–3612.

[7] Muhammad Asif and John Krogstie. 2012. Taxonomy of personalization in mobile services. In *Proceedings of 10th IADIS International Conference e-Society*. IADIS – International Association for the Development of the Information Society, -, 343–350.

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

[9] Emily P Bernier and Brian Scassellati. 2010. The similarity-attraction effect in human-robot interaction. In *2010 IEEE 9th International Conference on Development and Learning*. IEEE, IEEE Computer Society, USA, 286–290.

[10] Michael Bonfert, Nima Zargham, Florian Saade, Robert Porzel, and Rainer Malaka. 2021. An Evaluation of Visual Embodiment for Voice Assistants on Smart Displays. In *CUI 2021-3rd Conference on Conversational User Interfaces*. ACM, New York, NY, USA, 1–11.

[11] Guy Boy. 2012. *Orchestrating human-centered design*. Springer Science & Business Media, Berlin/Heidelberg.

[12] Michael Braun, Anja Mainz, Ronee Chadowitz, Bastian Pfleging, and Florian Alt. 2019. At your service: Designing voice assistant personalities to improve automotive user interfaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–11.

[13] Cynthia Breazeal. 2003. Toward sociable robots. *Robotics and autonomous systems* 42, 3-4 (2003), 167–175.

[14] Angelo Cafaro, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2016. First impressions in human–agent virtual encounters. *ACM Transactions on Computer-Human Interaction (TOCHI)* 23, 4 (2016), 1–40.

[15] John M. Carroll. 1999. Five Reasons for Scenario-Based Design. In *Proceedings of the Thirty-Second Annual Hawaii International Conference on System Sciences-Volume 3 - Volume 3 (HICSS '99)*. IEEE Computer Society, USA, 3051.

[16] Marcus Carter, Fraser Allison, John Downs, and Martin Gibbs. 2015. Player Identity Dissonance and Voice Interaction in Games. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play* (London, United Kingdom) *(CHI PLAY '15)*. Association for Computing Machinery, New York, NY, USA, 265–269. https://doi.org/10.1145/2793107.2793144

[17] Yuting Chen, Adeel Naveed, and Robert Porzel. 2010. Behavior and Preference in Minimal Personality: A Study on Embodied Conversational Agents. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction* (Beijing, China) *(ICMI-MLMI '10)*. Association for Computing Machinery, New York, NY, USA, Article 49, 4 pages. https://doi.org/10.1145/1891903.1891963

[18] Dasom Choi, Daehyun Kwak, Minji Cho, and Sangsu Lee. 2020. *"Nobody Speaks That Fast!" An Empirical Study of Speech Rate in Conversational Agents for People with Vision Impairments.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376569

[19] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, and et al. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article 475, 12 pages. https://doi.org/10.1145/3290605.3300705

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

[21] Benjamin R Cowan, Derek Gannon, Jenny Walsh, Justin Kinneen, Eanna O'Keefe, and Linxin Xie. 2016. Towards Understanding How Speech Output Affects Navigation System Credibility. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, USA, 2805–2812.

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

[23] Scott Davidoff, Min Kyung Lee, Anind K Dey, and John Zimmerman. 2007. Rapidly exploring application design through speed dating. In *International Conference on Ubiquitous Computing*. Springer, Springer, Berlin/Heidelberg, 429–446.

[24] Boele De Raad. 2000. *The big five personality factors: the psycholexical approach to personality.* Hogrefe & Huber Publishers, Göttingen, Germany.

[25] Colin G DeYoung. 2015. Openness/intellect: A dimension of personality reflecting cognitive exploration. *APA handbook of personality and social psychology* 4 (2015), 369–399.

[26] Jide S. Edu, Jose M. Such, and Guillermo Suarez-Tangil. 2020. Smart Home Personal Assistants: A Security and Privacy Review. *ACM Comput. Surv.* 53, 6, Article 116 (dec 2020), 36 pages. https://doi.org/10.1145/3412383

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

[28] Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in personality* 37, 6 (2003), 504–528.

[29] John HL Hansen, Xianxian Zhang, Murat Akbacak, Umit H Yapanel, Bryan Pellom, Wayne Ward, and Pongtep Angkititrakul. 2005. CU-MOVE: Advanced in-vehicle speech systems for route navigation. In *DSP for in-vehicle and mobile systems*. Springer, Berlin/Heidelberg, 19–45.

[30] Martina Hasenjäger and Heiko Wersing. 2017. Personalization in advanced driver assistance systems and autonomous vehicles: A review. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE Computer Society, USA, 1–7. https://doi.org/10.1109/ITSC.2017.8317803

[31] Chen-Wei Hsieh and Sherry Y Chen. 2016. A cognitive style perspective to handheld devices: Customization vs. personalization. *International Review of Research in Open and Distributed Learning* 17, 1 (2016), 1–22.

[32] Martin Jentsch, Maresa Biermann, and Evelyn Schweiger. 2019. Talking to Stupid?!? Improving Voice User Interfaces. In *Mensch und Computer 2019 - Usability Professionals*. Gesellschaft für Informatik e.V. Und German UPA e.V., Bonn.

[33] Oliver P John, Sanjay Srivastava, et al. 1999. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research* 2, 1999 (1999), 102–138.

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

[35] Irene Lopatovska. 2020. Personality Dimensions of Intelligent Personal Assistants. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval* (Vancouver BC, Canada) *(CHIIR '20)*. Association for Computing Machinery, New York, NY, USA, 333–337. https://doi.org/10.1145/3343413.3377993

[36] Silvia Lovato and Anne Marie Piper. 2015. "Siri, is This You?": Understanding Young Children's Interactions with Voice Input Systems. In *Proceedings of the 14th International Conference on Interaction Design and Children* (Boston, Massachusetts) *(IDC '15)*. ACM, New York, NY, USA, 335–338. https://doi.org/10.1145/2771839.2771910

[37] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf Between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[38] Graeme McLean and Kofi Osei-Frimpong. 2019. Hey Alexa… examine the variables influencing the use of artificial intelligent in-home voice assistants. *Computers in Human Behavior* 99 (2019), 28–37.

[39] Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava. 2000. Automatic personalization based on web usage mining. *Commun. ACM* 43, 8 (2000), 142–151.

[40] Kathleen K Molnar and Marilyn G Kletke. 1996. The impacts on user performance and satisfaction of a voice-based front-end interface for a standard software tool. *International Journal of Human-Computer Studies* 45, 3 (1996), 287–303.

[41] Christine Murad and Cosmin Munteanu. 2019. "I Don't Know What You're Talking about, HALexa": The Case for Voice User Interface Guidelines. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (Dublin, Ireland) *(CUI '19)*. Association for Computing Machinery, New York, NY, USA, Article 9, 3 pages. https://doi.org/10.1145/3342775.3342795

[42] Christine Murad, Cosmin Munteanu, Leigh Clark, and Benjamin R. Cowan. 2018. Design Guidelines for Hands-Free Speech Interaction. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct* (Barcelona, Spain) *(MobileHCI '18)*. Association for Computing Machinery, New York, NY, USA, 269–276. https://doi.org/10.1145/3236112.3236149

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

[44] Jakob Nielsen. 1998. Personalization is over-rated. *Jakob Nielsen's Alertbox for October* 4 (1998), 1998.

[45] Dennis Orth, Nadja Schömig, Christian Mark, Monika Jagiellowicz-Kaufmann, Dorothea Kolossa, and Martin Heckmann. 2017. Benefits of Personalization in the Context of a Speech-Based Left-Turn Assistant. In *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Oldenburg, Germany) *(AutomotiveUI '17)*. Association for Computing Machinery, New York, NY, USA, 193–201. https://doi.org/10.1145/3122986.3123004

[46] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 640, 12 pages. https://doi.org/10.1145/3173574.3174214

[47] Alisha Pradhan and Amanda Lazar. 2021. Hey Google, Do You Have a Personality? Designing Personality and Personas for Conversational Agents. In *CUI 2021 - 3rd Conference on Conversational User Interfaces* (Bilbao (online), Spain) *(CUI '21)*. Association for Computing Machinery, New York, NY, USA, Article 12, 4 pages. https://doi.org/10.1145/3469595.3469607

[48] Byron Reeves and Clifford Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places.* Cambridge University Press, New York, NY, USA.

[49] Ching-I Teng. 2010. Customization, immersion satisfaction, and online gamer loyalty. *Computers in Human Behavior* 26, 6 (2010), 1547–1554.

[50] Varun Tiwari, Mohammad Farukh Hashmi, Avinash Keskar, and NC Shivaprakash. 2020. Virtual home assistant for voice based controlling and scheduling with short speech speaker identification. *Multimedia tools and applications* 79, 7 (2020), 5243–5268.

[51] Omer Tsimhoni, Daniel Smith, and Paul Green. 2004. Address entry while driving: Speech recognition versus a touch-screen keyboard. *Human factors* 46, 4 (2004), 600–610.

[52] TM Van der Geest, JAGM van Dijk, WJ Pieterson, WE Ebbers, BM Fennis, NR Loorbach, MF Steehouder, E Taal, and PW de Vries. 2005. *Alter ego: State of the art on user profiling: An overview of the most relevant organisational and behavioural aspects regarding User Profiling.* Telematica Instituut, Netherlands.

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

https://doi.org/10.1145/3411764.3445536

[54] Sarah Theres Völkel and Lale Kaya. 2021. Examining User Preference for Agreeableness in Chatbots. In *CUI 2021 - 3rd Conference on Conversational User Interfaces* (Bilbao (online), Spain) *(CUI '21)*. Association for Computing Machinery, New York, NY, USA, Article 38, 6 pages. https://doi.org/10.1145/3469595.3469633

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

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

[57] Rainer Winkler, Matthias Söllner, Maya Lisa Neuweiler, Flavia Conti Rossini, and Jan Marco Leimeister. 2019. Alexa, Can You Help Us Solve This Problem?: How Conversations With Smart Personal Assistant Tutors Increase Task Group Outcomes. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. ACM, New York, NY, USA, Article LBW2311, 6 pages. https://doi.org/10.1145/3290607.3313090

[58] Maria Wolters, Kallirroi Georgila, Johanna D Moore, Robert H Logie, Sarah E MacPherson, and Matthew Watson. 2009. Reducing working memory load in spoken dialogue systems. *Interacting with Computers* 21, 4 (2009), 276–287.

[59] Benjamin A Wong, Thad Starner, and R Martin McGuire. 2002. *Towards Conversational Speech Recognition for a Wearable Computer Based Appointment Scheduling Agent.* Technical Report. Georgia Institute of Technology.

[60] Sarah Woods, Kerstin Dautenhahn, Christina Kaouri, Renete Boekhorst, and Kheng Lee Koay. 2005. Is this robot like me? Links between human and robot personality traits. In *5th IEEE-RAS International Conference on Humanoid Robots, 2005.* IEEE, IEEE Computer Society, USA, 375–380.

[61] Nima Zargham, Michael Bonfert, Robert Porzel, Tanja Doring, and Rainer Malaka. 2021. Multi-Agent Voice Assistants: An Investigation Of User Experience. In *20th International Conference on Mobile and Ubiquitous Multimedia* (Leuven, Belgium) *(MUM 2021)*. Association for Computing Machinery, New York, NY, USA, 98–107. https://doi.org/10.1145/3490632.3490662

[62] Nima Zargham, Michael Bonfert, Georg Volkmar, Robert Porzel, and Rainer Malaka. 2020. Smells Like Team Spirit: Investigating the Player Experience with Multiple Interlocutors in a VR Game.

[63] Nima Zargham, Johannes Pfau, Tobias Schnackenberg, and Rainer Malaka. 2022. "I Didn't Catch That, But I'll Try My Best": Anticipatory Error Handling in a Voice Controlled Game. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22).* Association for Computing Machinery, New York, NY, USA, –.

[64] Michelle X Zhou, Gloria Mark, Jingyi Li, and Huahai Yang. 2019. Trusting virtual agents: The effect of personality. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 9, 2-3 (2019), 1–36.

[65] John Zimmerman and Jodi Forlizzi. 2017. Speed dating: providing a menu of possible futures. *She Ji: The Journal of Design, Economics, and Innovation* 3, 1 (2017), 30–50.

[66] Andreas Zimmermann, Marcus Specht, and Andreas Lorenz. 2005. Personalization and context management. *User modeling and user-adapted interaction* 15, 3-4 (2005), 275–302.

Publication 8

May I Interrupt? Diverging Opinions on Proactive Smart Speakers

Leon Reicherts, Nima Zargham, Michael Bonfert, Yvonne Rogers, and Rainer Malaka

In Extended Abstracts of the 3rd Conference on Conversational User Interfaces (CUI '21). New York, NY, USA, 2022. Association for Computing Machinery.

Personal contribution to this work: Conceptualization, data curation, part of formal analysis, investigation, methodology, project administration, resources, validation, visualization, and contribution to all parts of the manuscript.

ISBN: 978-1-4503-8998-3/21/07 DOI: 10.1145/3469595.3469629

May I Interrupt?
Diverging Opinions on Proactive Smart Speakers

Leon Reicherts*
l.reicherts.17@ucl.ac.uk
University College London
United Kingdom

Nima Zargham*
zargham@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

Michael Bonfert*
bonfert@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

Yvonne Rogers
y.rogers@ucl.ac.uk
University College London
United Kingdom

Rainer Malaka
malaka@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

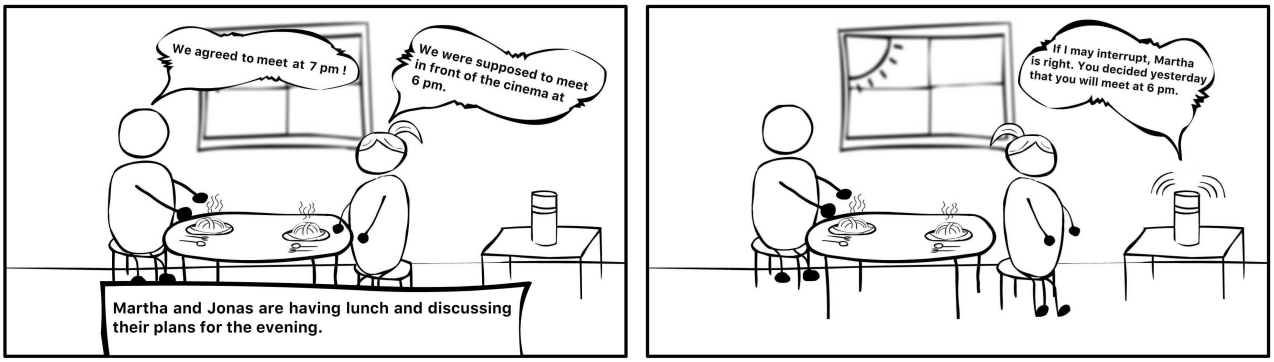


Figure 1: One of the storyboards used in the online survey presenting a scenario in which the voice assistant is proactively engaging in a conversation between two people to resolve their disagreement.

ABSTRACT

Although smart speakers support increasingly complex multi-turn dialogues, they still play a mostly reactive role, responding to user's questions or requests. With rapid technological advances, they are becoming more capable of initiating conversations by themselves. However, before developing such proactive features, it is important to understand how people perceive different types of agent-initiated interactions. We conducted an online survey in which participants (N = 47) rated 8 scenarios around proactive smart speakers on different aspects. Despite some controversy around proactive systems, we found that participants' ratings were surprisingly positive. However, they also commented on potential issues around user privacy and agency as well as undesirable interference with ongoing (social) activities. We discuss these findings and their implications for future avenues of research on proactive smart speakers.

CCS CONCEPTS

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

KEYWORDS

Proactive Agents, Voice Assistants, Conversational Agents, Smart Speakers, Smart Home

ACM Reference Format:
Leon Reicherts, Nima Zargham, Michael Bonfert, Yvonne Rogers, and Rainer Malaka. 2021. May I Interrupt? Diverging Opinions on Proactive Smart Speakers. In *3rd Conference on Conversational User Interfaces (CUI '21), July 27–29, 2021, Bilbao (online), Spain.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3469595.3469629

*The three authors contributed equally to this research

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

ACM ISBN 978-1-4503-8998-3/21/07…$15.00
https://doi.org/10.1145/3469595.3469629

1 INTRODUCTION

Smart speakers have become a mainstream technology in the home, commonly used for tasks such as searching for information, controlling internet of things devices, setting reminders, or asking for the weather [3]. Despite a large variety of use cases and increasingly sophisticated conversational abilities, smart speakers and the voice assistants (VAs) which they incorporate still follow a mostly reactive model where the user initiates the interaction and the VA responds. With rapid progress in sensing techniques and artificial intelligence, VAs become increasingly capable of understanding their surroundings and users' preferences, activities, and intentions which will enable them to become more proactive.

Proactive assistants have been proposed for specific situations, environments, and tasks [11, 14, 15], and some commercial assistants already support limited forms of proactivity [6], yet there is still a need to better understand people's views on such interactions. People may find the idea of a proactive VA too intrusive especially if it interferes in a conversation or the interruption is not helpful at that time. Others may welcome the interjections if it provides what they need to know at that moment. Therefore, proactive VAs need to strike the right balance between being helpful and being intrusive. Thus, the aim of this study is to investigate opinions on a range of everyday domestic situations, where a VA proactively addresses the user(s) in different ways based on their ongoing activities and conversations. An example is shown in Figure 1.

For the scope of this work, we consider proactive behaviour from VAs as agent-initiated interactions triggered by contextual and environmental events or user behaviours, opposed to user-initiated requests or pre-configured actions, such as reminders, alerts, or routines set by the user. Through an online survey we sought to answer the question: Which of the proposed interactions are considered most *useful*, *pleasant*, *appropriate*, and *overall positive*? In the survey, participants had to rate eight different scenarios on these dimensions and describe what they like and dislike about it. We found that most participants felt surprisingly positive about the proactive behaviour, although several people were generally skeptical. Various concerns were raised regarding privacy, timing of interventions, and appropriateness in certain contexts.

2 RELATED WORK

There has been extensive research around proactive services in various technologies and devices, for example for context-aware reminders or recommendations [19], for mental well-being [13], health [4], or in elder care [18]. Proactive or system-initiated interactions have been extensively studied over several decades in spoken dialogue systems (e. g. [9, 20]).

While proactive services can provide useful information for assisting, inspiring, and engaging users, the timing and relevance of interventions is critical to the user experience [2] and can often be challenging to achieve [14]. Proactive VAs and opportune moments for them to intervene have been studied in domestic settings [10], in vehicles [5, 15–17], as well as for performing manual do-it-yourself tasks [11], among others. The importance of timing and appropriateness of proactive interventions is even more pronounced for voice user interfaces (VUIs). Attending to GUI-based notifications can more easily be delayed until the user is ready, which is not possible with VUIs as speech demands immediate attention and can thus interfere with ongoing user activities or social interactions.

To examine opportune moments to intervene in domestic settings, Cha et al. [8] used a voice-based experience sampling method. In their study they found that the key determinants for opportune moments are closely related to personal contextual factors, including busyness or mood, as well as other factors associated with the everyday routines at home, such as social context, i. e., presence of other people, or user's movement.

Miksik et al. [14] describe a framework they developed for their proactive VA to determine opportune points to interrupt. Their system uses microphones and cameras to understand its context, e. g., presence and activity of people, using *Spatial AI*. In their user study, the VA provided simple day-to-day information which was generally perceived to be useful by participants. The authors note that more complex and more "social" proactive interventions would be the next development step, where the VA takes on a more human-like role considering the user's personality, current mood, and cultural and social context.

To create an understanding of how people may want to interact with prospective VAs, Völkl et al. [22] conducted an elicitation study through an online survey in which participants were presented with everyday scenarios. For each scenario they had to write down an imagined perfect dialog between the user and a VA. The VAs in participants' imagined dialogues were often proactive, anticipating possible next actions, and suggesting things without being requested by the user. 8.3 % of dialogues were even initiated by the VA and not by the user, which suggests that people may want future VAs to be more proactive in certain situations. However, the authors point out that for some of these imagined dialogues – including the proactive ones – participants assumed that the assistant would have substantial knowledge about both the user and the environment, which may lead to concerns around data collection and privacy.

3 STUDY DESIGN

The purpose of this study was to understand people's perceptions of proactive behaviour in different situations. Our approach was inspired by vignette experiments [1] and scenario-based design methods [7], which can be used to investigate (future) technologies despite current technical limitations. Participants are presented with a hypothetical scenario, which they are asked to reflect on and evaluate. Since the context and spatial configurations of smart speakers and users are relevant for each scenario, we used graphical storyboards to more effectively convey this information. Two exemplary scenarios are shown in Figure 2.

3.1 Online Survey

Through an online survey, for which ethics approval was obtained from University College London, we collected the opinions of 47 anonymised participants. After a welcome text and a short introduction, they gave informed consent. We then introduced the concept of a proactive VA and our fictional agent, whom we gave the gender-ambiguous name 'Jay' to reduce gender bias. We asked about the participants' typical usage of VAs and if they own a smart speaker. We then presented the eight scenarios one by one in randomised

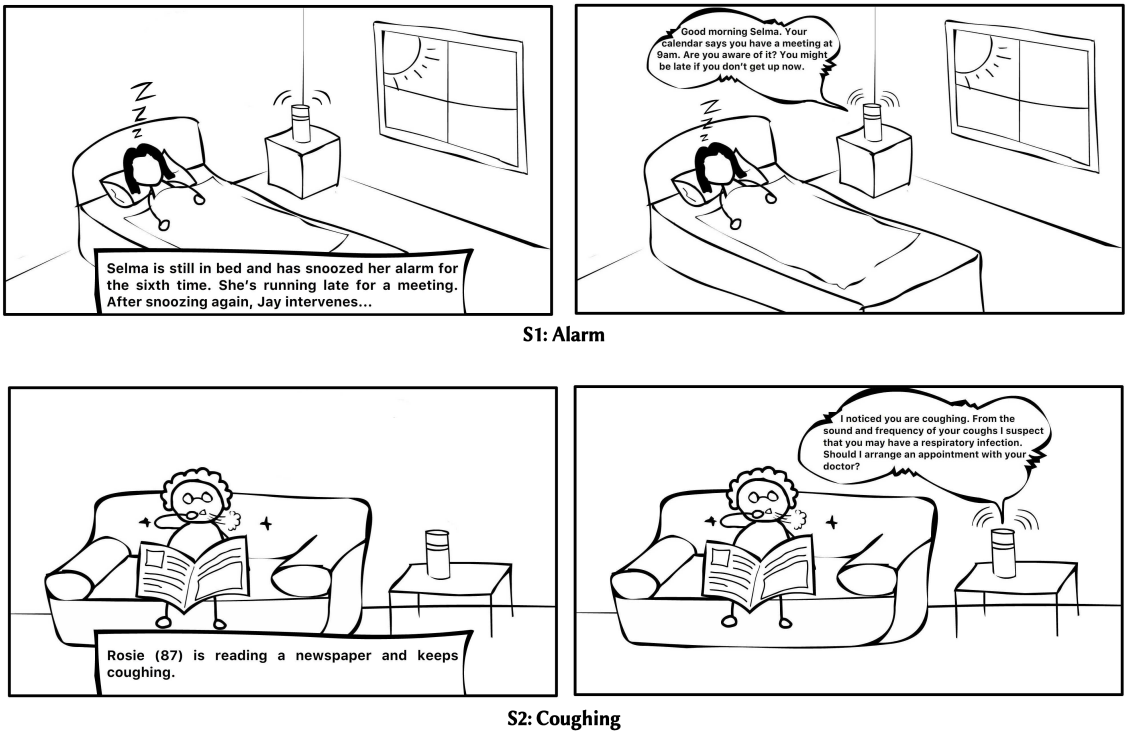


Figure 2: Two examples of the storyboards used in the survey.

order and asked participants to rate Jay's proactive interactions in terms of *usefulness*, *appropriateness*, *pleasantness*, and how positive or negative their *overall impression* is, using a five-point Likert scale for every scenario individually. Participants were then asked to share what they liked or disliked regarding Jay's proactive behaviour in open-ended questions. The survey concluded with a set of questions on demographics. Since current smart speakers are used by a wide range of users of different age groups, we did not have any inclusion criteria apart from being fluent in English.

3.2 Storyboards

Multiple brainstorming sessions were held with a group of three researchers in which 30 scenarios were conceived. The creation of the scenarios was based on what we imagined may be useful proactive interventions in everyday situations, which we were familiar with ourselves, or situations we knew about from other people. The scenarios were all situated in a home environment including a single person or multiple people. We classified the scenarios according to the interruption of a conversation, the number of people present, whether the action was imposed on the user or rather suggestive, and the potential to be perceived positively or negatively by the user(s). After several iterations, eight scenarios were selected covering the different categories – including one deliberately misplaced initiation of interaction – for which we then created graphical storyboards. We ran a pilot study on the final set of the storyboards with 3 participants to see if the scenarios successfully immersed participants and inspired them to contemplate.

All scenarios were presented as sketches in a comic style with two or three panels. Several different styles were explored with the aim to convey the situation without any ethnic or cultural cues so that all participants should be able to put themselves in the shoes of the characters. To avoid an influence from the reactions of the depicted characters on the participants' opinion, no facial expressions or responses to Jay's behaviour were included. The complete set of storyboards is in the appendix and briefly described in the following.

- *S1 Alarm*: After the user has repeatedly snoozed the alarm, Jay reminds her of an upcoming meeting.
- *S2 Coughing*: From the sound of the cough, Jay suspects an elderly user to have a respiratory infection and offers arranging a doctor's appointment.
- *S3 Tyre Change*: Based on past events in the calendar, Jay proposes to arrange an appointment at the car workshop.
- *S4 Historical Fact*: Three friends discuss a historical topic when Jay interrupts them to get a fact right.
- *S5 Time Clarification*: Two people remember differently what they agreed on, when Jay settles the disagreement by quoting what they said.
- *S6 Binge Watching*: When the user asks Jay to play a TV series, Jay suggests to stop earlier than last night.
- *S7 Headphones Setup*: A user asks a friend for help in setting up new headphones. As the friend is busy, Jay offers to assist.
- *S8 Quiz Spoiler*: During quiz night, Jay reveals the correct solution before the players had a chance to answer.

3.3 Participants

A quota sampling approach was used to recruit participants. The acquisition was based on mailing lists, social networks, and word-of-mouth. Participation was voluntary and uncompensated. Of the 47 participants 25 self-identified as female, 18 as male, 1 as non-binary, and 3 preferred not to say. The majority of our participants (72.3 %) ranged from 18 to 34 years of age, 14.9 % ranged from 35 to 54, and 12.8 % were older than that. 55.3 % of our participants have previously used VAs (10 rarely, 16 often). 25.5 % of participants owned a smart speaker.

4 RESULTS

The following findings give an impression of the participants' diverse opinions on the proactive abilities of Jay and are divided into first quantitative and then qualitative results. Due to the exploratory nature of this research we refrained from inference testing. Our aim was to identify trends as possible avenues for future research.

4.1 Heterogeneous Scenario Ratings

The participants rated the scenarios on average higher than we expected. Especially the *usefulness* of the interactions received high ratings with a mean of $M = 3.73$ out of 5 across all scenarios (including the misplaced initiation in *S8 Quiz Spoiler*) compared to how *pleasant* ($M = 2.95$), *appropriate* ($M = 2.94$), and *positive* ($M = 3.07$) the participants found the scenarios. The most popular interaction was *S1 Alarm* with an overall impression of $M = 3.89$. Similarly positive was the impression of the scenarios *S7 Headphones Setup* ($M = 3.77$), *S3 Tyre Change* ($M = 3.62$), and *S2 Coughing* ($M = 3.51$). The least popular interactions were *S8 Quiz Spoiler* with $M = 1.79$ and *S5 Time Clarification* with $M = 2.53$. The distribution of the Overall Impression ratings are shown in Figure 3.

The participants expressed widely varying opinions in the questionnaire regarding all interactions. Every scenario received the highest and the lowest possible ratings on all tested dimensions by at least one participant. The only exceptions are *S2 Coughing* with a minimum rating of 2 for *pleasantness* and overall impression, and *S8 Quiz Spoiler* with a maximum overall impression of 4. With the exception of *S8*, the standard deviation for all ratings of the overall impression was larger than 1.0 which indicates a notable variance considering the five-point scale. The scenario with the largest disagreement among the sample was *S6 Binge Watching* with a standard deviation of $SD = 1.40$.

We could not identify a relationship between the wide spread in attitudes and the demographic data. There were no differences depending on gender or age. Likewise, we did not find differences depending on the participants' usage of VAs, or whether they own a smart speaker.

4.2 Predictors for Positive Overall Impression

In contrast to the interpersonal dissent, the uniformity of the dimensions *pleasantness*, *appropriateness*, and overall impression was strikingly high. The data distribution for the single scenarios are noticeably similar for these items. This consistency in the data is also evident in the strong correlation between the dimensions. Meaningful indicators for a *positive overall impression* seem to be *appropriateness* with a Pearson's coefficient of $r = .925$ and *pleasantness* with $r = .817$. *Usefulness* appears to be a less decisive factor for predicting the overall assessment of proactive VA behaviour ($r = .517$, all correlations one-sided with $p < .001$).

The classifications of whether there was an interruption of a conversation, a single person or multiple people, and whether the action was imposed on the user or rather suggestive seemed to have an influence on participants' ratings. This is depicted exemplary for *appropriateness* in Figure 4 but it applies similarly to how *useful*, *pleasant*, or *positive* the interaction was perceived. The scenarios in which Jay addressed the user in reaction to an ongoing conversation were rated worse than when the user was not engaged in a conversation. Similarly, the interactions in which the user was alone when being addressed by Jay received better ratings than when being with others. Further, the scenarios in which Jay framed the assistance as a suggestion, instead of imposing the help onto the user, were judged better by the participants.

4.3 Participants' Reflections on Proactivity

To evaluate the answers to the open-ended questions, three researchers agreed on a coding system that was generated from a random selection of ten participants' responses. Subsequently, all responses were coded along this categorisation and summarised.

Overall, participants found the proactive behaviour of Jay helpful. The most favoured aspect of Jay were the proactive reminders. 20 people mentioned that they would benefit from such a feature. On the other hand, one person had concerns if this would become a habit: "I think it will make me lazy and will have a bad effect on my memory overall". Four participants pointed out that the ability to provide personalised suggestions is an important factor to enhance the usability of the system. One participant mentioned "Jay can definitely improve certain aspects of life, but it has to be well calibrated and personalised so it only assists when you really want it to." At the same time, the personalisation aspect generally raised many privacy concerns. Although we explained in the survey introduction that Jay would protect the users' personal data by processing it locally on the device, twelve participants still raised doubts regarding the privacy protection by a proactive VA. One user said: "Only proactive behaviours that do not require constant listening are acceptable". Another user even recommended that such systems should proactively provide suggestions regarding privacy: "Jay appears to always be listening, but does Jay ever say 'Please turn off the microphone when you don't want me to hear what you are saying'?"

Proactive instructions, where Jay guides users through a task with a sequence of steps, was a feature that was favoured by eleven participants. 15 participants pointed out that the timing for initiating a proactive action is crucial. One participant mentioned: "When Jay is proactive, it should basically behave like a person. Jumping in every discussion or argument is going to be annoying." Another one said: "I like the idea of Jay asking if it should suggest something later." Four people stated that Jay's proactive behaviour is fine only when being alone. When more people are present, they would not like to be interrupted by the VA: "If I am in the middle of an interaction with one or more persons, I do not want Jay to interrupt." One person raised concerns about proactive behaviour

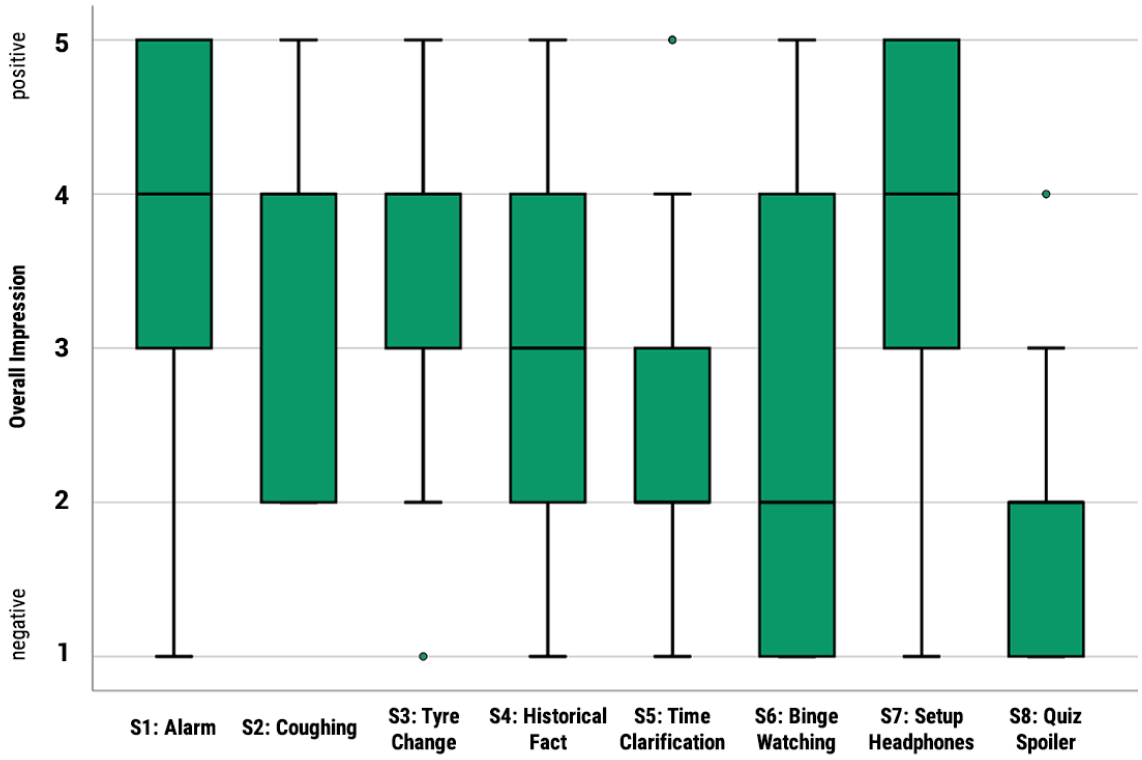


Figure 3: Box plot of *Overall Impression* ratings for all scenarios on a Likert scale from 1 to 5.

of Jay in front of children: "If [the assistant is] proactively speaking, you will always be worried that Jay says something unpleasant in front of a 5-year-old child."

Five participants were skeptical about the social sensitivity of a proactive smart speaker. They raised concerns about an AI's understanding of the conversational context which sometimes can be even difficult for humans. One user mentioned: "It would be great if Jay could learn some basic good manners and develop a certain level of social sensitivity by interacting with humans like children do. I could easily imagine a young kid interrupting a social interaction and being told off by his parents." Seven participants pointed out that certain proactive behaviours could damage human-human interaction. A participant speculated: "If the relationships in the household are suffering from lack of time spent together, it may exacerbate the circumstances by taking time away from the families." Six participants raised concerns regarding their agency. They found some proactive behaviours of Jay intrusive and did not like that the assistant takes control of certain aspects of their lives. For instance, one user said: "I am already annoyed by my phone [automatically] turning down the volume on my headphones, because it feels intrusive."

5 DISCUSSION

Our survey results suggest that many people think rather positively about proactive behaviours and consider them useful. However, there were various concerns about privacy, timing of interventions, and appropriateness in certain contexts, which resonates with previous studies [2, 8, 12].

The quantitative analysis revealed interesting tendencies for the different types of scenarios. Interactions where users were alone with the VA were generally rated more positive than with other people present, which corresponds with various comments by participants. Somewhat unsurprisingly, quantitative and qualitative findings suggest that reminders were the most favoured type of intervention, which may be partly due to people already being familiar with various types of reminders from existing devices and services they use. Other well-received types of behaviour were proactive instructions on a task the user is performing or providing health-related suggestions. When the VA interfered in personal conversations and provided evidence from previous conversations or knowledge graphs, participants perceived it as less appropriate. However, with a set of only eight scenarios and an exploratory study design, it is too early for generalising comparisons between the different types of interactions. Future research should verify these conjectures systematically and include further use cases, (social) contexts, and ways the VA initiates interactions. Based on that, the classifications of these situations and VA behaviours could also be further refined and extended towards a taxonomy of VA proactivity types regarding content and form of interventions.

Several privacy concerns were raised, since our proposed VA would need to continuously analyse its environment and users' activities. This concern has already been raised for existing smart

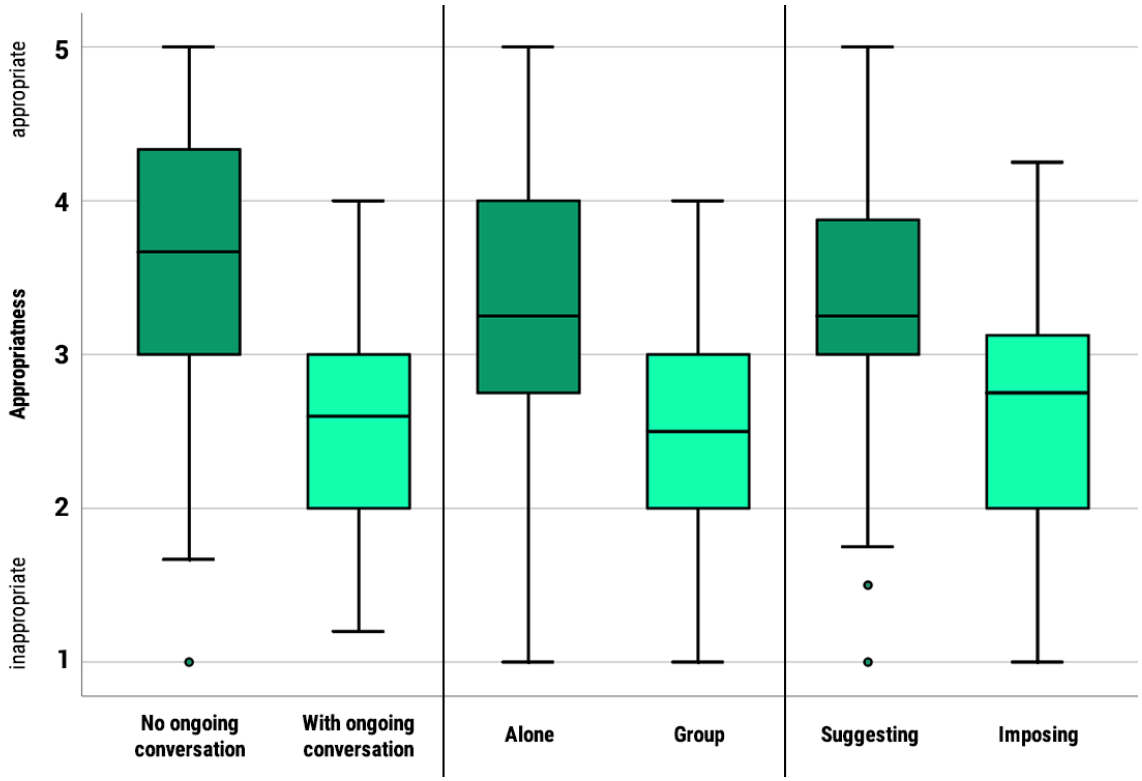


Figure 4: Box plot of *Appropriateness* ratings comparing the three scenario classifications on a Likert scale from 1 to 5.

speakers [12, 21] but will be even more pronounced for proactive ones, due to the data collection that is required to determine opportune moments for VA interactions. A participant claimed that proactive behaviour would only be acceptable if it does not require constant listening. One idea was that the speaker could remind users about how they can configure it or temporarily turn it off.

Much critique concerned the social awareness of the VA and questioned sufficient understanding of context and intentions, e. g., that not all questions are meant to be answered. Social skills such as when to speak or when to approach others are complex abilities that are difficult for computer systems to master. A possible approach that was suggested, which could reduce inappropriateness in social situations, is that the assistant would ask more gently if it should suggest or remind about something, e. g., "Would you like me to help you with that?" or "May I suggest something concerning ... ?".

The strong correlations between the rating dimensions were expected. It is interesting to see though that the overall impression of the scenarios correlated more strongly with the aspects *appropriateness* and *pleasantness* than with *usefulness*. Future research could examine these relationships further to confirm the tendencies found here and suggest social or situational appropriateness as a primary design guideline for proactive VAs.

6 CONCLUSION

Our scenario-based study, in which participants were shown a series of storyboards in which smart speakers proactively addressed users in everyday situations, was successful in eliciting a broad range of reactions. In particular, it enabled participants to reflect on the *usefulness*, *pleasantness*, and *appropriateness* of VA-initiated interactions. Our findings show that people generally found them useful but many raised concerns around timing of interventions, privacy protection, and loss of control. This further resonates with our finding that a positive perception of a proactive VA behaviour seemed to be less related to its perceived usefulness and more to its appropriateness. Furthermore, the diverging opinions suggest that proactive smart speakers may be desirable only in certain situations and for some users. The study findings underline that although future smart speakers will most likely involve a combination of reactive and proactive interactions, people will need to keep a certain level of agency over when they allow the VA to observe the environment and to be proactive.

ACKNOWLEDGMENTS

This work was partially funded by the Leverhulme Trust as part of the Doctoral Training Programme for the Ecological Study of the Brain (DS-2017-026) and the Klaus Tschira Foundation.

REFERENCES

[1] Herman Aguinis and Kyle J. Bradley. 2014. Best Practice Recommendations for Designing and Implementing Experimental Vignette Methodology Studies. *Organizational Research Methods* 17, 4 (2014), 351–371. https://doi.org/10.1177/1094428114547952

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[3] 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

[4] Daisuke Asai, Jarrod Orszulak, Richard Myrick, Chaiwoo Lee, Joseph F Coughlin, and Olivier L De Weck. 2011. Context-aware reminder system to support medication compliance. In *2011 IEEE international conference on systems, man, and cybernetics*. IEEE, New York, USA, 3213–3218.

[5] André Berton, Dirk Bühler, and Wolfgang Minker. 2006. SmartKom-Mobile Car: User Interaction with Mobile Services in a Car Environment. In *SmartKom: Foundations of Multimodal Dialogue Systems*, Wolfgang Wahlster (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 523–537. https://doi.org/10.1007/3-540-36678-4_33

[6] Ian Carlos Campbell. 2021. Amazon's Alexa can now act on its own hunches to turn off lights and more. https://www.theverge.com/2021/1/25/22249044/amazon-alexa-update-proactive-hunches-guard-plus-subscription

[7] John M. Carroll. 1999. Five Reasons for Scenario-Based Design. In *Proceedings of the Thirty-Second Annual Hawaii International Conference on System Sciences-Volume 3 - Volume 3 (HICSS '99)*. IEEE Computer Society, USA, 3051.

[8] Narae Cha, Auk Kim, Cheul Young Park, Soowon Kang, Mingyu Park, Jae-Gil Lee, Sangsu Lee, and Uichin Lee. 2020. Hello There! Is Now a Good Time to Talk? Opportune Moments for Proactive Interactions with Smart Speakers. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 74 (Sept. 2020), 28 pages. https://doi.org/10.1145/3411810

[9] Kristiina Jokinen and Michael McTear. 2009. Spoken dialogue systems. *Synthesis Lectures on Human Language Technologies* 2, 1 (2009), 1–151.

[10] Mitsuki Komori, Yuichiro Fujimoto, Jianfeng Xu, Kazuyuki Tasaka, Hiromasa Yanagihara, and Kinya Fujita. 2019. Experimental Study on Estimation of Opportune Moments for Proactive Voice Information Service Based on Activity Transition for People Living Alone. In *Human-Computer Interaction. Perspectives on Design*, Masaaki Kurosu (Ed.). Springer International Publishing, Cham, 527–539.

[11] Matthias Kraus, Marvin Schiller, Gregor Behnke, Pascal Bercher, Michael Dorna, Michael Dambier, Birte Glimm, Susanne Biundo, and Wolfgang Minker. 2020. "Was That Successful?" On Integrating Proactive Meta-Dialogue in a DIY-Assistant Using Multimodal Cues. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20)*. Association for Computing Machinery, New York, NY, USA, 585–594. https://doi.org/10.1145/3382507.3418818

[12] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, Are You Listening? Privacy Perceptions, Concerns and Privacy-Seeking Behaviors with Smart Speakers. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 102 (Nov. 2018), 31 pages. https://doi.org/10.1145/3274371

[13] Uichin Lee, Kyungsik Han, Hyunsung Cho, Kyong Mee Chung, Hwajung Hong, Sung Ju Lee, Youngtae Noh, Sooyoung Park, and John M. Carroll. 2019. Intelligent positive computing with mobile, wearable, and IoT devices: Literature review and research directions. *Ad Hoc Networks* 83 (Feb. 2019), 8–24. https://doi.org/10.1016/j.adhoc.2018.08.021

[14] O. Miksik, I. Munasinghe, J. Asensio-Cubero, S. Reddy Bethi, S-T. Huang, S. Zylfo, X. Liu, T. Nica, A. Mitrocsak, S. Mezza, R. Beard, R. Shi, R. Ng, P. Mediano, Z. Fountas, S-H. Lee, J. Medvesek, H. Zhuang, Y. Rogers, and P. Swietojanski. 2020. Building Proactive Voice Assistants: When and How (not) to Interact. arXiv:2005.01322 [cs.HC]

[15] Maria Schmidt and Patricia Braunger. 2018. A Survey on Different Means of Personalized Dialog Output for an Adaptive Personal Assistant. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization* (Singapore, Singapore) *(UMAP '18)*. Association for Computing Machinery, New York, NY, USA, 75–81. https://doi.org/10.1145/3213586.3226198

[16] Maria Schmidt, Wolfgang Minker, and Steffen Werner. 2020. User Acceptance of Proactive Voice Assistant Behavior. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2020*, Andreas Wendemuth, Ronald Böck, and Ingo Siegert (Eds.). TUDpress, Dresden, Germany, 18–25.

[17] Maria Schmidt, Daniela Stier, Steffen Werner, and Wolfgang Minker. 2019. Exploration and assessment of proactive use cases for an in-car voice assistant. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, Peter Birkholz and Simon Stone (Eds.). TUDpress, Dresden, Germany, 148–155.

[18] Candace L. Sidner, Timothy Bickmore, Bahador Nooraie, Charles Rich, Lazlo Ring, Mahni Shayganfar, and Laura Vardoulakis. 2018. Creating New Technologies for Companionable Agents to Support Isolated Older Adults. *ACM Trans. Interact. Intell. Syst.* 8, 3, Article 17 (July 2018), 27 pages. https://doi.org/10.1145/3213050

[19] Timothy Sohn, Kevin A. Li, Gunny Lee, Ian Smith, James Scott, and William G. Griswold. 2005. Place-Its: A Study of Location-Based Reminders on Mobile Phones. In *Proceedings of the 7th International Conference on Ubiquitous Computing* (Tokyo, Japan) *(UbiComp'05)*. Springer-Verlag, Berlin, Heidelberg, 232–250. https://doi.org/10.1007/11551201_14

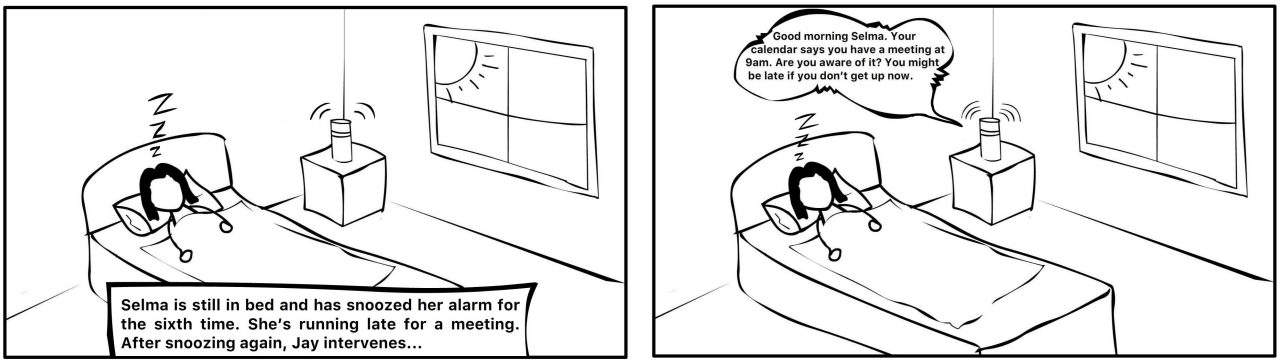
[20] Petra-Maria Strauss and Wolfgang Minker. 2010. *Proactive spoken dialogue interaction in multi-party environments*. Springer, Heidelberg, Germany.

[21] Madiha Tabassum, Tomasz Kosiński, and Heather Richter Lipford. 2019. "I Don't Own the Data": End User Perceptions of Smart Home Device Data Practices and Risks. In *Proceedings of the Fifteenth USENIX Conference on Usable Privacy and Security* (Santa Clara, CA, USA) *(SOUPS'19)*. USENIX Association, USA, 435–450.

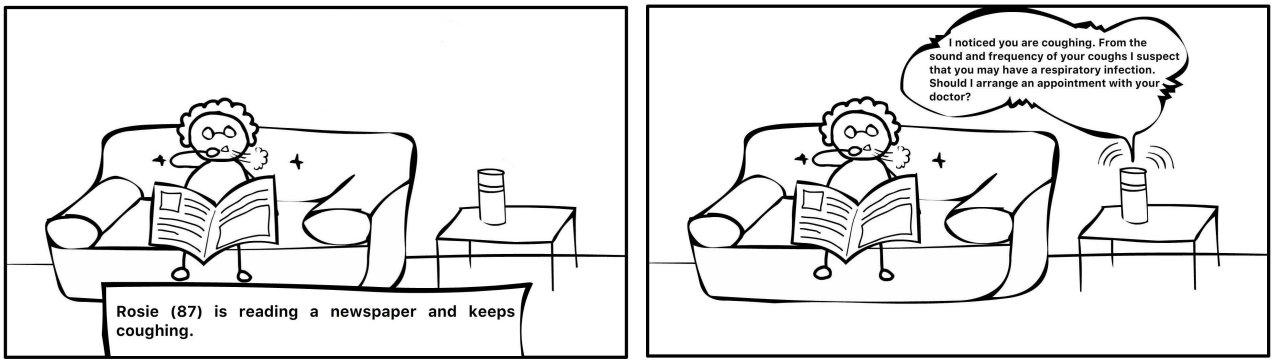
[22] Sarah Theres Völkel, Daniel Buschek, Malin Eiband, Benjamin R. Cowan, and Heinrich Hussmann. 2021. Eliciting and Analysing Users' Envisioned Dialogues with Perfect Voice Assistants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association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

A APPENDIX

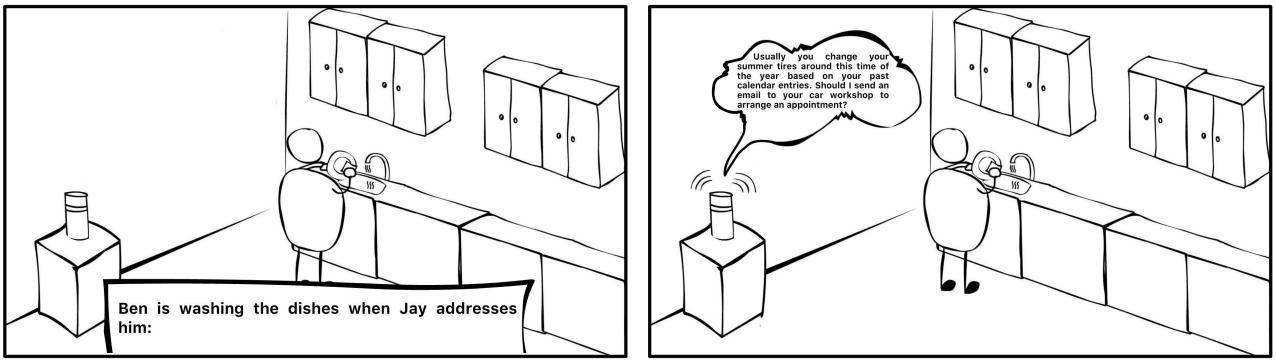
Scenario 1: Alarm



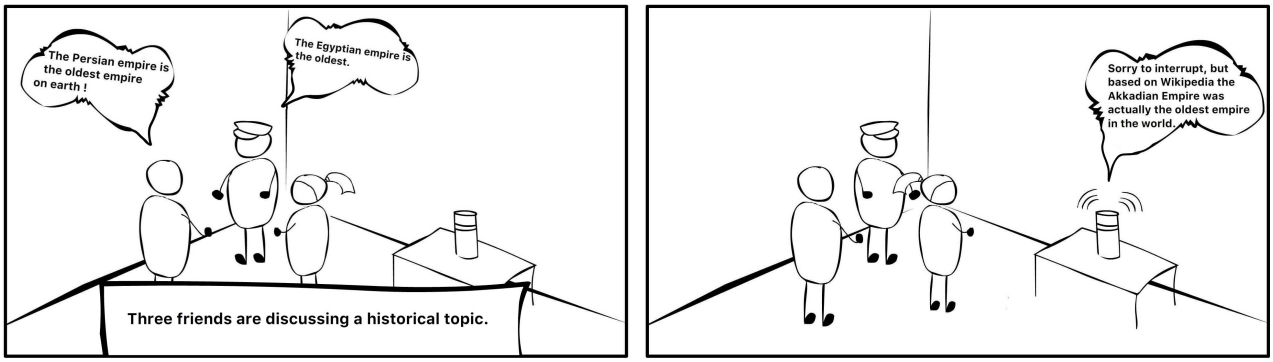
Scenario 2: Coughing



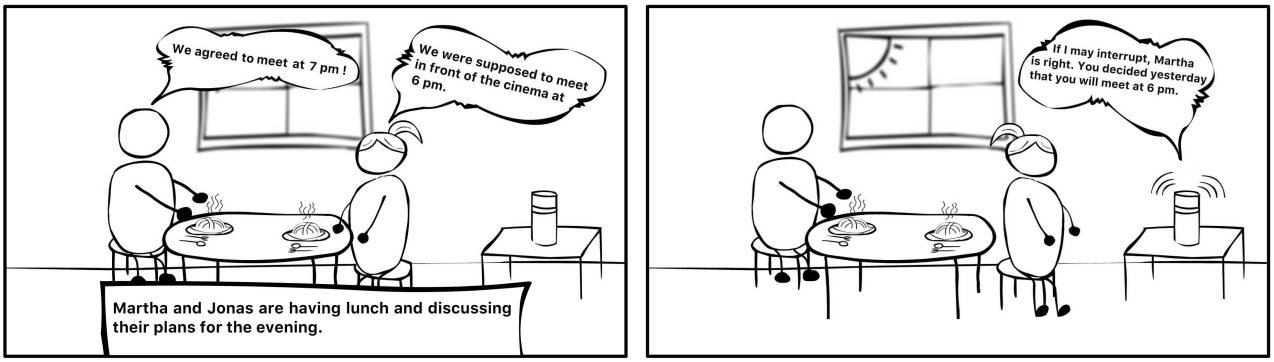
Scenario 3: Tyre Change



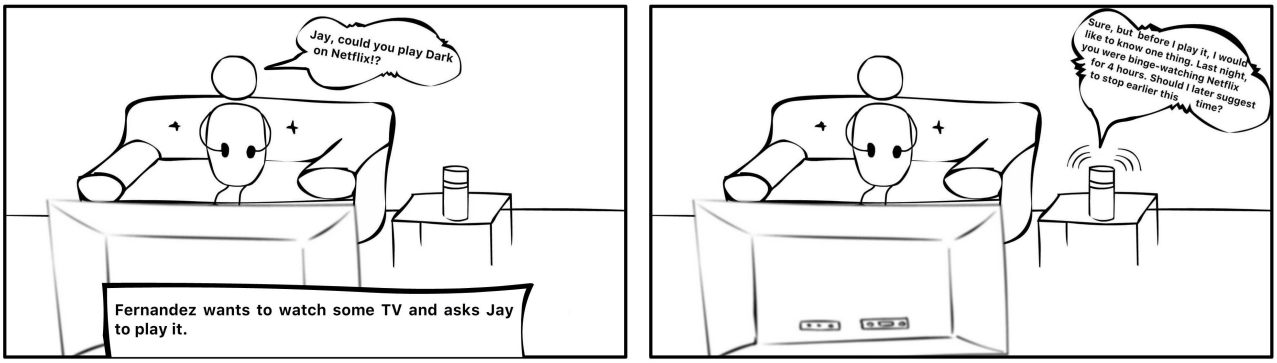
Scenario 4: Historical Fact



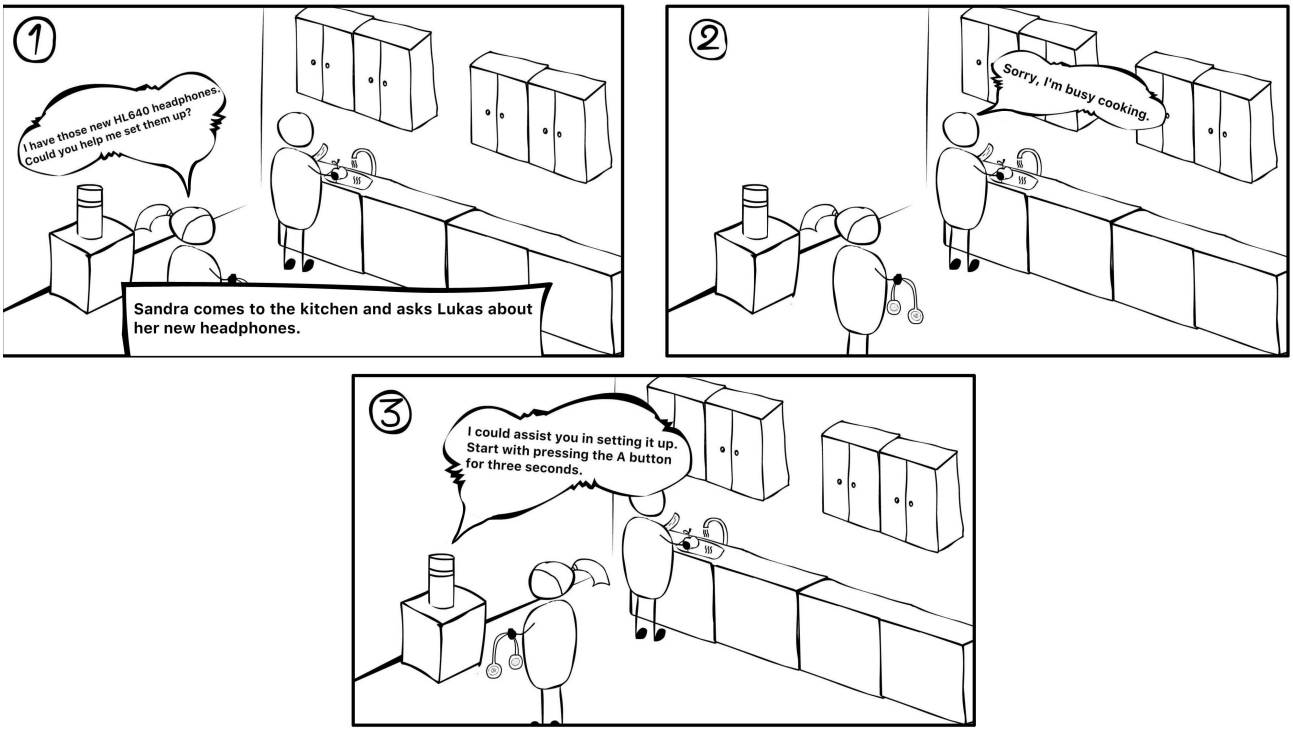
Scenario 5: Time Clarification



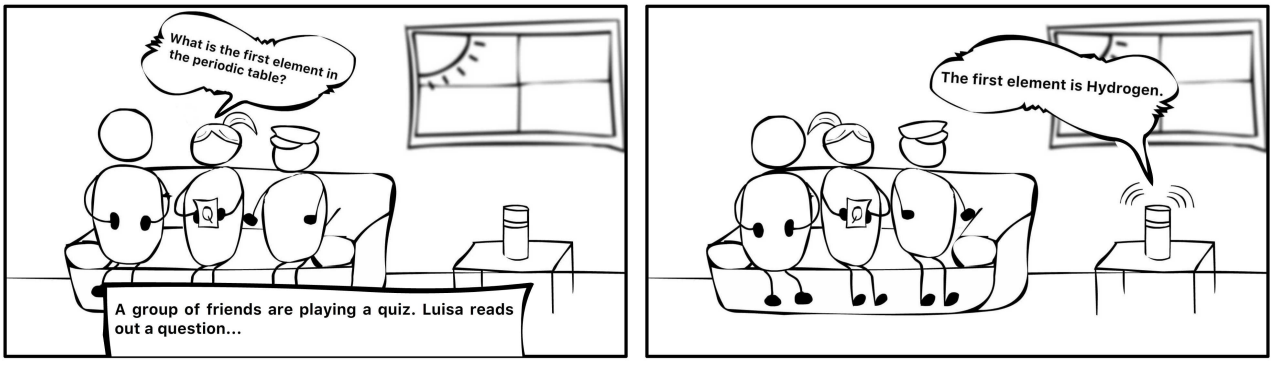
Scenario 6: Binge Watching



Scenario 7: Headphones Setup



Scenario 8: Quiz Spoiler



Publication 9

Understanding Circumstances for Desirable Proactive Behaviour of Voice Assistants: The Proactivity Dilemma

Nima Zargham, Leon Reicherts, Michael Bonfert, Sarah Theres Völkel, Yvonne Rogers, Johannes Schöning, and Rainer Malaka

In Proceedings of the 4th Conference on Conversational User Interfaces (CUI '22). New York, NY, USA, 2022. Association for Computing Machinery.

Personal contribution to this work: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, validation, visualization, and contribution to all parts of the manuscript.

ISBN: 978-1-4503-9739-1/22/07 DOI: 10.1145/3543829.3543834

Understanding Circumstances for Desirable Proactive Behaviour of Voice Assistants: The Proactivity Dilemma

Nima Zargham
zargham@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

Leon Reicherts
l.reicherts.17@ucl.ac.uk
University College London
United Kingdom

Michael Bonfert
bonfert@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

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

Johannes Schöning
johannes.schoening@unisg.ch
University of St. Gallen
Switzerland

Rainer Malaka
malaka@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

Yvonne Rogers
y.rogers@ucl.ac.uk
University College London
United Kingdom

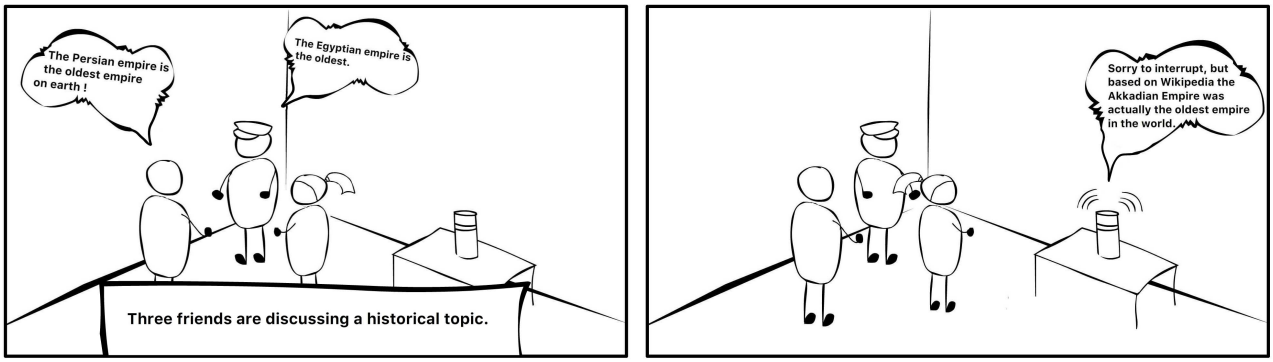


Figure 1: An example storyboard used in the study. In this scenario (S4), the agent proactively approaches users based on their conversation.

ABSTRACT

The next major evolutionary stage for voice assistants will be their capability to initiate interactions by themselves. However, to design proactive interactions, it is crucial to understand whether and when this behaviour is considered useful and how desirable it is perceived for different social contexts or ongoing activities. To investigate people's perspectives on proactivity and appropriate circumstances for it, we designed a set of storyboards depicting a variety of proactive actions in everyday situations and social settings and presented them to 15 participants in interactive interviews. Our findings suggest that, although many participants see benefits in agent proactivity, such as for urgent or critical issues, there are concerns about interference with social activities in multi-party settings, potential loss of agency, and intrusiveness. We discuss our implications for designing voice assistants with desirable proactive features.

CUI 2022, July 26–28, 2022, Glasgow, United Kingdom

ACM ISBN 978-1-4503-9739-1/22/07…$15.00
https://doi.org/10.1145/3543829.3543834

CCS CONCEPTS

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

KEYWORDS

Proactive Agents, Voice Assistants, Conversational Agents, Smart Speakers, Smart Home

ACM Reference Format:
Nima Zargham, Leon Reicherts, Michael Bonfert, Sarah Theres Völkel, Johannes Schöning, Rainer Malaka, and Yvonne Rogers. 2022. Understanding Circumstances for Desirable Proactive Behaviour of Voice Assistants: The Proactivity Dilemma. In *4th Conference on Conversational User Interfaces (CUI 2022), July 26–28, 2022, Glasgow, United Kingdom.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3543829.3543834

1 INTRODUCTION

Voice assistants (VAs) are becoming more intelligent and capable of supporting increasingly complex tasks and dialogues. People use them for controlling smart home devices, information seeking, entertainment, shopping, and activity management, among others [30]. As more and more VAs are finding their way into our homes in the form of smart speakers, they play a greater role as digital everyday helpers. However, despite the broad range of use cases and increasingly advanced language understanding and "conversational abilities" of smart speakers, these devices are still interacting mainly in a reactive manner, responding to the users' inquiry. The interaction starts with users saying the wake word followed by an inquiry, and only then the agent responds to the user.

With rapid advances in artificial intelligence, natural language understanding, and sensing techniques, VAs are becoming more capable of understanding users' behaviours, preferences, intentions, and surroundings, which opens up a broad landscape of opportunities for proactive interactions. Proactive behaviours from VAs are considered as agent-initiated interactions triggered by events related to the user(s) and their environment, as opposed to user-initiated inquiries or pre-configured actions, such as reminders, alerts, or routines set by the user [31]. In several studies, researchers have begun to examine proactive behaviour of VAs and proposed to use it for specific situations and environments [20, 27, 33]. Others have looked into the timing of proactive VA interruptions [7] and how such interruptions should be designed [11]. Moreover, certain commercial assistants already support some basic proactive services, such as reminding users of their schedule[1], automating home routines[2], and supporting home safety and security[3].

Nevertheless, the current state of research still lacks a deeper understanding of how people perceive and feel about such interactions at home. As proactive VAs require to monitor and process users' behaviour constantly, privacy concerns are likely to intensify. Tabassum et al. [38] highlighted that privacy remains a key concern that limits users' willingness to use proactive VAs. However, other factors driving users' acceptance of proactive VAs beyond privacy concerns, such as the usefulness and appropriateness of situations to be interrupted, remain unexplored.

To close the gap in understanding people's perceptions of proactive VAs in a domestic setting, we present an elicitation study to investigate the desirability of agent-initiated interventions, i.e., high usefulness, high appropriateness, and low invasiveness. Therefore, we address the following research questions:

RQ1: Under which circumstances is proactive behaviour by a voice assistant perceived as desirable?
RQ2: How should proactive interventions be initiated by the voice assistant?

To answer our research questions, we designed a set of storyboards illustrating a range of possible proactive interventions in a home environment. An example storyboard is shown in Figure 1. In an online elicitation study, we conducted interactive interviews. The participants went through a series of tasks based on these storyboards using a virtual, collaborative whiteboard to evaluate and contemplate on the concept from different perspectives. Our results show that key factors for a desirable proactive intervention are the people in the room, the type of ongoing activity, the urgency of the topic, the user's current emotional state, and the agent's initiation and phrasing of the intervention. The main contribution of this work is empirical evidence for the situational desirability of proactive VA behaviour, thereby providing a deeper understanding of factors influencing user acceptance. Our findings point toward a dilemma. As such, interactions are often perceived as useful but at the same time invasive or inappropriate. Therefore, we propose an initiation process model for minimizing the intrusiveness of useful features.

2 RELATED WORK

Although current commercial smart speakers support a limited set of proactive features such as Amazon Echo displaying a specific light pattern to visualise notifications and messages, or Google Home delivering proactive reminders for upcoming meetings based on the user's calendar, such devices remain primarily reactive with users initiating interactions. However, proactive interactions in such devices can open up new opportunities for supporting, probing, or inspiring their users [42]. In this section, we summarise related work on proactivity in VAs, opportune moments for VAs to proactively engage with the user, privacy concerns, and appropriateness of proactive interventions.

2.1 Proactivity in VAs

A survey with 1,550 participants by Schmidt and Braunger [33] revealed that proactivity is a favoured feature by users. Similarly, an elicitation study by Völkel et al. [41] on users' envisioned dialogues with a perfect voice assistant showed that many participants imagined proactive voice assistant behaviour to be desirable. In particular, the envisioned dialogues pointed to agents being able to anticipate the next possible actions and to give suggestions without being requested by the users. Andolina et al. [3] designed a proactive search agent that would listen to users conversing and present information about their conversation based on entities detected in their dialogue. They found that their agent could effectively support the conversation with facts and ideas without causing much interruption to the conversation's flow.

On the other hand, there are also potential downsides of proactive voice assistants, in particular concerning privacy [38], which we discuss below. In a study about in-car assistants, Braun et al. [5] reported that people have mixed opinions on whether the voice assistant should initiate conversations and that they only accept

[1]https://www.techrepublic.com/article/google-nest-smart-speakers-a-cheat-sheet, last accessed 2022-02-28
[2]https://www.theverge.com/2021/1/25/22249044/amazon-alexa-update-proactive-hunches-guard-plus-subscription, last accessed 2022-02-28
[3]https://www.cnet.com/home/smart-home/what-amazon-alexa-will-tell-us-in-2019, last accessed 2022-02-28

proactivity if the assistant can act like an authentic, human co-driver [5]. However, they did not investigate what factors influence some users' reluctance to engage with proactive VAs.

While there seems to be a demand for proactivity, there is little knowledge about what makes a proactive voice assistant desirable [2]. Since today's conversations with voice assistants are highly constrained, task-oriented, and impersonal [8–10, 12, 22, 29], proactive interactions in such devices could open up new opportunities and potentially empower a broad range of applications [42].

2.2 Opportune Moments for Proactive VAs

Opportune moments for interaction refer to moments where the disruption of the user's current activity is at a minimum level [39]. Although it is fairly easy and natural for people to assess another person's current activity before starting to interact with them, it is a big challenge to design such behaviour for machines [15, 32, 39]. Identifying these moments for VAs to start interacting with a user is particularly challenging, as speech interaction requires immediate attention and can easily interrupt users with their current activities or social interactions [37]. To achieve proactivity, the voice assistant needs to be context-aware and detect opportune moments to initiate interactions. Previous studies have explored opportune moments for VA interventions in homes [7, 16, 19, 43], cars [17, 18, 33–36], and other settings [20].

Conducting an experiencing sampling study with smart speakers in people's homes, Cha et al. [7] found that the key determinants for opportune moments are linked to personal contextual factors such as busyness, mood, and urgency, as well as the other contextual factors related to everyday routines at home, including social context such as presence of other people, and user mobility. Similarly, a study by Nothdurft et al. [28] suggests that the most important factors to decide if proactive behaviour is desired are the importance of the intervention for the user, users' surrounding and their mental state, and the accurate placement of the interaction.

Apart from identifying opportune moments for proactivity, researchers have also examined *how* the agent should initiate a conversation. Edwards et al. [11] looked at how people interrupt another person who is engaged in a complex task, as an approach to inform the design of proactive VAs. Their results showed that the level of urgency significantly affects how long it takes for people to start interrupting. Arias et al. [4] suggest that agents should notify users before proactively engaging with them to make sure they are willing to interact at the specific moment. Moreover, users should be in charge of deciding which information they are proactively told by an agent [4].

These studies underline that a lack of contextual knowledge is detrimental to users' acceptance of proactive VAs. While previous work has pointed out influencing factors such as the urgency of the task, little is known about specific situations in which users find a proactive VA appropriate.

2.3 Privacy Concerns in Voice Assistant Use

Preservation of privacy is a key to the users' acceptance of smart speakers. Specifically, in a home environment, stressing the importance of user privacy and security is crucial. A study by Lau et al. [21] showed that many people refrain from adopting smart speakers because they have privacy concerns or distrust the companies offering smart speakers. Malkin et al. [24] surveyed smart speaker owners and found that users are not comfortable with permanently preserving user recordings, especially those that include children and guests. Moreover, users were strongly opposed to the use of their data by third parties or for advertising.

When it comes to proactive services, such concerns intensify as the agents need to be more context-aware, have access to more personal data which are usually uploaded to and processed on companies' cloud services, and act out of the user's control. Previous research has shown that privacy represents the key challenge for proactively initiated interactions [27]. Tabassum et al. [38] conducted an online survey to explore user preferences and expectations of proactive VAs and found that, even though users perceived the services as useful, they were uncomfortable with the always-listening nature of such systems. Yet, many users were willing to share even sensitive conversations to receive more personalised and contextual services. Likewise, Cha et al. [7] found that users willingly accept to compromise their privacy in exchange for a smart speaker that offers personalised care.

Including privacy-preserving features is essential when designing proactive VAs. Previous work shows that giving users the option to examine the recordings and actions taken by the systems [25] as well as transparency on the recorded data are decisive factors for the acceptance of such proactive technologies [26]. But even with full control over what private data is shared or stored, the VA's active role and interference in the private sphere in domestic situations might be experienced as inappropriate, which needs further investigation.

2.4 Appropriate Proactive Interventions

At times, certain proactive behaviour can cause discomfort and be perceived as disruptive and invasive [4]. For successful proactive interventions, not only users' current mood but also cultural and social context need to be considered – in particular if the agent takes on a more human-like role, like a personal coach [27].

Luria et al. [23] identified three thresholds of agent proactivity including reactive to user requests, proactive by providing information, or proactive by providing recommendations for a course of action, with users differing in their comfort levels with each threshold. In a study to explore socially sophisticated agents in a domestic setting, they witnessed that most participants were open to the idea of a proactive agent in a multi-user situation, but nobody wanted the agent to enforce recommendations such as preventing them from ordering unhealthy food.

In a previous study [31], we conducted an online survey in which we asked users to rate a series of hypothetical storyboards depicting situations at home where an agent proactively addresses users, based on the criteria of *usefulness*, *pleasantness*, *appropriateness*. We found that even when participants perceived the agent interventions as useful in general, the ratings for appropriateness were much lower, suggesting that appropriateness given (social) context is a crucial factor for the overall acceptability of the interactions. While the quantitative study design could reveal interesting differences in people's ratings along those criteria, the purpose of

the present study is to gain a comprehensive understanding of the reasons behind these differences through a qualitative approach.

Overall, despite the popularity of proactive features in VAs, previous literature still lacks an understanding of user perceptions of desirable proactive behaviour in domestic settings considering situational factors. In this paper, we build upon prior work by engaging users to reflect on the contexts in which they would consider proactive interventions useful and appropriate.

3 STUDY DESIGN

In this work, we sought to investigate circumstances for a desirable proactive voice assistant in a home environment. We used an approach inspired by scenario-based design methods [6] and vignette experiments [1], which allows us to investigate (future) technologies despite current technological limitations. In our online interviews, we present participants with different hypothetical scenarios, which are illustrated by graphical storyboards to better visualize the situation and spatial configuration of the specific home environment, the user(s), and the smart speaker. We developed an interactive task-based interview procedure, designed to elicit participants' reflections on the scenarios from different perspectives. Hence, in addition to asking participants for their general thoughts on the scenarios, they were asked to complete different tasks on a virtual whiteboard. This allowed us to explore in-depth deliberations around proactive features and collect richer data. Ethical approval was received for the study from University College London.

3.1 Storyboards

The initial set of scenarios about proactive VAs in domestic settings was based on the eight scenarios from our previous study [31] where we used them to collect participant ratings across different dimensions such as perceived usefulness and appropriateness in an online study. In the present study, we reuse most of these scenarios and investigate them with a qualitative approach to shed light on *why* some proactive behaviours are seen as more or less desirable in certain contexts. However, we initially added eight additional scenarios to expand the range and the diversity of scenarios, based on further classification criteria we considered relevant for covering the large spectrum of conceivable circumstances, such as varying levels of urgency, the number of people present, or the extent of interruption (e.g., of an ongoing human-human interaction). We presented the extended set of 16 scenarios and classifications to two VA experts and asked them to add further scenarios and classifications they think are missing or might complement the existing ones. Considering their feedback, we refined the scenario selection and the classification scheme and asked three HCI researchers to independently code the scenarios using our scheme. Based on the coded scenarios, we selected nine scenarios that covered all the classifications. Seven of these nine scenarios were identical or almost identical to those of our previous research [31], including a scenario which highlighted potential challenges of proactive VAs (S8).

All scenarios were presented as cartoon sketches with two panels. The storyboards were designed in a way that should minimize any cultural and ethnic cues so that participants could put themselves in the shoes of the characters. The characters were designed without any facial expressions to avoid influencing participants' interpretation of the scenarios. As in the original storyboards, the cylinder-shaped appearance of the voice assistant was similar to a conventional smart speaker. The fictional agent was given the gender-ambiguous name "Jay" to reduce gender bias. The complete set of storyboards used for this study can be found in the Appendix and is briefly described in the following list.

- *S1 Meeting Reminder*: After the user has repeatedly "snoozed" the alarm, Jay reminds her of an upcoming meeting.
- *S2 Health Risk*: From the sound of the cough, Jay suspects an elderly user to have a respiratory infection and offers to arrange a doctor's appointment.
- *S3 Cooking Inspiration*: Two friends are contemplating about dinner when Jay offers to suggest recipes based on what is in the fridge.
- *S4 Fact Checking*: Three friends discuss a historical topic when Jay interrupts them to get a fact right.
- *S5 Disagreement Clarification*: Two people remember differently what they agreed on when Jay settles the disagreement by quoting what they said.
- *S6 Nudging*: When the user asks Jay to play a TV series, Jay suggests stopping earlier than last night.
- *S7 Technical Support*: A person asks their friend for help in setting up new headphones. As the friend is busy, Jay offers to assist.
- *S8 Fact Spoiler*: During quiz night, Jay reveals the correct solution before the players had a chance to answer.
- *S9 Emergency*: Jay detects a fire in the apartment, immediately calls the fire department and warns the sleeping residents.

3.2 Participants

15 people participated in the study, of which seven self-identified as female and eight as male. They were between 22 and 35 years of age ($M = 27.86$, $SD = 4.47$). Five participants had a bachelor's degree, nine had a master's degree, and one had a PhD. Participants were recruited using convenience sampling. The participation was voluntary and uncompensated. The recruitment continued until data saturation was reached, satisfying the recommended sample sizes of theoretical saturation from the literature [13, 40]. Two-thirds of our participants have previously used VAs (four rarely, six often). Seven participants (46.6 %) owned a smart speaker. All participants were proficient in English.

3.3 Procedure

Every study session was held remotely via video calls. The participants were asked to give informed consent and fill in the demographics questionnaire prior to the session. At the beginning of each session, participants were informed about the study procedure and the concept of a proactive VA. The study tasks were performed through the virtual whiteboard tool Miro[4]. All participants had a short familiarization phase with Miro and the virtual board. During

[4]https://miro.com, last accessed June 18, 2022

the sessions, the participants would share their screens with the interviewer to be guided through the tasks.

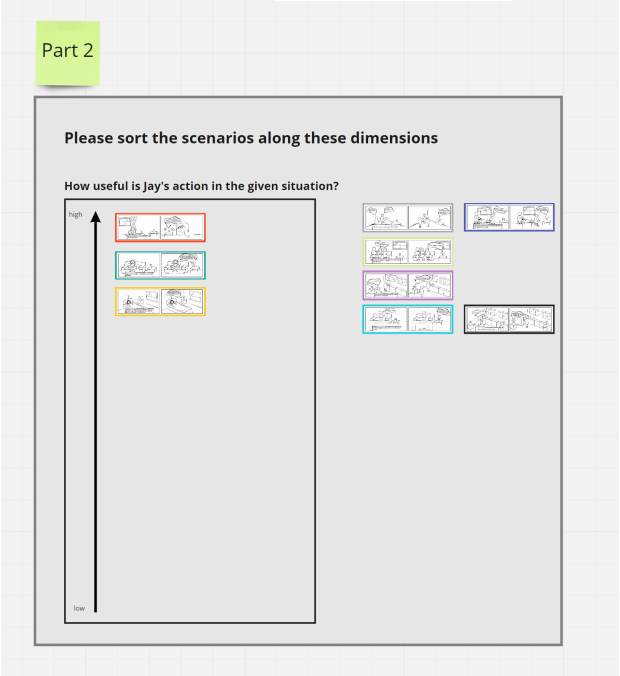


Figure 2: A partial screenshot of the card sorting task on Miro where a user is bringing the storyboards into order regarding their usefulness.

We designed a sequence of different interview parts combined with specific tasks to elicit how participants perceive the depicted (social) situation and how they think Jay's intervention affects it, as well as to understand how proactive interventions need to be designed to mitigate any negative effects on people's (social) activities. Before starting the interview, the interviewer explained that participants should assume the data is processed locally on the device. While some of Jay's features may not yet be feasible today with offline/on-device processing, we wanted to avoid participants solely worrying about data privacy, as this aspect is well researched [27, 38]. In an initial short interview, we gathered first impressions of the individual scenarios. Afterwards, participants were asked to sort the scenarios in terms of usefulness, appropriateness, and invasiveness in a card-sorting task as shown in Figure 2. After that, they speculated how each scenario may evolve and how the characters may respond to Jay's intervention. In the third task, participants were asked to choose the most invasive and the most inappropriate scenarios to then re-imagine an improved intervention. In the final task, participants were asked to decide for each scenario how they would like the VA to initiate the interaction and if it should provide a cue before starting to speak. After going through all the tasks, the session concluded with a short semi-structured interview in which participants gave their overall impression and elaborated on the potentials and challenges of proactive smart speakers. All sessions were audio-recorded for later analysis. The sessions took approximately 51.3 minutes on average ($SD = 10.6$).

3.4 Data Analysis

Our data analysis focused on two parts: content from the virtual whiteboards and spoken statements from the interviews. The information from the completed tasks in each participant's board was extracted and consolidated. The resulting data set was reviewed and discussed by three researchers. Some tasks were designed to produce categorical data, such as the card sorting and the cue selection tasks, which were examined using descriptive statistics. The interview segments were analysed for subsequent triangulation with the data from the boards. The transcripts of the interviews were independently coded by two researchers using inductive coding, where a single quote could be assigned to multiple codes. Codes were merged and consolidated by the two researchers. Three researchers discussed the codes, resolved disagreements, and derived themes which can be categorized into (I) perceived helpfulness, (II) privacy and mistrust, (III) consideration of social context, (IV) configuration and control, (V) and initiation and phrasing of interventions.

4 FINDINGS

Overall, participants had diverse opinions on proactive behaviours of a smart speaker. Some generally liked proactive interventions and valued the additional features, while others disliked them: "I would rather ask [for help] than getting help without asking" (P6). Some had mixed feelings: "It's like a double-edged sword: both helps and can intrude" (P5). In this section, we will share details about the conflicting appreciation and concerns of our participants. In each subsection, the results are presented first, followed directly by an interpretation in which we also discuss potential design implications.

4.1 Perceived Helpfulness

Participants sorted the scenarios in terms of usefulness, appropriateness, and invasiveness of the assistant's behaviour. The median rank of the scenarios in the order from 1st to 9th place is shown in Table 1. The scenario *Emergency* was considered most useful ($Medianrank = 1$), most appropriate ($Md = 1$), and least invasive ($Md = 9$) by most participants. On the other hand, *Fact Spoiler* was ranked least useful ($Md = 9$) and least appropriate ($Md = 9$). Participants ranked *Disagreement Clarification* most invasive, with 86 % sorting it within the last three ranks, and highly inappropriate ($Md = 8$). We observed considerable similarities between the outcome of the three factors. The median ranks of how useful and appropriate the scenarios were assessed strongly correlate ($r_{Spearman} = 0.911$). The usefulness is furthermore negatively correlated with the invasiveness ($r_s = -0.830$). Similarly, this strong negative correlation occurs between appropriateness and invasiveness ($r_s = -0.902$). That is, the more useful and appropriate a situation is perceived, the less invasive it is ranked in general. However, there are several exceptions regarding invasiveness that we discuss below. All correlations are statistically significant with $p < .006$ on a Bonferroni-corrected alpha level of $\alpha = .016$.

An important factor for the proactive assistants' perceived helpfulness was the amount of time its intervention could save the user.

Scenario	Useful	Appropriate	Invasive
Emergency	1	1	9
Health Risk	2	4	5
Meeting Reminder	3	3	5
Cooking Inspiration	4	3	7
Technical Support	4	4	7
Nudging	6	6	4
Fact Checking	7	7	3
Disagreement Clarification	8	8	1
Fact Spoiler	9	9	3

Table 1: Median of how useful, appropriate, and invasive the scenarios were ranked in the card sorting task (1 being the highest and 9 the lowest rank). Scenarios are sorted based on their usefulness rankings.

The time saving aspect was mentioned in particular for the *Cooking Inspiration* (four times), the *Meeting Reminder* (five times), and – unsurprisingly – by almost all participants for the *Emergency* scenario. Also, the urgency of an agent's intervention appeared to be a key determinant for how (positively) it was perceived. One person said about the *Emergency*: "As long as someone's health is in danger, privacy would not be my priority" (P6). Similarly, regarding the *Health Risk*, 12 participants found the agent's intervention helpful as it is beneficial for the user's health: "I wouldn't mind [Jay] intruding in such cases. It's more important than me not wanting to be interrupted" (P6). For most participants, agent-initiated interactions that are time-critical but without dangerous consequences were still perceived as appropriate. Regarding the *Meeting Reminder*, one user said: "This is a good feature since [Jay] is making sure that the user won't be late for her meeting" (P3). Others concurred: "a good reason to interact" (P2). For two participants, emergency situations were the only acceptable instances for proactive interventions: "In other cases, it's just annoying" (P4). Participants also pointed out benefits for certain user groups: "This can really help with accessibility, especially for elderly and people with physical disabilities" (P10). One participant found the verbal support in the *Emergency* situation "especially helpful for children or the elderly. The system can also further instruct them" (P15).

Further, the proactive assistance for the *Technical Support* was perceived positively: "[Jay] was smart enough to understand the initial question was aimed at another person. After seeing that no solution can be found, it jumps in and helps" (P6). Reacting to indirect calls for assistance was also highlighted for the *Cooking Inspiration* scenario: "The character is mentioning that she has no clue, and she needs help" (P5) without addressing the VA. "The system was smart enough to detect a problem. It's not just answering a question, but rather trying to solve a problem it has detected" (P6). This was considered a meaningful "entry point" for the agent to proactively intervene. Speculating about the continuation of these two scenarios, all participants but two described that the proactive offer was accepted gladly by the users.

Overall, a common feeling observed during the interviews was the indecisiveness of participants to find proactivity helpful or not, when they found interactions intrusive but at the same time useful. About the *Disagreement Clarification*, one user said: "Very useful but very scary. It can destroy you but it will also cut the discussion short" (P8). In the speculated scenario continuation task, participants often thought the characters would feel violated, but still find the agent's intervention helpful, e.g., regarding *Health Risk*: "Even though she feels violated, she agrees to set an appointment" (P7). Similarly, for *Nudging*, "The user would get offended and say 'leave me alone!' But he would think about it and reflect on it later" (P6).

Interpretation. These results show that there are several situations in which users find the proactivity both useful and appropriate. However, a common pattern that we noticed was the dilemma of proactive interventions being perceived as helpful but at the same time disproportionately intrusive. We call this the proactivity dilemma. For several scenarios, participants were ambivalent about whether the intervention was overall desirable or not: a double-edged sword. This conflict of useful but invasive interventions, such as regarding health risks, is also visible in the quantitative results shown in Figure 3 (right) where one can clearly see that the relationship between both dimensions is not as uniform on the right graph (invasiveness-usefulness) as it is on the left (usefulness-appropriateness) and that the former also shows a somewhat wider spread.

Further, our results confirmed that urgency plays a big role in the appropriateness of proactive interventions, supporting the findings of previous works [7, 11, 28]. We observed throughout the study that the agent's proactive intervention was perceived as highly useful when users' health might be at risk. In such cases, people would not prioritise their privacy but were still concerned about insensitive interventions, aligning with previous research by Tabassum et al. [38]. Generally, the more serious and urgent the topic, the more useful and appropriate it was found to provide proactive assistance, e.g., when facing potential financial or professional damage. Proactively reminding users about their important upcoming activities or events was also highlighted as an appropriate and useful intervention. The familiarity of such interactions through existing digital services could be a reason for the acceptance of this form of proactive intervention.

4.2 Privacy and Mistrust

Even though we asked our participants to put aside any privacy and data protection concerns, they were the biggest worries among participants. One user mentioned: "I don't want the big companies to use all my data" (P4). A common demand amongst the interviewees was transparency and control in data processing: "If I know where my information is being processed and used, I can decide better to use such systems or not" (P12). Some participants were concerned about the misuse of personal data for hidden agendas or providing proactive advertisements: "[the agent] might give me suggestions that are influenced by political reasons or advertisements and try to control my behaviour based on that" (P10). Another concern was about an entity intruding into the private environment: "It's like another person is always at your home" (P12). They found it "really scary that everything could be monitored" (P8). These participants argued that people would constantly feel "observed"

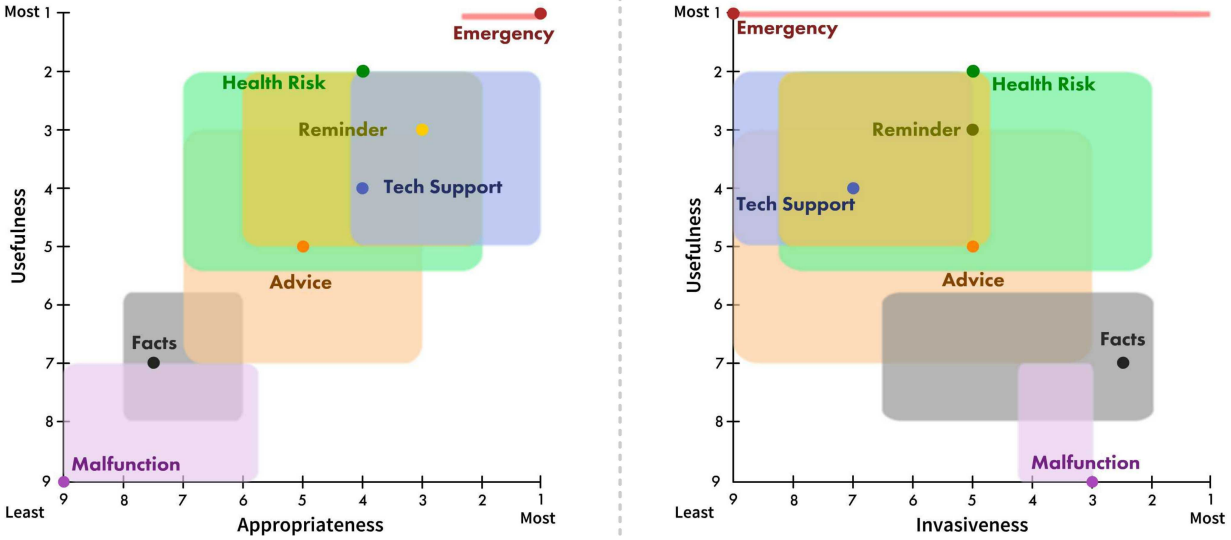


Figure 3: The medians and 75% ranges of the card sorting ranks of mapping usefulness to appropriateness (left) and to invasiveness (right) for different types of interactions: Malfunction (S8), Facts (S4, S5), Advice (S3, S6), Technical Support (S7), Reminder (S1), Health Risk (S2), Emergency (S9).

or "judged". This was especially prominent for scenarios where the agent interrupted a conversation. Mistrust was further expressed about "false alarms" and "misinterpretations" of certain situations and user states or behaviours by the agent which might create anxiety or cause stress in people: "If it's diagnosing you and it's wrong, it will create additional anxiety" (P9). Two even indicated mistrust about the reliability of the *Emergency* alarm.

Interpretation. In order for VAs to be proactive, they require more information about users' environment and behaviour, meaning more personal data needs to be processed to provide such services. During our sessions, it became evident that participants' main hurdle for adopting proactive VAs was the privacy aspect, in line with previous literature [23, 38]. Participants were worried about the misuse of their personal data by companies providing such assistants and third parties. Another concern was related to having an additional entity in the home that is not just a passive servant – like current VAs – but rather a character that aims at taking an active role in users' private space and family life. The participants associated these interferences with paternalism and a lack of control over the device, fearing negative social repercussions in multi-user settings. Therefore, to build trust and set boundaries, one approach could be that proactive VAs initially (e.g., in the first weeks of use) initiate proactive conversations in single-user contexts only.

4.3 Consideration of Social Context

Generally, participants were sceptical about the agent's social awareness. Seven participants found Jay's interventions disruptive and intrusive when they interfered with ongoing conversations:"[Jay] should not stop the thinking process and break conversations. It damages the human-human interaction" (P1). The proactive intervention was then considered "ruining the magic of the discussion" (P15). Two participants even perceived these interruptions as "creepy". The interjections were considered unwelcome because the agent was seen "as a tool rather than an equal conversation partner" (P7). With this unassailable interlocutor present, it felt to one participant "like a contract: everything is noted down. That's very stressful" (P15). The content of the conversation was described as an important factor for proactivity by seven participants: "If it is an intimate conversation, [Jay] should not really intervene" (P10). Two participants were concerned about the missed opportunity of socializing and bonding with another person due to the imposed help by the agent: "This is not received as an act of helping, but rather programmed" (P15). Further, the presence of people in the room was a common theme: "Emotional connection between me and my visitors is the key factor" (P3). In the presence of other people, 12 participants preferred the agent to be proactive only if it was an urgent matter.

Moreover, most participants found it frustrating or unpleasant when the agent corrected users: "People would feel bad about it. No one wants to be corrected" (P14). One person was torn as "this can be helpful, but it can hurt people's feelings" (P3). When the agent was contradicting one user while supporting another, participants found it even more insensitive. Regarding the *Disagreement Clarification*, verifying what was previously agreed was seen as the assistant was taking sides and seven people suspected dissatisfaction of one of the parties. They believed that such well-intended interventions "can potentially cause users to argue" (P13) and that "this could add more oil to the flame" (P1). About the *Fact Checking*, however, one participant assumed: "I think in this case, none of [the users] is correct, so the speaker was being helpful. If one of those people

was right, then the others would feel bad" (P14). Four participants speculated that the users in this scenario might feel offended, and three presumed that the proactive intervention would cause social awkwardness. In contrast, a small number of participants were in favour of these interventions, because "it's nice to be corrected" (P7) or "it's factual and cuts the discussion short" (P8). Similarly, two people appreciated the *Disagreement Clarification*: "I love this example. I think these arguments come up quite often and everyone thinks *they* are right. Personally, in this situation, I would like to have that. I always dreamed about having such a system to check for the truth" (P6).

Interpretation. In multi-user scenarios, the interventions in which the agent would help people to resolve an issue and save time were perceived positively. However, other than emergency situations, these were only perceived to be appropriate when the people had a chance to first try to resolve the matter by themselves. Participants generally thought that when the agent detected a question that was aimed at other people, responding to such questions before the intended person got a chance to respond was perceived as annoying and interfering. However, if the intended person could not properly respond to these questions or inquiries, the agent's intervention was considered useful and appropriate. For example, in the *Technical Support* scenario, the agent intervenes based on a request for help but only does so after the addressed person says they are not able to help at that point. Participants assumed that the agent was aware of the context and could appropriately detect an opportune moment to engage in the ongoing conversation. However, participants raised a concern about the agent taking away an opportunity of bonding, even if it is being helpful. They frequently mentioned that the agent's intervention in social situations is disruptive and could potentially damage human-human interaction. In accordance with previous research [27, 42], understanding the relationship between the people who are co-located, as well as the seriousness and intimacy of the conversation, were pointed out as important factors for the appropriateness of the agent's intervention in these situations.

Moreover, when the agent corrected people, some participants found it inappropriate, annoying, patronising or even insulting. The *Disagreement Clarification* scenario was rated most invasive and ranked second to last in terms of appropriateness. One reason for this was that in this scenario, the conversation was perceived as private. Additionally, the agent's intervention contradicts one of the people present and approves of the other, which resolves the disagreement but could further fuel the conflict. Nevertheless, some participants still found this highly useful and wished for such systems in their households, e.g., to cut discussions short. This example illustrates well that there seem to be major individual differences in how the proactive interventions are perceived.

4.4 Configuration and Control

A common desire amongst the participants was the ability to control and configure the system's proactive actions, in particular concerning the timing and topics. Three participants suggested the possibility to switch proactivity off temporarily. Four wanted to regulate interventions based on who is present in the room. Limiting proactive interventions at specific times of the day was suggested by three participants. One proposed to set the agent's proactivity extent using a slider in the settings. Hence, the users' agency was raised as a concern among participants. They found certain proactive interventions of Jay patronising and imperious. Participants did not like the assistant playing the role of someone who is controlling certain aspects of their lives: "I'm a person and I decide for my life. AI should not decide for me" (P4). This was particularly the case for the *Nudging* scenario. Eight participants explained that proactive features without prior approval would not be acceptable, in particular when the agent tries to nudge users towards a healthier behaviour: "If I have activated this in the settings, I would be more open to it. But if it is unasked for, I would be really annoyed" (P10). Without having asked for advice, a participant had the impression as if the agent "is judging you" (P13). Accordingly, ten participants expected users to ignore the intervention, seven said users would get frustrated, and two thought users would even disconnect the intrusive device. For one participant, the *Meeting Reminder* scenario was all about who is in control: "I feel like the system is forcing you to be productive and be a useful part of society. It takes my mind to dark places where people cannot control the system anymore. Autonomy is more important to me" (P7) Beyond customisation, participants also hoped for the system to automatically adjust over time. Whether manual or automatic, for one person "it needs to be adapted enough to the user's needs in order to understand when it's really needed – and when not" (P9).

Interpretation. Participants were concerned about their possible loss of agency. The feeling of being controlled and patronised by an agent was expressed as a worry. Similar to the findings by Luria et al. [23], several participants did not like it when the agent was suggesting healthier behaviour, i.e., avoiding extensive binge-watching. Based on our observations, the factors that would increase the chance of appropriateness for such interventions were the phrasing and the predictability of the interaction based on pre-configuration by the users. It was recommended for the agent's phrasing to be polite, calming, and suggestive rather than imposing. Correspondingly, participants wanted to have control over proactive interventions and be able to configure times and topics so that they could anticipate interactions to some extent and have more authority. To this end, such proactive VAs are ideally highly customisable and personalised based on individual user needs and preferences as suggested by previous research, such as regarding how short users want their VA's responses to be [14].

4.5 Initiating and Phrasing Interventions

How to introduce proactive interventions was a recurring theme during the interviews. For most of the interactions, participants suggested that the agent should ask for permission or give some kind of cue before speaking: "Maybe it is more acceptable if [Jay] says 'sorry to interrupt' " (P14). Some thought it would be a good compromise to first announce the subject without being too specific yet, such as: "I noticed something about your TV usage. Would you like me to share it?" In the proposed solutions, we identified three levels of initiation:

- *Non-Verbal Cues*: The agent indicates an intervention with a visual or auditory signal but then waits for the user's prompt to proceed.

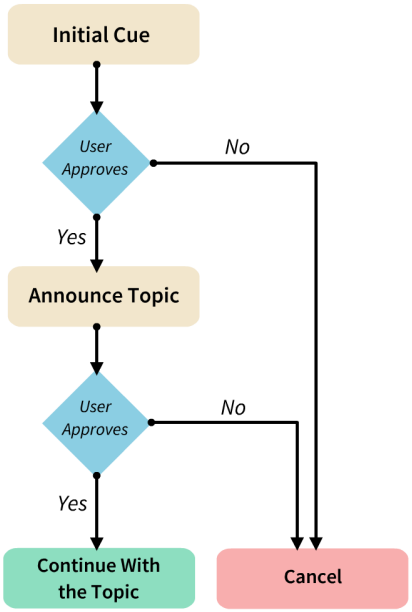


Figure 4: The figure shows the proposed initiation process model to proactively interact with people.

- *Verbal Cue*: The agent announces the subject but waits for the user's permission to proceed.
- *Direct Interventions*: The agent brings up the subject directly.

Direct interventions were mainly suggested for urgent or health-related scenarios but also for saving time. When interrupting conversations between people, non-verbal cues were preferred as they are the least distracting. In the interview task in which we asked participants to re-imagine the agent interventions to improve invasive or inappropriate scenarios, they either wanted the system to give a cue first or to be reactive. The intentionally misplaced *Fact Spoiler* was strongly criticised by the participants. Twelve people speculated that people would disconnect the device in such situations. Specifically, because the agent would not ask for the users' permission to speak, participants found it highly invasive and most inappropriate. One person was reminded of "the annoying kid in the class that screams the answer" (P10).

Further, when re-imagining scenarios, the wording was often adjusted. Concerning health-related issues, participants proposed phrasing the suggestion more cautiously: "Some people may perceive such news as shocking and get some other effects from it. It can create anxiety" (P11). Others did not want the agent to sound patronising or judgmental: "I have a tip for you regarding your health, do you want me to share it with you?" (P1). Instead of assuming a medical diagnosis and booking a doctor's appointment, two participants recommended asking clarifying questions, e.g., how they are feeling or if they have any other symptoms: "It is better if [the agent] gathers more information before making a conclusion and providing suggestions" (P13). Two participants indicated that, where possible, the agent should even help the user deal with stress, such as: "You don't have to worry, I can help you with that" (P8). Overall, the participants phrased their suggested initiations in a polite and calming manner, gently "building up" potentially distressing topics while keeping them goal-oriented and succinct.

Interpretation. From these insights, we can conclude that in most situations, participants expected the agent to ask for permission before conversing. This is in line with results by Arias et al. [4] who suggested that the agent should make sure the users are willing to interact at the specific moment. This permission request could be communicated in various forms. Verbal cues would have high conversational fidelity in relation to human conversations, such as addressing the user by name ("Excuse me, Alex?") or polite phrases ("May I interrupt?") as we suggested in our previous work [31]. A more subtle approach could be non-verbal cues of different modalities, such as abstract audio or light indicators. Depending on the ongoing activity, the preferences of our participants differed. The cue should not distract people from their activity unless it is an urgent matter requiring a striking cue. Verbal cues were described as most distracting, followed by audible cues. Visual cues were described as the least distracting. Based on our findings, we propose an initiation process model for VAs to proactively approach users in non-urgent situations, as illustrated in Figure 4. It starts with an *initial cue*, where the agent indicates that it would like to speak. After user approval, the agent moves on with introducing the *topic of intervention*. If that is also approved by the user, the agent can proceed with the action. In urgent cases, for less sensitive topics or in single-user settings, the second step could be skipped or combined with the first. Although this model could help make certain proactive behaviours more acceptable, the configuration of and control over the types of proactive behaviours as outlined in the previous section must always come first when designing such systems.

5 LIMITATIONS AND FUTURE WORK

In this research, we investigated proactive VA behaviours in a home environment as one of the most predominant use cases for VAs through a selection of storyboards depicting everyday situations. Although the broader insights of this evaluation can be applied to other settings, future work should investigate proactive VAs in other environments such as work and public environments. Moreover, the sample was skewed toward young ($M = 27.86$) and on average, more educated users, and therefore, may not be fully representative of possible VA users. This is particularly relevant when considering that in the scenarios, users with various demographics were present (e.g., the elderly person in the *Health Risk* scenario). Future studies could validate our findings by investigating a wider population and the specific user groups that certain proactive features may be designed for. In our study, we witnessed that individual personal differences can also be a decisive factor in terms of finding proactive VAs appropriate. Differences in user traits (e.g., personality) may lead to different preferences on proactive VAs, which should be incorporated and reflected in further studies.

The method applied in this investigation has its limitations and advantages. Since proactive VAs that have comparable capabilities to those illustrated in our storyboards are not yet available in the

market, we explored people's opinions on these features by presenting hypothetical scenarios. We conducted interactive interviews involving various tasks on a digital whiteboard that engaged participants to contemplate about the presented design space from different angles. As our method requires the participants to immerse and speculate, it enables the investigation of interactions with future technologies that would be intricate or expensive to build. It further enables evaluating aspects of the system that would be impossible to simulate realistically, such as emergency situations or delicate private settings. However, since participants did not experience the situations and proactive behaviours themselves, their perceptions may not reflect real-world experience. Furthermore, it is important to note that some of the services presented in the storyboards may also be supported by other technologies and not solely VAs. In this study, we sought to explore what needs to be considered when developing such features for Voice Assistants.

6 CONCLUSION

This research explores desirable circumstances for proactive interventions by VAs in domestic settings. The findings of our scenario-based study show that people see great benefit in proactivity, specifically in cases of important reminders, time-saving interventions, or emergency support. However, great concerns such as privacy implications, potential loss of agency, and interference with social activities may inhibit the adoption of such systems. Based on the interpretation of our results, we believe that the desirability of proactive interactions depends on the following factors.

Significance. The more urgent or critical the topic, the more appropriate it is for a VA to proactively intervene. The desirability is high under circumstances with a large scope or grave consequences.

Social Context and Environment. Proactive VAs should accurately identify the environmental and social context including the presence of other users or guests, the closeness of their relationships, the type and sensitivity of the ongoing activity, and the time of the day.

Agency and Control. The user needs to be able to adjust and configure proactive features including the times and topics for interventions. Users should have control over when the agent is allowed to listen and observe its environment, and when it is allowed to intervene so that they could anticipate such interactions.

Individual User Factors. As there seem to be major differences between individuals in how certain interventions are perceived, proactive VAs should be able to consider individual user factors such as physical and cognitive abilities (e.g., of young children or elderly users), the current physical and emotional state (e.g., stress level, sadness, or fatigue), and the personality and preferences of the user (e.g., privacy needs or agency).

Form of Execution. When initiating an interaction, the agent should generally first request permission using verbal or non-verbal cues, and announce the topic of intervention – unless it is time-critical as in an emergency. Furthermore, the intent should be phrased so that it is polite, not imposing, and does not create a feeling of unease, while at the same time being goal-oriented and concise. When users got used to certain interventions over time or gave permission, the VA may get right to the point.

Altogether, as long as the proactivity dilemma is carefully considered by finding a positive balance with suggestions that are perceived as more helpful than invasive, there seems to be great potential in proactive VAs.

ACKNOWLEDGMENTS

This work was partially funded by the Leverhulme Trust as part of the Doctoral Training Programme for the Ecological Study of the Brain (DS-2017-026) and the Klaus Tschira Foundation.

REFERENCES

[1] Herman Aguinis and Kyle J. Bradley. 2014. Best Practice Recommendations for Designing and Implementing Experimental Vignette Methodology Studies. *Organizational Research Methods* 17, 4 (2014), 351–371. https://doi.org/10.1177/1094428114547952 arXiv:https://doi.org/10.1177/1094428114547952

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[3] Salvatore Andolina, Valeria Orso, Hendrik Schneider, Khalil Klouche, Tuukka Ruotsalo, Luciano Gamberini, and Giulio Jacucci. 2018. Investigating Proactive Search Support in Conversations. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 1295–1307. https://doi.org/10.1145/3196709.3196734

[4] Kika Arias, Sooyeon Jeong, Hae Won Park, and Cynthia Breazeal. 2020. Toward Designing User-centered Idle Behaviors for Social Robots in the Home.

[5] Michael Braun, Anja Mainz, Ronee Chadowitz, Bastian Pfleging, and Florian Alt. 2019. At your service: Designing voice assistant personalities to improve automotive user interfaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–11.

[6] John M. Carroll. 1999. Five Reasons for Scenario-Based Design. In *Proceedings of the Thirty-Second Annual Hawaii International Conference on System Sciences-Volume 3 - Volume 3 (HICSS '99)*. IEEE Computer Society, USA, 3051.

[7] Narae Cha, Auk Kim, Cheul Young Park, Soowon Kang, Mingyu Park, Jae-Gil Lee, Sangsu Lee, and Uichin Lee. 2020. Hello There! Is Now a Good Time to Talk? Opportune Moments for Proactive Interactions with Smart Speakers. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 74 (Sept. 2020), 28 pages. https://doi.org/10.1145/3411810

[8] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300705

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

[10] Philip R. Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R. Cowan. 2019. Mapping Perceptions of Humanness in Intelligent Personal Assistant Interaction. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services* (Taipei, Taiwan) *(MobileHCI '19)*. ACM, New York, NY, USA, Article 5, 12 pages. https://doi.org/10.1145/3338286.3340116

[11] Justin Edwards, Christian Janssen, Sandy Gould, and Benjamin R. Cowan. 2021. Eliciting Spoken Interruptions to Inform Proactive Speech Agent Design. In *CUI 2021 - 3rd Conference on Conversational User Interfaces* (Bilbao (online), Spain) *(CUI '21)*. Association for Computing Machinery, New York, NY, USA, Article 23, 12 pages. https://doi.org/10.1145/3469595.3469618

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

[13] Greg Guest, Arwen Bunce, and Laura Johnson. 2006. How many interviews are enough? An experiment with data saturation and variability. *Field methods* 18, 1 (2006), 59–82.

[14] 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* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 321, 12 pages. https://doi.org/10.1145/3491102.3517684

[15] Scott Hudson, James Fogarty, Christopher Atkeson, Daniel Avrahami, Jodi Forlizzi, Sara Kiesler, Johnny Lee, and Jie Yang. 2003. Predicting Human Interruptibility with Sensors: A Wizard of Oz Feasibility Study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 257–264. https://doi.org/10.1145/642611.642657

[16] Soowon Kang, Heepyung Kim, Youngtae Noh, and Uichin Lee. 2021. *Poster: Toward Context-Aware Proactive Conversation for Smart Speakers*. Association for Computing Machinery, New York, NY, USA, 38–40. https://doi.org/10.1145/3460418.3479306

[17] Auk Kim, Woohyeok Choi, Jungmi Park, Kyeyoon Kim, and Uichin Lee. 2018. Interrupting Drivers for Interactions: Predicting Opportune Moments for In-Vehicle Proactive Auditory-Verbal Tasks. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 175 (Dec. 2018), 28 pages. https://doi.org/10.1145/3287053

[18] Auk Kim, Jung-Mi Park, and Uichin Lee. 2020. Interruptibility for in-vehicle multitasking: influence of voice task demands and adaptive behaviors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–22.

[19] Mitsuki Komori, Yuichiro Fujimoto, Jianfeng Xu, Kazuyuki Tasaka, Hiromasa Yanagihara, and Kinya Fujita. 2019. Experimental Study on Estimation of Opportune Moments for Proactive Voice Information Service Based on Activity Transition for People Living Alone. In *Human-Computer Interaction. Perspectives on Design*, Masaaki Kurosu (Ed.). Springer International Publishing, Cham, 527–539.

[20] Matthias Kraus, Marvin Schiller, Gregor Behnke, Pascal Bercher, Michael Dorna, Michael Dambier, Birte Glimm, Susanne Biundo, and Wolfgang Minker. 2020. "Was That Successful?" On Integrating Proactive Meta-Dialogue in a DIY-Assistant Using Multimodal Cues. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20)*. Association for Computing Machinery, New York, NY, USA, 585–594. https://doi.org/10.1145/3382507.3418818

[21] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, Are You Listening? Privacy Perceptions, Concerns and Privacy-Seeking Behaviors with Smart Speakers. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 102 (Nov. 2018), 31 pages. https://doi.org/10.1145/3274371

[22] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[23] 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*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376311

[24] Nathan Malkin, Joe Deatrick, Allen Tong, Primal Wijesekera, Serge Egelman, and David Wagner. 2019. Privacy attitudes of smart speaker users. *Proceedings on Privacy Enhancing Technologies* 2019, 4 (2019), 250–271.

[25] Donald McMillan. 2017. Implicit Interaction Through Machine Learning: Challenges in Design, Accountability, and Privacy. In *Internet Science*, Ioannis Kompatsiaris, Jonathan Cave, Anna Satsiou, Georg Carle, Antonella Passani, Efstratios Kontopoulos, Sotiris Diplaris, and Donald McMillan (Eds.). Springer International Publishing, Cham, 352–358.

[26] Donald McMillan, Antoine Loriette, and Barry Brown. 2015. Repurposing Conversation: Experiments with the Continuous Speech Stream. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3953–3962. https://doi.org/10.1145/2702123.2702532

[27] O. Miksik, I. Munasinghe, J. Asensio-Cubero, S. Reddy Bethi, S-T. Huang, S. Zylfo, X. Liu, T. Nica, A. Mitrocsak, S. Mezza, R. Beard, R. Shi, R. Ng, P. Mediano, Z. Fountas, S-H. Lee, J. Medvesek, H. Zhuang, Y. Rogers, and P. Swietojanski. 2020. Building Proactive Voice Assistants: When and How (not) to Interact. arXiv:2005.01322 [cs.HC]

[28] Florian Nothdurft, Stefan Ultes, and Wolfgang Minker. 2015. Finding appropriate interaction strategies for proactive dialogue systems—an open quest. , 73–80 pages.

[29] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal, QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174214

[30] Aung Pyae and Tapani N. Joelsson. 2018. Investigating the Usability and User Experiences of Voice User Interface: A Case of Google Home Smart Speaker. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct* (Barcelona, Spain) *(MobileHCI '18)*. Association for Computing Machinery, New York, NY, USA, 127–131. https://doi.org/10.1145/3236112.3236130

[31] Leon Reicherts, Nima Zargham, Michael Bonfert, Yvonne Rogers, and Rainer Malaka. 2021. May I Interrupt? Diverging Opinions On Proactive Smart Speakers. In *CUI 2021 - 3rd Conference on Conversational User Interfaces* (Bilbao (online), Spain) *(CUI '21)*. Association for Computing Machinery, New York, NY, USA, Article 34, 10 pages. https://doi.org/10.1145/3469595.3469629

[32] A Joy Rivera. 2014. A socio-technical systems approach to studying interruptions: Understanding the interrupter's perspective. *Applied ergonomics* 45, 3 (2014), 747–756.

[33] Maria Schmidt and Patricia Braunger. 2018. A Survey on Different Means of Personalized Dialog Output for an Adaptive Personal Assistant. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization* (Singapore, Singapore) *(UMAP '18)*. Association for Computing Machinery, New York, NY, USA, 75–81. https://doi.org/10.1145/3213586.3226198

[34] Maria Schmidt, Wolfgang Minker, and Steffen Werner. 2020. User Acceptance of Proactive Voice Assistant Behavior. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2020*, Andreas Wendemuth, Ronald Böck, and Ingo Siegert (Eds.). TUDpress, Dresden, Dresden, Germany, 18–25.

[35] Maria Schmidt, Daniela Stier, Steffen Werner, and Wolfgang Minker. 2019. Exploration and assessment of proactive use cases for an in-car voice assistant. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, Peter Birkholz and Simon Stone (Eds.). TUDpress, Dresden, Dresden, Germany, 148–155.

[36] Rob Semmens, Nikolas Martelaro, Pushyami Kaveti, Simon Stent, and Wendy Ju. 2019. Is Now A Good Time? An Empirical Study of Vehicle-Driver Communication Timing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300867

[37] Petra-Maria Strauss and Wolfgang Minker. 2010. *Proactive spoken dialogue interaction in multi-party environments*. Springer, Cham.

[38] Madiha Tabassum, Tomasz Kosiński, Alisa Frik, Nathan Malkin, Primal Wijesekera, Serge Egelman, and Heather Richter Lipford. 2019. Investigating Users' Preferences and Expectations for Always-Listening Voice Assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–23.

[39] Liam D. Turner, Stuart M. Allen, and Roger M. Whitaker. 2015. Interruptibility Prediction for Ubiquitous Systems: Conventions and New Directions from a Growing Field. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 801–812. https://doi.org/10.1145/2750858.2807514

[40] Konstantina Vasileiou, Julie Barnett, Susan Thorpe, and Terry Young. 2018. Characterising and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period. *BMC Medical Research Methodology* 18, 1 (Nov. 2018), 148. https://doi.org/10.1186/s12874-018-0594-7

[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. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445536

[42] Jing Wei, Tilman Dingler, and Vassilis Kostakos. 2021. Developing the Proactive Speaker Prototype Based on Google Home. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–6.

[43] Jing Wei, Tilman Dingler, and Vassilis Kostakos. 2022. Understanding User Perceptions of Proactive Smart Speakers. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 185 (dec 2022), 28 pages. https://doi.org/10.1145/3494965

A APPENDIX

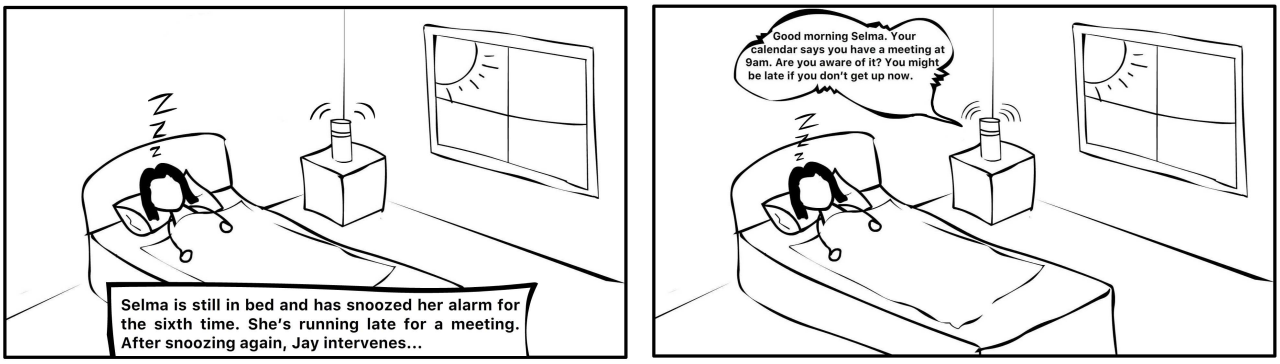


Figure 5: Scenario 1: Meeting Reminder

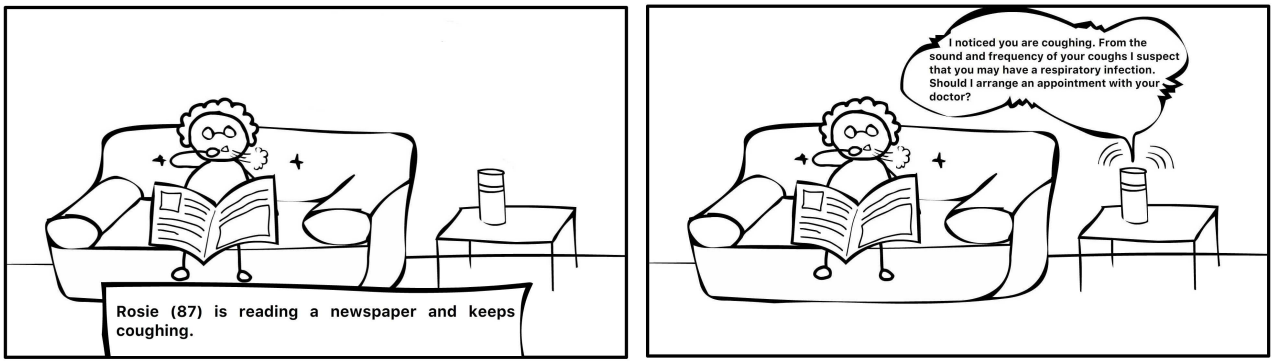


Figure 6: Scenario 2: Health Risk

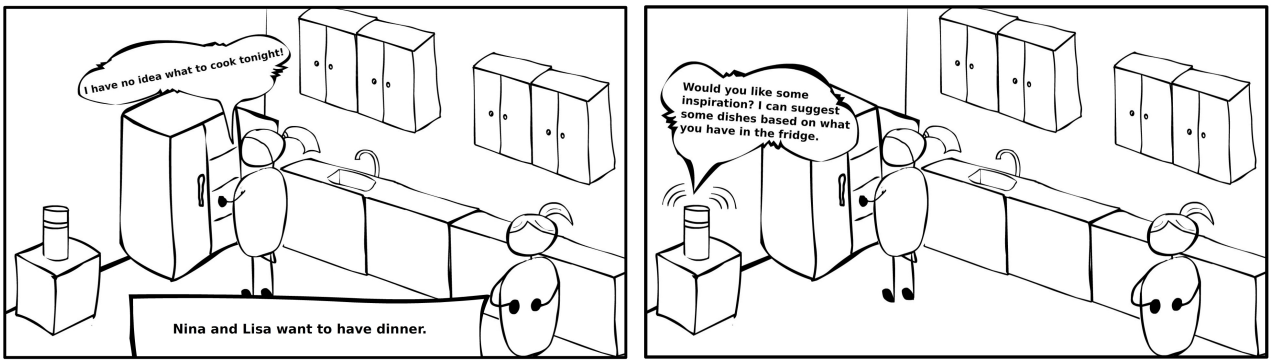


Figure 7: Scenario 3: Cooking Inspiration

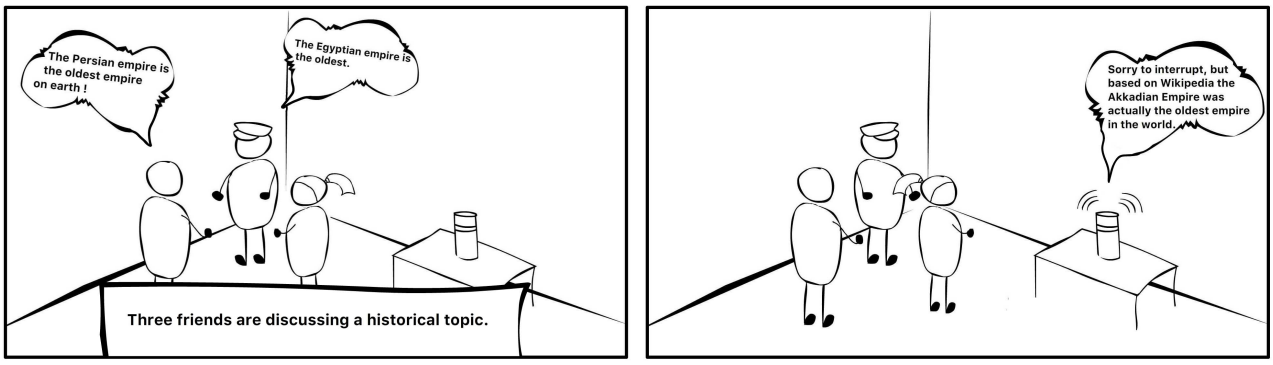


Figure 8: Scenario 4: Fact Checking

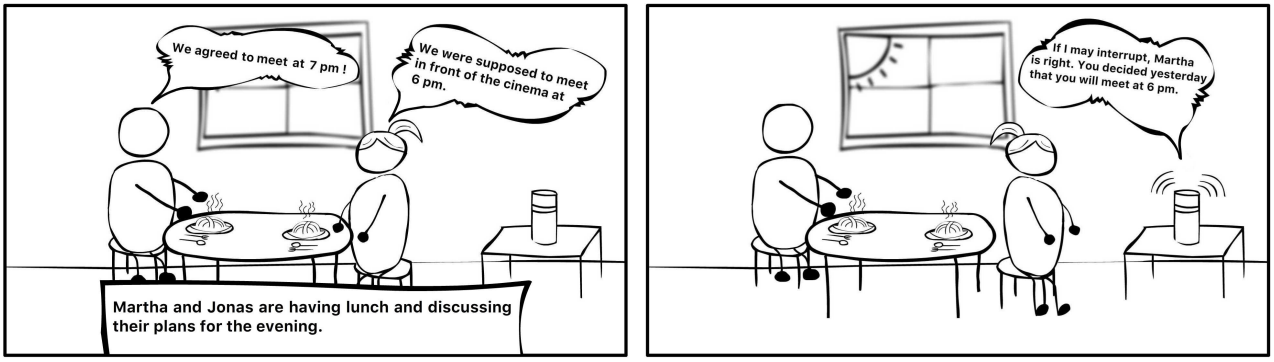


Figure 9: Scenario 5: Disagreement Clarification

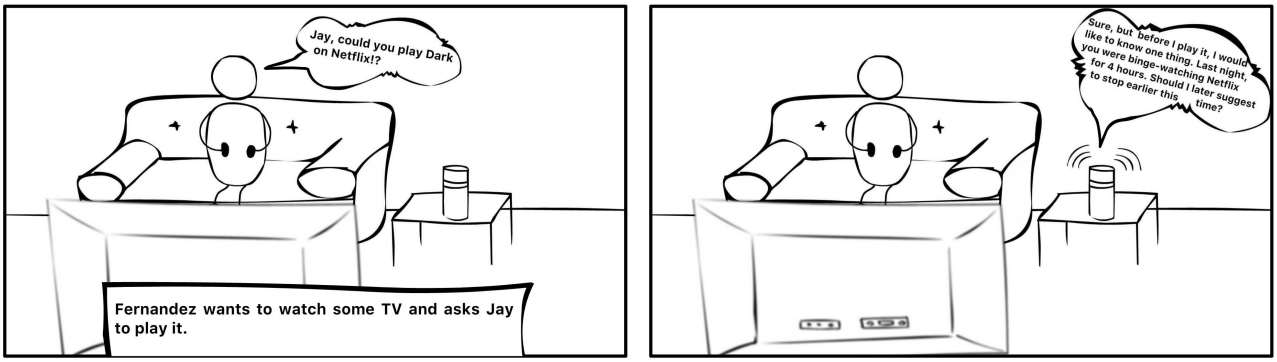


Figure 10: Scenario 6: Nudging

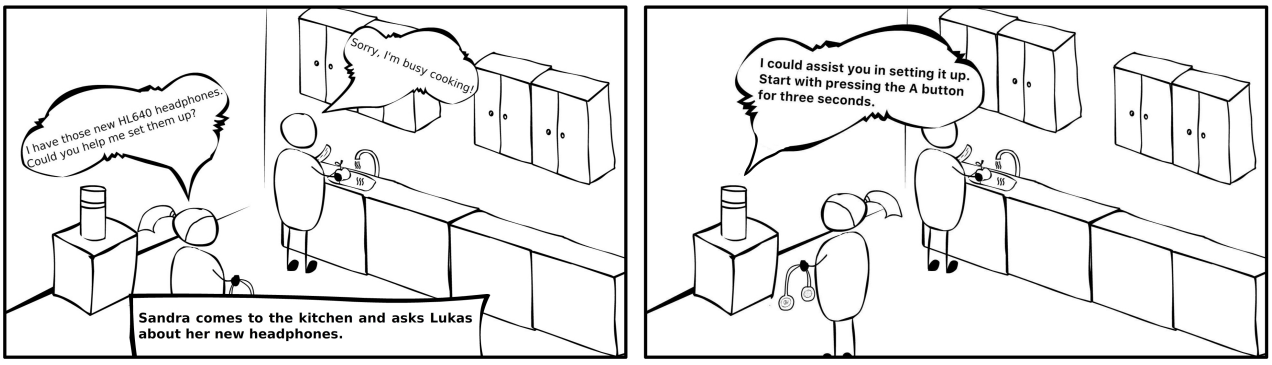


Figure 11: Scenario 7: Technical Support

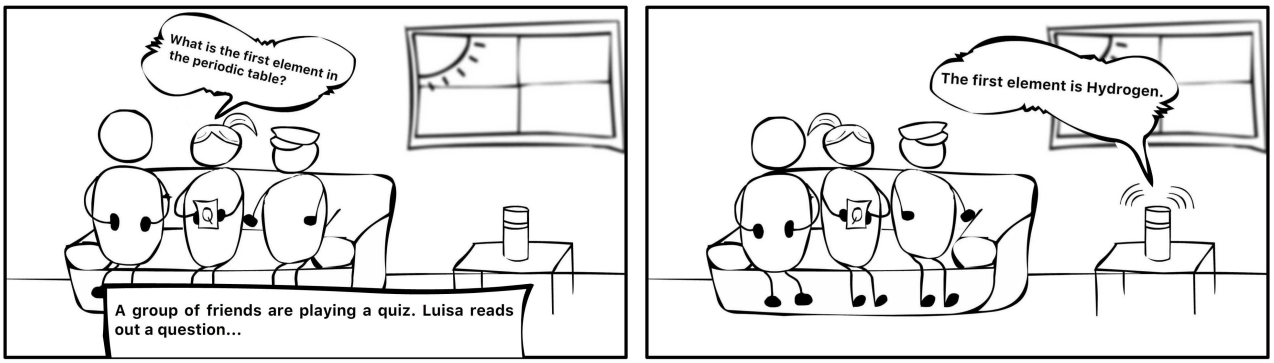


Figure 12: Scenario 8: Fact Spoiler

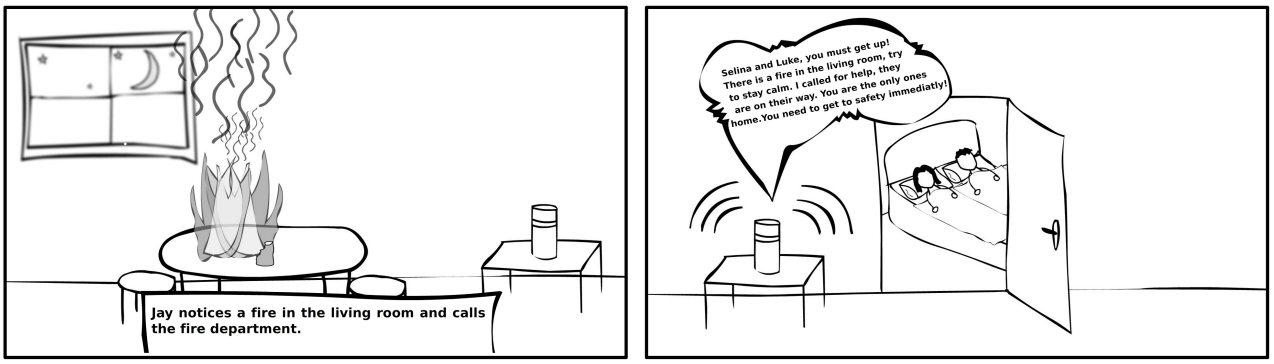


Figure 13: Scenario 9: Emergency

Publication 10

Tickling Proactivity: Exploring the Use of Humor in Proactive Voice Assistants

Nima Zargham, Leon Reicherts, Vino Avanesi, Yvonne Rogers, and Rainer Malaka

In Proceedings of the 22th International Conference on Mobile and Ubiquitous Multimedia (MUM '23). New York, NY, USA, 2023. Association for Computing Machinery.

Personal contribution to this work: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, validation, visualization, and contribution to all parts of the manuscript.

ISBN: 979-8-4007-0921-0/23/12 DOI: 10.1145/3626705.3627777

Tickling Proactivity: Exploring the Use of Humor in Proactive Voice Assistants

Nima Zargham
zargham@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

Leon Reicherts
l.reicherts.17@ucl.ac.uk
University College London
United Kingdom

Vino Avanesi
avanesi@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

Yvonne Rogers
y.rogers@ucl.ac.uk
University College London
United Kingdom

Rainer Malaka
malaka@uni-bremen.de
Digital Media Lab
University of Bremen
Germany

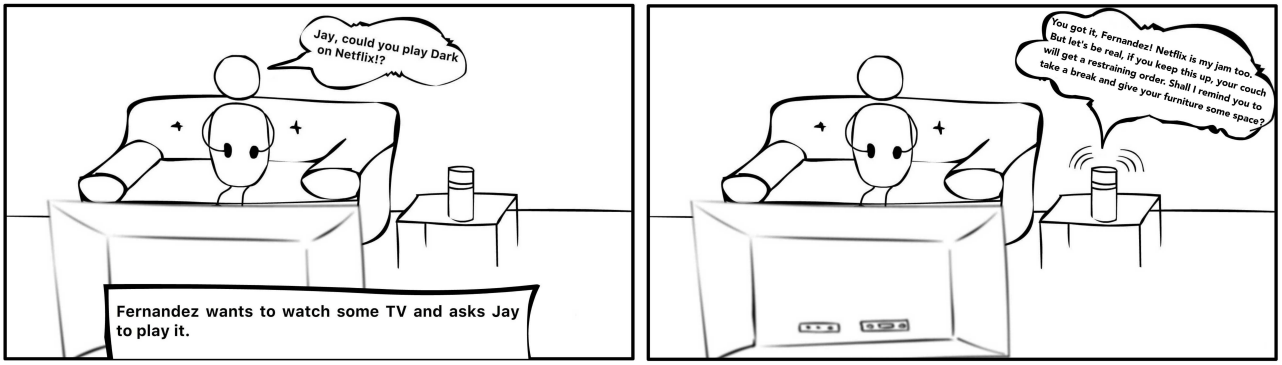


Figure 1: An example storyboard used in our online questionnaire presenting a scenario in which the voice assistant uses humor to respond to the user.

ABSTRACT

With rapid advances in artificial intelligence and natural language processing, voice assistants are evolving into advanced digital personal assistants capable of complex tasks. As they become more proficient at understanding people's behaviors, preferences, intentions, and surroundings, opportunities for proactive interactions emerge. However, despite their potential benefits, people still find certain proactive agent interactions inappropriate and invasive, such as correcting or nudging the user. This study investigates humor's potential to enhance the desirability of proactive agent comments, given its stress-relieving and acceptance-promoting characteristics. We investigate how infusing humor into VA statements affects perceptions of appropriateness and desirability in proactive interventions. We designed storyboards showcasing voice assistants' proactive actions in everyday situations and social contexts. Participants ($N = 50$) assessed these scenarios in an online questionnaire across multiple criteria. Our results reveal that humor's impact on proactive statement desirability is contingent on participants' perceptions of voice assistants and their subjective judgment of the humor.

MUM '23, December 03–06, 2023, Vienna, Austria

ACM ISBN 979-8-4007-0921-0/23/12…$15.00
https://doi.org/10.1145/3626705.3627777

CCS CONCEPTS

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

KEYWORDS

Conversational Agents, Voice Assistants, Home Assistants, Proactivity, Humor, Computational Humor

ACM Reference Format:
Nima Zargham, Leon Reicherts, Vino Avanesi, Yvonne Rogers, and Rainer Malaka. 2023. Tickling Proactivity: Exploring the Use of Humor in Proactive Voice Assistants. In *International Conference on Mobile and Ubiquitous Multimedia (MUM '23), December 03–06, 2023, Vienna, Austria.* ACM, New York, NY, USA, 27 pages. https://doi.org/10.1145/3626705.3627777

1 INTRODUCTION

Voice assistants (VAs) are becoming more advanced and capable of handling complex tasks and conversations. They are commonly used for controlling smart home devices, information gathering, entertainment, online shopping, and time management [60]. With the rise of products such as ChatGPT [13] or smart speakers in homes, conversational agents (CAs) are becoming increasingly important as digital personal assistants. VAs like Apple's Siri, Google Assistant, Microsoft's Cortana (now Microsoft Copilot), and Amazon's Alexa are accessible on various devices such as smartphones, tablets, computers, cars, and smart home devices like Apple HomePod, Google Home, or Amazon Echo. As AI, natural language processing, and sensing technologies advance, researchers predict that these systems will become increasingly proactive [23, 40, 55, 64, 71, 103]. In our previous work [103], we defined proactivity of VAs as "agent-initiated interactions which are triggered by events related to the user(s) and their environment, as opposed to user-initiated inquiries or pre-configured actions, such as reminders, alerts, or routines set by the user." Previous literature has highlighted the opportunities and benefits that proactive VAs can offer to support, probe, or inspire people [64, 93]. Research has shown that people find proactive VAs highly beneficial, specifically in cases of important reminders, time-saving interventions, or emergency support [103]. Despite the benefits that proactivity can bring, there are also potential challenges, in particular concerning privacy [82], lack of interlocutor authenticity [11], or potential loss of agency [103]. Furthermore, in our previous study [103], we witnessed that proactive interventions for correcting people or nudging them for positive behavior change are often perceived as inappropriate and invasive. In general, CAs often fail to meet consumer expectations [76] and are commonly perceived as machine-like, cold, socially inept, untrustworthy, and incompetent [25, 29, 77].

Humor has been shown to be effective in reducing stress [57] and increasing feelings of well-being [52, 53]. Furthermore, research suggests that humor can make difficult or unpleasant information easier to 'digest' [27, 45, 62, 74]. Humor has also shown to be an effective tool in persuasion [50, 95]. Recent research has shown that using humor by CAs enhances service satisfaction [77] and can potentially improve user engagement [78].

Current VAs often use humor to keep people engaged and entertained and compensate for performance limitations [28, 30, 47]. Research on the use of humor in VAs recommends that among the common existing systems, Siri is considered the funniest by people [41, 48]. However, the type of humor used and jokes generated by such systems is often perceived as corny, which can break the illusion of human-likeness, leaving people unhappy, frustrated, and disappointed [47, 75], and damage the emotional connection between humans and the agent. The humor of current VAs is primarily communicated through a number of prescripted jokes, often leading to repetition. One of the most critical elements of humor is timing [58]. Central to its effectiveness are the elements of unpredictability and surprise [6, 88]. The essence of humor lies in its well-timed delivery, aligning appropriately with the situation at hand. This requires agents to possess prior knowledge (e.g., about the user and environment), emotional awareness, situational comprehension, and cultural sensitivity, which often entails proactive actions [101]. Despite all these challenges, previous research highlights that people wish for more humor in VAs, as evident from the requests for jokes from the agent [10]. In a study conducted by Völkel et al. [90], an elicitation study was undertaken to explore users' expectations in interactions with an ideal voice assistant. The study revealed that proactivity was an aspect that users wished for voice assistants to exhibit, as well as the use of humor in some cases. Yet, despite the variety of studies on humor and the proactivity of VAs, none have specifically explored the potential of using humor for proactive interventions by VAs.

To build on the previous work about proactive interactions of VAs, in this work, we aim to explore how humor can impact the desirability of proactive VA statements. We further examine the elements that highlight the appropriateness of using humor by identifying in which context and environment such agent interventions are desirable.

We pursue the following research questions:

RQ1: Can the use of humor by a VA increase the desirability of its proactive interventions?
RQ2: In which situations and context can humor be perceived as more appropriate?

To address our research questions, we employed scenarios presented in our previous study [103] that showcase various proactive actions of a voice assistant in a home setting. We modified the voice assistant's comments in a three-step process to make them humorous, and presented people with two versions of each scenario, once with the use of humor and once without, and asked them to rate the scenarios regarding *usefulness, appropriateness, invasiveness*, and how likely they think the user in the scenario will *consider* what the VA says.

Our findings indicate that humor did not consistently improve the desirability of proactive interventions, and where it was not perceived as humorous, it had diminishing effects. However, desirability can be increased depending on participants' perceptions of VAs and their assessment of whether the VA's humor was actually humorous.

This research addresses the need for more engaging and desirable interactions with VAs. By exploring the use of humor by a proactive voice assistant in a domestic setting, we contribute to understanding when and in which context humor is perceived as desirable. The findings of this study provide insights for designers and developers to create VAs that effectively incorporate humor, leading to an improved user experience with voice assistants. As highlighted by the literature, perception of humor is highly subjective and depends on one's socio-cultural background [10, 99]. Nevertheless, there is potential that certain social and environmental aspects of humor can be explored collectively for enhanced utilization, ultimately improving the user experience with voice assistants.

2 RELATED WORK

Previous research has examined proactive services in various applications and technologies such as context-aware reminders or recommendations [79, 86], health and mental well-being [4, 44], or self-tracking to improve productivity [37, 96]. This section provides an overview of related work on proactivity in VAs, humor in

human-computer interaction (HCI), and the role of humor in social interactions.

2.1 Proactive Voice Assistants

Extensive research has been conducted on system-initiated (proactive) interactions within spoken dialogue systems [34, 63, 81]. Although previous research has shown that proactive interactions can open up new opportunities for supporting, probing, or inspiring people [93], current commercial smart speakers remain primarily reactive with users initiating interactions and support only a minimal set of proactive features [64]. Proactive interactions have demonstrated their capacity to be beneficial across various domains, aiding and engaging users. A survey conducted by Schmidt and Braunger [71] involving 1,550 participants indicated that proactivity is a highly valued attribute of voice assistants among users. Additionally, a study by Völkel et al. [89], exploring people's envisioned interactions with an ideal voice assistant, revealed that many participants expressed a preference for proactive voice assistant behavior.

However, one of the biggest challenges with these systems, which is critical to the user experience, is the timing of the interventions [1, 55, 64, 103]. Since speech responses demand immediate attention, they can interfere with people's ongoing activities. This is unlike GUI-based alerts, where users can often delay it until they are ready to take action [63]. Several researchers have looked into opportune moments to proactively interact with people [7, 38, 40, 63, 71–73, 94]. Opportune moments for interaction refer to moments where the disruption of the user's current activity is at a minimum level [85]. Even though it is a fairly easy task for humans to assess another person's current activity before initiating a conversation, designing such behaviors for agents is very challenging [33, 67, 85]. In addition to pinpointing opportune moments for proactive interactions, one crucial aspect is how the agent would deliver them [20, 23, 103]. An adequate delivery could sometimes mitigate the negative effects when the timing might not be perfect. One possible approach for delivering proactive interventions might be the use of humor, which, to the best of our knowledge, is yet to be explored.

One of the major barriers to users' acceptance of VAs is the topic of privacy [16, 51, 101, 103]. A study by Lau et al. [43] revealed that many individuals hesitate to embrace smart speakers due to concerns about privacy and a lack of trust in the companies behind these devices. Adapting proactive services necessitates a higher level of context awareness and access to more personal data, intensifying people's privacy concerns even further [55]. This concern is particularly pertinent in a home environment, where emphasizing the importance of user privacy and security becomes paramount. A study by Tabassum et al. [82] showed that, while users perceived proactive services useful, they were uncomfortable with the always-listening nature of such systems.

Reviewing the literature on proactive interventions of voice assistants reveals that despite some proactive behaviors causing discomfort and being viewed as disruptive and invasive [3], people still recognize many benefits associated with these types of interactions. Previous works suggest taking into account individual user factors, including their current physical and emotional state (e.g., stress level, sadness, or fatigue), as well as the surrounding environmental and social context, such as the presence of other people or guests, the closeness of relationships, and the nature and sensitivity of ongoing activities, to foster more favorable interactions [55, 103].

The need for VAs to consider the psycho-social context of their operations to minimize disruptions caused by proactive interventions aligns with the approach required for implementing computational humor. This entails a sensitivity to the social context, which will be elaborated upon in the following sub-section.

2.2 Humor in HCI

Humor plays a crucial role in influencing human behavior and promoting positive social interactions across diverse cultures and societies globally [65]. It is a powerful communication tool, allowing individuals to foster connections and navigate social interactions more effectively [26]. Despite the extensive body of literature exploring humor from various disciplines, such as philosophy, literature, and psychology, there remains a lack of consensus regarding a unified theory of humor [65]. Researchers concur that humor represents a cognitive state of joy, often manifested through facial and vocal expressions like smiles and laughter [47]. A previous study suggests that making creative connections, whether understanding jokes or solving math problems, is an innately pleasurable experience [84]. It is recognized as an inherently ambiguous and context-dependent phenomenon, where its interpretation is contingent upon the specific context in which it occurs [17]. Correspondingly, Martin et al. [53] note that four distinct styles of humor are used in human interaction. Two of these are adaptive (*Affiliative* and *Self-Enhancing humor*), and two are maladaptive (*Aggressive* and *Self-defeating humor*). Further studies have supported the existence and impact of these styles across diverse groups [42, 48].

Within the field of human-computer interaction, humor is recognized as a feature that can enhance engagement, usability, and the personification of technology [47, 58, 75]. Moreover, humor has proven effective in facilitating learning, reducing stress, and fostering intrinsic motivation in various contexts [5, 18, 30, 46, 102]. Using humor in machines aims to imbue them with anthropomorphic qualities, creating a sense of relatability and human-like attributes [47, 101]. By incorporating humor, conversational agents strive to connect with users, evoking perceptions of the agents as more human-like and likable [22, 58]. Consequently, humor becomes a means for CAs to foster attachment [47]. It has also been demonstrated that humor can be used to effectively handle situations where the system is unable to respond to users appropriately [10, 48]. Wei et al. [92] found that users find humorous agents more friendly, intimate, and similar to themselves compared to their non-humorous counterparts. Yet, humor is also often acknowledged as one of the most intricate human qualities to replicate in AI agents [58, 59, 101]. Its multifaceted nature makes crafting even a simple joke complex, necessitating various cognitive abilities, including language skills, theory of mind, symbolism, abstract thinking, and social perception. The challenges in teaching computers to comprehend humor stem from its inherent contextual dependencies, encompassing assumptions, morals, attitudes, and taboos deeply ingrained within humanity's history and cultures [31]. Implementing humor in computers involves three fundamental components:

detection (semantic understanding), generation, and delivery [58]. Even though there have been notable advances in these three areas of computational humor, the development of an agent fully capable of recognizing, generating, and using humor is still not achieved [47, 58, 101]. As such, it has been reported that VA companies often employ professional writers to create comedic responses [35, 47, 56]. This suggests that the current state of technology is still not yet at the level where it can produce sufficiently humorous interactions without the help of humans.

Taking a closer look at each of these components confirms this observation. Regarding humor detection, computational algorithms have been developed to identify humor created by humans. Some studies have focused on simpler forms of humor, such as one-liners [66, 83, 87], while others have explored detecting more intricate expressions like sarcasm, which can be challenging even for humans [36, 61, 98]. Concerning humor generation, HCI researchers argue that AI systems still struggle to consistently produce humorous interactions that meet user expectations [48, 58]. However, it must also be noted that recent advancements in generative AI technologies, such as ChatGPT, have shown promising improvements in this area [14]. And finally, the delivery of humor is arguably still the most challenging aspect of computational humor [58]. To deliver humor effectively, agents need to possess substantial background knowledge about the user, their environment, emotional intelligence, and an awareness of social context and culture.

Even though there are several challenges in integrating humor for agents, the literature argues that people wish for more humor in VAs [10]. However, several studies suggest that humor in VAs depends on the individual and is only appreciated by a subset of users [19, 90, 91]. Research by Völkel et al. [90] suggests that the incorporation of humor by a voice assistant is greatly dependent on individual user preferences. The study observed a disparity in user reactions, with some individuals enjoying humor while others disliking it. Consequently, the authors suggest a cautious approach when integrating humor into voice assistant interactions.

2.3 Humor and Social Interactions

As addressed earlier, beyond its entertainment value, humor plays a crucial role in shaping social dynamics, influencing perceptions, and even challenging societal norms [24, 49, 65]. VAs generally exhibit a socially adaptive style of humor, as demonstrated by Kubert and Korshakova [41]. Their study on humor styles employed by VAs revealed that the prevailing style, across all devices, is *affiliative* humor. This humor style seeks to establish connections and foster bonds between individuals [53]. Furthermore, research by Shin et al. [77] has shown that using affiliative humor by chatbots enhances service satisfaction, as opposed to aggressive humor. This aligns with the idea that affiliative humor is not only suitable in terms of psycho-social sensitivity for incorporating humor into VAs, but it also holds the potential for implementing proactive interventions by fostering a social bond between users and agents. In social interactions, humor appreciation is influenced by the group context within which it occurs [21], including the characteristics of the humor initiator [8, 97]. Previous literature emphasizes that the humor initiator's social status and perceived authority influence how their humor is perceived [68]. For instance, humor delivered by someone in a position of power might be interpreted differently than if a peer presented the same humor.

In societal relationships, social status and power are highly sought after, motivating individuals to maintain or elevate their position within the hierarchy [2]. Previous research underscores humor's influence on social status [26]. Effective humor can elevate status in new and established relationships, while failed attempts, like inappropriate jokes, can harm it [9]. Romero and Cruthirds [68] suggest that self-enhancing humor can foster positive connections with higher-status individuals, aiding in establishing rapport with superiors or groups like upper management.

In human-agent interaction, research indicates that the more social agency attributed to artificial agents, the greater the reactance displayed by users [69, 70]. Social agency refers here to the perception of the VA as being capable of social behavior resembling human-human interaction [69, 80].

These observations underscore the importance of understanding users' perceived social equality attributed to VAs for comprehending how humor is received from VAs to users. The characteristics of the humor initiator and users' perception of the artificial agent's social attributes play a significant role in understanding humor's impact in such interactions.

From our examination of existing literature, we establish the following research hypotheses for our study:

- **H1:** The desirability of a proactive intervention is affected depending on how humorous it is perceived.
- **H2:** A correlation exists between how people perceive a VA regarding its social equality and how humorous they find its interventions.

3 METHOD

We conducted an exploratory study consisting of an online survey to examine the impact of humorous proactive interventions by a voice assistant in a domestic setting. Drawing inspiration from scenario-based design methods [15, 64, 103], we used a series of hypothetical storyboards and asked participants to reflect upon and evaluate them. This approach allows us to investigate upcoming technologies despite existing technological constraints. We utilized graphical storyboards to better visualize the situation and spatial configuration of the specific home environment, the user(s), and the smart speaker within the home environment. The efficacy of this method in gaining a good understanding of user perceptions has been demonstrated previously [64, 100, 103].

3.1 Storyboards

We used the scenarios created in our previous study [103] as our initial reference point, depicting proactive VA interventions in a home environment. From their final selection of nine scenarios, we identified five that fell within a moderate range of appropriateness and usefulness and used them as neutral variants for our evaluation.

We employed a three-step approach to design the humorous versions of the scenarios. It is important to note, our aim was not to produce a version of each scenario that would reliably be perceived as humorous by every single individual. It would be unrealistic to try to do so given humor's highly subjective and context-dependent

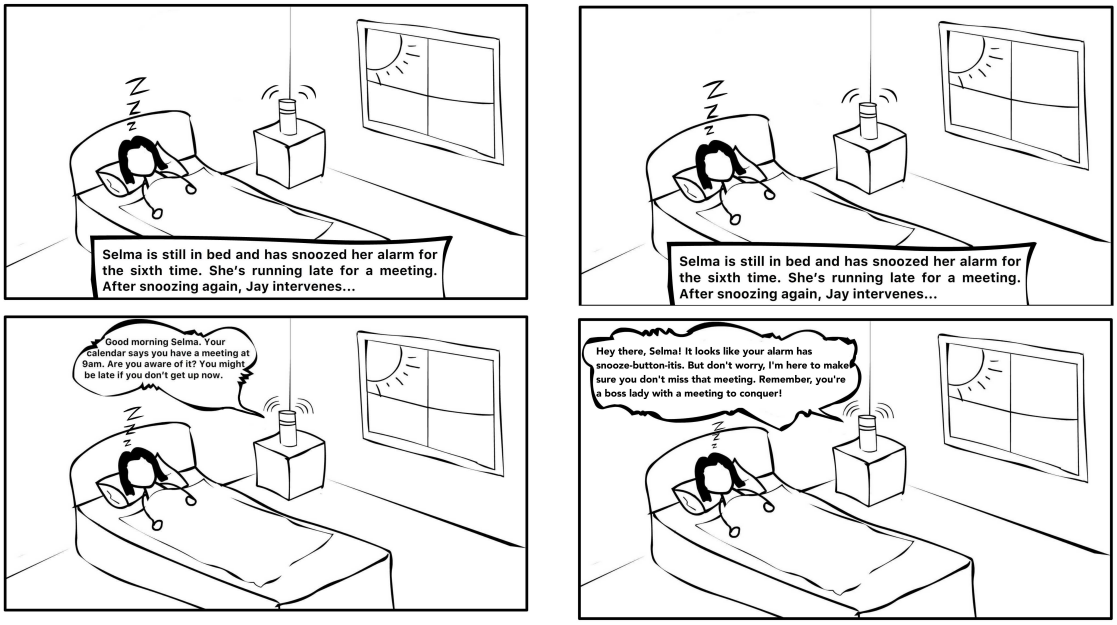


Figure 2: Both versions of the scenario *Meeting Reminder (S1)*. On the left, the neutral version, and on the right, the humorous version is shown. Both versions were evaluated in the survey.

nature. Thus, our aim was rather to produce a version for each scenario that would most likely be seen as *more* humorous than the initial/neutral version of the scenario *on average*. Regarding the type of humor, we exclusively utilized affiliative humor for the agent's comments. As discussed in the related work, prior studies indicate that employing this kind of humor could improve user satisfaction, in contrast to aggressive humor [77]. Initially, for each selected scenario, we generated ten humorous comments using ChatGPT [13]. These generated comments were reviewed by three authors, who assessed their humor and chose their favourite five comments for each scenario. The comments were further refined in an attempt to make them more humorous. Lastly, we presented the selected and refined five humorous comments along with their respective scenarios to a panel of four HCI researchers who were not involved in this project. Based on their feedback, one humorous comment was chosen for each scenario, with some of the selected comments undergoing further modification based on the group's input. This additional step of filtering and refinement by the panel was done with the aim to increase the likelihood that the interventions could be seen as humorous by a wider population in the following study.

Moreover, we included two additional scenarios from the final set of nine scenarios in our previous study [103]. However, unlike the other scenarios, we intentionally left the remark by the agent blank, allowing participants to come up with their own proactive VA interventions.

The scenarios were presented in the form of two-panel cartoon sketches. The design of the storyboards aimed to minimize cultural and ethnic cues to ensure participants could relate to the characters regardless of their backgrounds. To avoid any potential influence on participants' interpretation of the scenarios, the characters were intentionally designed without facial expressions. Consistent with the original storyboards, the VA in the sketches had a cylinder-shaped appearance resembling a conventional smart speaker. To reduce gender bias, the fictional agent was given the gender-ambiguous name "Jay". For the complete selection of storyboards used in the questionnaire, please refer to the Appendix.

Here is a brief description of each of the scenarios. Both a neutral and a humorous version of each scenario were used in the questionnaire:

- *S1 Meeting Reminder*: After the user has repeatedly "snoozed" the alarm, Jay reminds her of an upcoming meeting.
- *S2 Health Risk*: From the sound of the cough, Jay suspects an elderly user to have a respiratory infection and offers to arrange a doctor's appointment.
- *S3 Fact Checking*: Three friends are discussing a historical topic when Jay interrupts them to get a fact right.
- *S4 Disagreement Clarification*: Two people remember differently about what they agreed on when Jay settles the disagreement by quoting what they said.
- *S5 Nudging*: When the user asks Jay to play a TV series, Jay suggests stopping earlier than last night.

Moreover, here is a brief description of the scenarios where participants had to fill in the agent's proactive comment:

- *S6 Cooking Inspiration*: Two friends are deciding about dinner when Jay proactively intervenes.

- *S7 Technical Support*: A person asks their friend for help setting up new headphones, but the friend is busy cooking. Jay proactively intervenes.

3.2 Online Questionnaire

Participants' responses were collected through the online survey platform Qualtrics [1]. The questionnaire began with a welcome text and a brief introduction about the procedure and the research purpose. Participants were then informed about their rights and were required to provide informed consent before proceeding. Afterward, the concept of proactive VAs and the fictional agent "Jay" were introduced to the participants. The initial part of the questionnaire involved participants answering questions about their experience and usage of voice assistants, including their level of interest, enjoyment, and perceived usefulness of VAs. We also provided a clear definition of AI agents and asked participants to indicate their perception of these agents concerning social equality compared to themselves. This evaluation was conducted on a scale of −50 to 50, with 0 representing the agent's equality, −50 representing significant inferiority, and 50 representing significant superiority (to investigate *H2*). Additionally, we asked participants how humorous they would like a VA to be while additionally collecting data on participants' self-assessments of their own humor and how important they find humor in general.

In the subsequent part of the questionnaire, participants were presented with ten scenarios, consisting of five neutral and five humorous scenarios, in a randomized order. Participants were asked to rate each scenario in terms of *usefulness*, *appropriateness*, and *invasiveness*, as well as indicate the likelihood of the user in the storyboard considering the assistant's proposition (following, below: *consideration*). We will in the following generally refer to these variables as the *'four key dimensions'* related to the overall 'desirability' of the interventions (see RQs). Ratings were given using a seven-point Likert scale. Note that higher ratings reflect better perceptions for all four dimensions, including *invasiveness*, for which the scale was inverted to simplify the data analysis and presentation of results (hence, a rating of 1 refers to **most** invasive, and 7 to **least** invasive).

Before this section, participants were informed that, for the purposes of the study, they could assume the fictional agent (Jay) protects their personal data, processes it on the device, and does not share it with any third parties. By pointing this out, we intentionally aimed to alleviate participants' concerns primarily focused on data privacy, as this aspect has been extensively studied in existing research [55, 82]. Participants were encouraged to read the scenarios carefully, as they would be repeated, but the agent's comments would differ.

The next part of the questionnaire involved two scenarios where the agent's comment was left blank. Participants were asked to write their ideal statement for the agent in each scenario and provide their thoughts on the potential impact of a humorous comment from the agent in that context. Participants were then asked to rate the humor of the five humorous scenarios on a Likert scale ranging from 1 to 7 ("How humorous do you find Jay's interaction in this scenario?"). This rating aimed to subsequently examine how the key dimensions might get affected depending on how humorous participants found the VA's humorous interventions (to investigate *H1*). It was expected that there would be adverse effects on aspects covered by our four key dimensions, such as *appropriateness*, if the humor used should not perceived as humorous. Subsequently, participants were asked to share their thoughts on the type of humor used in the agent's interventions, including aspects they liked or disliked about the humor. They were also asked to indicate situations where the agent should or should not use humor. The questionnaire concluded with a set of demographic questions about participants' age, gender, nationality, country of residence, and fluency in English.

Prior to running the main study, a pilot study was conducted with two participants. The primary objectives of the pilot study were to identify any potential issues within the questionnaire and assess the scenarios' effectiveness in immersing participants and stimulating contemplation. Subsequently, minor adjustments were made to the questionnaire based on the feedback received, and the main study was conducted. On average, the questionnaire took approximately 20 minutes to complete ($M = 20.03$, $SD = 9.84$). The complete list of questions can be found in the supplementary material.

3.3 Participants

Participants were recruited using convenience sampling, which involved reaching out through mailing lists, social networks, internet forums, and word-of-mouth. Participation in the survey was voluntary and uncompensated. Initially, we obtained a total of 102 responses to our questionnaire. Out of these, 46 responses were excluded due to incompleteness. Furthermore, six participants were excluded from the analysis because their responses consistently lacked informative content, which indicated their unsuitability for our study. These exclusions were made based on their tendency to engage in straight-lining or consistently providing responses that were not pertinent to the questionnaire's content. The final sample consisted of $N = 50$ participants, with 22 identifying as male, 24 as female, three as non-binary, and one participant not specifying their gender. Our study encompassed participants from 16 distinct countries, with the majority residing in the UK (32%), followed by the US (22%), Canada (10%), Germany (8%), Netherlands (8%), and Switzerland (6%). The average age of participants was $M = 33.50$ ($SD = 0.707$). All participants were proficient in English. Among them, 17 have not previously used a voice assistant, while the remaining participants reported rarely (17), sometimes 8, and often (8) using them. 17 participants reported that they own a smart speaker. With regards to participants' self-assessed humor, the items covered *Humor Self* ("How humorous do you think you are?"), *Humor General* ("How important do you find humor in general?"), and *Humor Relationship* ("How important is humor for you in your relationships with other people?"), which participants rated with $Mdn = 5$ ($IQR = 1.25$), $Mdn = 6$ ($IQR = 2$), $Mdn = 6$ ($IQR = 2$), respectively. In addition to the self-assessments, participants rated how humorous they would like a VA to be with $Mdn = 4$ ($IQR = 3$) slightly lower than the previous items.

[1] https://www.qualtrics.com

Table 1: Medians and IQRs of the sums of participants' ratings of the four *key dimensions* for the scenarios without humor (baseline) and the scenarios with humor (intervention). On the right side of the table are the *Mann-Whitney U* test statistics; significant results with Bonferroni-corrected $\alpha = .013$ are marked with asterisks. Higher *median* values are 'better' - incl. *invasiveness*, hence higher values mean *less* invasive.

	Without Humor (Baseline)		With Humor (Intervention)		Wilcoxon Signed-Rank Test Statistics		
Dimension	*Mdn*	*IQR*	*Mdn*	*IQR*	*U Statistic*	*p − value*	*ES*
Usefulness	26.00	7.00	23.00	10.75	795	<.001*	0.680
Invasiveness	15.50	7.75	16.00	9.00	497	0.348	-0.156
Appropriateness	19.00	10.50	17.50	10.00	696	<.001*	0.617
Consideration	22.00	9.75	20.50	9.00	937	<.001*	0.593

3.4 Data Analysis

The questionnaire responses are analyzed and presented both quantitatively and qualitatively to provide a comprehensive understanding of the participants' views on humorous proactive interventions. These results offer insights into the diverse range of opinions expressed by the participants.

Based on visual inspection of our data and the Shapiro–Wilk statistic, we could not assume normally distributed data. Due to this, as well as the ordinal scale level of most of our items, we conducted non-parametric tests. We used *Spearman* correlations to explore relationships, *Wilcoxon Signed-Rank* tests to compare the difference between baseline and intervention data, and *Mann-Whitney U* tests to compare specific subgroups in our sample. We applied an alpha level of .05 for all our statistical tests.

The open-ended responses were systematically analyzed using a conventional content analysis approach [32]. The analysis began with data familiarisation [12], where two researchers read through all the responses to get a sense of the content and context to understand the patterns, ideas, and concepts present in the responses. Afterward, to develop a coding system, a subset of responses from 10 randomly selected participants were independently coded by two researchers using an inductive coding approach, where a single quote could be assigned to multiple codes, including descriptive (e.g., privacy concerns), conceptual (e.g., benefits of humorous responses), or emotional (e.g., frustration) codes. The researchers engaged in extensive discussions to reach a consensus and establish a coding system. In cases of disagreements, a third author was consulted to ensure agreement. Subsequently, an iterative discussion process between the two authors resulted in the creation of a codebook. One researcher coded the remaining responses individually, employing the established codebook. As the evaluation proceeded, some new codes emerged, requiring the codebook to be adjusted accordingly. This process resulted in extracting key insights and findings from the analyzed responses, presented in section 5.

4 QUANTITATIVE FINDINGS

In this section, we present the quantitative analysis of the questionnaire responses. Variable names are typically presented in *italics*. Descriptive statistics will be reported using *median (Mdn)* and *Interquartile Range (IQR)*. Exceptions are continuous variables like *Age*, for which we will utilize *Mean (M)* and *Standard Deviation (SD)*.

4.1 Perspectives on VAs

To provide a contextual backdrop to our findings, we asked a series of questions from participants regarding their experiences with and attitudes toward VAs. We measured participants' interest in VAs, their enjoyment while using them, and their perceived usefulness of these systems using a Likert scale ranging from 1 to 7. The participants' interest in VAs ($Mdn = 5$, $IQR = 4$), enjoyment of using them ($Mdn = 5$, $IQR = 3$), and perceived usefulness ($Mdn = 4$, $IQR = 2.25$) indicate a mostly balanced distribution of general perceptions about VAs among the participants. However, the relatively large IQRs also suggest diverse viewpoints within the sample.

Additionally, we inquired how participants perceived VAs from a 'social hierarchy' perspective ("What is your perception of AI agents in comparison to you? – They feel ... to me"). Respondents indicated their perception using a slider with the midpoint representing 'equality' (corresponding to a value of 0), the left end signifying 'highest inferiority' (corresponding to a value of −50), and the right end representing 'highest superiority' (corresponding to a value of 50) in relation to themselves. The $Mdn = -20$ ($IQR = 32.50$) indicates a rating between inferior and equal, slightly 'leaning towards' equal. The following sections will refer to this variable as *Social Equality*.

4.2 Comparing the Baseline with the Humorous Scenarios

The scenarios with humor were rated lower than scenarios without humor for *usefulness*, *appropriateness*, and *consideration* – this difference was found to be significant with a *Wilcoxon Signed-Rank* test (see Table 1 for corresponding descriptive and inference statistics). The only dimension that tended to have higher ratings for the scenarios with humor was *invasiveness*; however, the difference was not significant. This suggests that, overall, the humor used by the VA – in the given scenarios – does not seem to affect the four key dimensions positively. The lower ratings of the scenarios with humor could be due to participants not finding the humor used in the scenarios humorous. The humor in the humorous scenarios was rated with $Mdn = 2.5$ for scenario 1 ('Meeting Reminder') ($IQR = 4$), and all the other scenarios were rated with $Mdn = 3$ and IQRs ranging from 2.5 to 4 (see Table 3 in Appendix A for descriptive statistics for all scenarios). Overall, this suggests that most participants did not find the scenarios with humor that humorous. However, the high spread (i.e., IQRs) underlines that there are marked individual

Table 2: Medians and IQRs of the *rating deltas* between the scenarios without (baseline) and with humor (intervention) – grouped by participants who found the scenarios more humorous (left side) versus those who found them less humorous (middle). On the right side are the *Mann-Whitney U* test statistics; significant results with Bonferroni-corrected $\alpha = .013$ are marked with asterisks. Higher *median* values are 'better' - incl. *invasiveness*, hence higher values mean *less* invasive.

	Above Average Humor Rating		Below Average Humor Rating		Mann-Whitney U Test Statistics		
Dimension	*Mdn*	*IQR*	*Mdn*	*IQR*	*W Statistic*	$p-value$	*ES*
Usefulness	0	5	-6	9	149.0	.002*	0.532
Invasiveness	2	4	-1	5	173.0	.007*	0.446
Appropriateness	0	3	-4	6	240.5	.163	0.230
Consideration	1	4	-6	5	89.5	<.001*	0.714

differences between participants and that they have perceived the humor in the scenarios very differently. This leads to the question of what effects humor might have had on the four key dimensions for participants who found the scenarios **more** humorous compared to those who found them **less** humorous. In other words, in case the VA's intervention is found to be humorous, could this positively affect how *invasive* the intervention is perceived? We will explore this question in the following subsection by investigating how the key dimensions might be affected depending on the participants' humor ratings.

4.3 Effects of Humor When it is Considered Humorous

This section explores how participants' baseline and humorous scenario ratings (for the four key dimensions) differ depending on how humorous they found the latter. To explore this, the sample was split into two halves (ex post) based on their overall ratings (using the sum of humor ratings of all five scenarios for each participant). One group was defined containing all participants above the median ($Mdn = 16.5$) of the humor rating sums (which we will refer to as *Higher Humor Rating Group*, $n = 25$) and the other group below the median (*Lower Humor Rating Group*, $n = 25$). Using the median instead of the scale's midpoint ensured that both sub-samples were equally sized. However, it is important to stress that the upper half does not exclusively comprise participants who found the scenarios humorous overall. This is due to the median ratings being positioned below the midpoint of the 7-point Likert scale (see also Table 3 in the Appendix A).

When inspecting Table 2, it can be seen that the *Lower Humor Rating Group* (who found the scenarios with humor less humorous) consistently rated them worse across all *four key dimensions* than the scenarios without humor – thus presenting a similar picture as in the previous section (subsection 4.2) but with the negative effects being even more pronounced. However, a different picture emerges when considering the *Higher Humor Rating Group*, where there seemed to be no adverse effects on the *four key dimensions* (with rating deltas ranging between 0 to 2) and for *invasiveness* and *consideration* there even seemed to be positive effects (see also Figure 3).

Taken together, the deltas thus were all < 0 for the participants who found the scenarios **less** humorous and >= 0 for participants who found the scenarios **more** humorous. To investigate if the differences between the two groups are significant, a *Mann-Whitney U* test was conducted for each of the four key dimensions, which was significant for *usefulness* ($p = .002$), *invasiveness* ($p = .007$), and *consideration* ($p = < .001$), but not significant for *appropriateness* ($p = .163$).

Given this significant difference in the *invasiveness* and *consideration* rating deltas and since the deltas were positive for the *Higher Humor Rating Group*, an exploratory analysis was conducted to examine if the increase from baseline to intervention is significant for these two dimensions when only considering this group. For *invasiveness* the difference is indeed significant (Wilcoxon Signed-Rank, $W = 47$, $p = .002$, $Effectsize = -0.687$) while for *consideration* it is not ($W = 144$, $p = .579$, $Effectsize = -0.044$).

4.4 Perceived Social Equality of VA

We expected that participants would find the scenarios with humor more humorous if they perceive the VA more *socially equal* to them. Indeed, there seems to be a significant correlation ($p = .043$, $r_{Spearman} = 0.288$). This is further corroborated when examining the *VA Social Equality* ratings of participants who stated that they preferred the scenarios with humor over those without humor ("In general, did you prefer the humorous interactions over the non-humorous ones?"). A marked difference can be observed in participants' *VA Social Equality* ratings for those who prefer the scenarios with humor $Mdn = 47.5$ ($IQR = 32.5$) compared to those who prefer those without $Mdn = 25$ ($IQR = 21.3$), see also Figure 4. This difference was found to be significant using a *Mann-Whitney U* test with $U = 146$, $p = .004$, and $Effect\ size = 0.493$. This suggests that the more people see VAs at a similar social level to themselves, the more they are open to the VA using humor.

5 QUALITATIVE FINDINGS

Within our sample of 50 participants, nearly half of them (23) expressed their dislike for the style of humor used in the scenarios. They perceived the humor as inappropriate, forced, lacking personal connection, and bothersome. For instance, one participant remarked: "None of the characters in the scenarios were joking around with their friends. I would find the comments irritating if an actual human had made them. Not only is it irritating, but it makes it far less clear what the AI is actually saying or offering to do." (P14). On the other hand, 14 participants embraced the humor,

Figure 3: Boxplots of baseline to intervention rating deltas for *invasiveness* and *consideration* grouped by participants below and above the average scenario humor rating, showing median, IQR, and maximum and minimum values (with three outliers represented as dots).

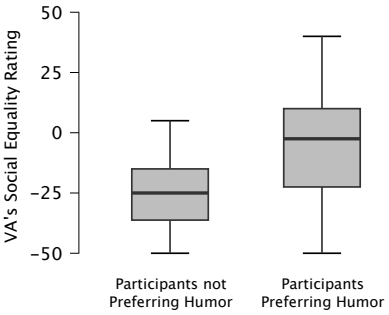


Figure 4: Boxplot of *VA Social Equality* ratings grouped by participants who preferred scenarios with/without humor, showing median, IQR, and max/min.

finding it both enjoyable and intriguing. One participant articulated: "[The humor] makes the intervention more natural." (P36). Seven participants underscored the subjective essence of humor, acknowledging the challenge of crafting humorous comments for voice assistants.

Moreover, four participants expressed concerns regarding continuous monitoring of the auditory environment by the agent. One participant stated: "It raises security concerns about the constant surveillance of household audio." (P43). These concerns were raised even though participants were explicitly requested to temporarily set aside privacy and data protection considerations during the survey.

5.1 Humor Ranking

Participants indicated their favorite humorous scenario and provided the rationale behind their choice (see Figure 5).

The *Meeting Reminder* emerged as the favorite among 12 (24%) of participants. Participants found humor in this scenario to be encouraging, a blend of entertainment and utility, as well as inspiring and motivational. One participant mentioned: "Calling the user a 'boss' is a colloquial and personable interaction that does not feel forced and is motivating. It is how a friend would speak to you." (P43).

As for the *Fact Checking* and *Nudging* scenarios, 11 (22%) participants favored them. In the *Fact Checking* scenario, the humorous agent intervention was perceived as 'funny yet factual', with enjoyment derived from a historical reference, and an opinion that it alleviates tension. One participant mentioned: "It is humorously delivering the fact while not making the situation unnecessarily awkward." (P19). In the *Nudging* scenario, participants viewed the comment as both humorous and effective, suitable as it aligned with the entertainment context (two), and not detracting attention from the issue (one): "It brings humor without undermining the seriousness of the matter." (P5).

Seven participants favored *Disagreement Clarification*, mainly citing its tension-relieving aspect (three). One participant pointed out:

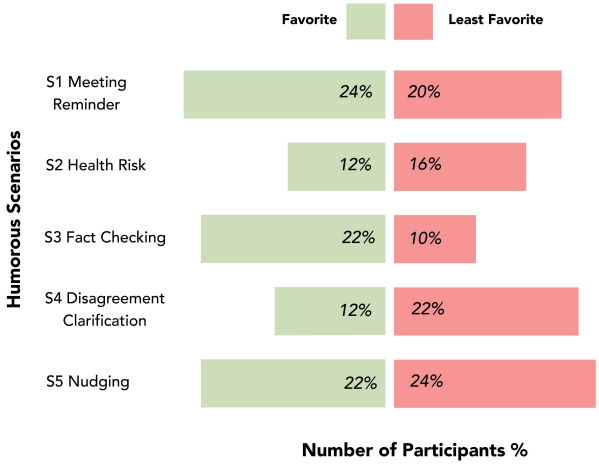


Figure 5: Bar charts displaying the percentage of participants selecting a humorous scenario as their favorite and least favorite for each scenario.

"It breaks up an awkward/tense scenario, and the subject matter is pretty light and inconsequential." (P33). The *Health Risk* scenario was favored by six participants, primarily as it was perceived as unintrusive: "It is a touch of humor without sounding condescending or juvenile." (P30). Three participants did not select a favorite scenario.

In terms of participants' least favorite humorous scenario and the reasoning behind it, the *Nudging* scenario garnered the highest number of votes for being the least favored by 12 people. Participants expressed concerns about the agent's remark in this scenario being impolite, inappropriate, and overly lengthy. One participant pointed out: "It feels a bit weird that a piece of technology would be questioning what I'm doing." (P11).

The *Disagreement Clarification* scenario was chosen as least favorite for 11 respondents, due to perceptions of the interaction as intrusive, impolite, and pedantic. A participant highlighted: "[The agent] is negative towards one person." (P42). Another participant noted that "the agent's involvement felt intrusive in a personal relationship."

Meeting Reminder was selected as the least favorite by ten participants. Two found the agent's behavior insensitive to context, two others thought it is exerting pressure on the user to be productive, and three participants specifically disliked the employed humor. One participant commented: "It comes across as a company trying far too hard." (P25).

Eight chose *Health Risk* as their least favorite. Participants thought that the humor employed could detract from the gravity of the health concern. Five participants highlighted that humor is inappropriate when dealing with health matters. One participant stated: "It's not good to add that level of humor into matters related to people's health, especially when coming from an AI." (P33).

Finally, *Fact Checking* was chosen by five respondents as the least favorite, primarily due to the comment being perceived as intrusive: "Nobody asked Jay's opinion. Maybe it could light up to show it has something to contribute." (P10). Two participants disliked the humor employed in this scenario. Four participants did not select a favorite scenario.

5.2 What Would the Agent Say?

Participants filled in their ideal agent's comment for the two scenarios of *Cooking Inspiration* and *Technical Support*.

Regarding *Cooking Inspiration*, a significant majority (40 out of 50) offered supportive comments without incorporating humor, by suggesting food ideas or facilitating online food ordering. Three participants expressed the belief that the agent should not engage in such a context. Only five participants chose to introduce humor into their agent's comment. They either used a humorous food recommendation, or a humorous comment followed by supportive guidance: "Not this again, I can't remember the last time you knew what to eat. Luckily, I can help you – how about some spaghetti Bolognese?" (P28). Two left this question unanswered. Regarding the impact of humor, within this scenario, 16 people thought that incorporating humor would not have any influence on the situation. In contrast, 15 participants believed that humor might exacerbate the situation, associating it with potential annoyances, distractions, time wastage, diminished seriousness, or elements of irritation and condescension. One participant mentioned: "It would sound more invasive and less like a service." (P16). On the contrary, eight participants thought that humor could enhance the situation, being seen as 'encouraging' or 'inspiring' for users. One participant said: "I think humor would make it more light-hearted and pleasant." (P8). 11 participants did not offer a response to this aspect.

Similarly, regarding *Technical Support*, the majority (32) offered a supportive comment without incorporating humor, three believed the agent should not engage, and five left this question unanswered. For this scenario, ten participants used humor in their comment. Such comments were either a humorous statement, or a humorous statement followed by supportive guidance: "Sandra, let me be the chivalrous one here and help with the headphones." (P13). About the impact of humor on the situation for this scenario, 18 found the use of humor to be beneficial, as it could lighten the mood, help release tension, and make the interaction more appropriate. One participant stated: "It diffuses a potentially tense situation by lightening the mood" (P8). 16 people thought it will make the situation worse as it can be annoying, distracting, inappropriate, or it can decrease the seriousness of the situation. One participant said: "It would be inappropriate to joke as everyone is busy." (P3) Eight participants believed humor would not have any impact on the situation, and ten did not respond to this question.

5.3 Opportune Time for Humor

When considering appropriate times for an agent to employ humor, 15 participants indicated that it should be employed during non-serious and playful instances, such as when people are in a playful mood or laughter is detected. Six people mentioned that humor should be utilized exclusively when explicitly requested by the user. Additionally, three suggested its usage when users are in the company of close friends or family members. Three proposed its application during moments of perceived tension to alleviate stress. Two participants recommended a consistent humorous approach, while one participant suggested leveraging humor to motivate people toward healthier behaviors.

On the contrary, four people expressed a preference for the agent to refrain from using humor altogether. Regarding contexts where participants felt humor should be avoided, half of the people (25) noted that the agent should abstain from using humor during discussions of serious topics such as health, work, or finances. An additional five participants emphasized that humor should not be used during time-critical situations, while another five highlighted that humor should be avoided in socially tense situations.

6 DISCUSSION

Our exploratory investigation delved into incorporating humor in proactive VA statements within a home environment.

We interpreted the results of this evaluation to provide answers to the following comprehensive questions:

RQ1: Can humor increase the desirability of proactive interventions of VAs?

RQ2: In which situations and context is humor perceived as appropriate?

6.1 Impact of Humor Reception on Desirability

Overall, regarding **RQ1**, our questionnaire results demonstrated that the humor used in our scenarios did not affect aspects of *usefulness*, *appropriateness*, *invasiveness*, and *consideration* positively.

However, there were marked differences between the participants. Some found the humor heart-warming and pleasant, while others considered it distracting and inappropriate. This once again underscores the inherent subjectivity of humor [10, 99, 101]. However, we observed that certain factors impacted the desirability of humorous proactive comments by VAs, which we discuss here.

We witnessed that around half of our participants did not like the humor used in our scenarios. For this subset of participants, the inclusion of humor predominantly resulted in a negative influence on the proactive interventions made by the VA in terms of *usefulness*,

appropriateness, *invasiveness*, and *consideration*. On the contrary, another subgroup of participants generally enjoyed the humor incorporated into the scenarios. For this category, humor within our scenarios generally positively impacted the VA's proactive interventions concerning *invasiveness*. Taken together, these findings suggest that when the humor used fails to resonate with users, it is likely to adversely affect the user's perception of the VA's proactive statement. This is in line with prior research indicating that humor carries inherent risks, and if a humorous attempt falls short, it can lead to worse outcomes [8, 9, 39]. Conversely, if the humor used is indeed perceived as humorous by the user, it has the potential to mitigate the *invasiveness* of the comment. In such cases, humor can act as a buffer, making people more receptive to proactive interventions. This aligns with existing literature on humor, suggesting that humor can enhance the reception of information [27, 45, 62, 74]. In effect, we can accept our first hypothesis:

H1: The desirability of a proactive intervention is affected depending on how humorous it is perceived.

6.2 Perceived Social Equality and Humorous Interventions

We witnessed that the effectiveness of delivering a humorous intervention can be heightened when the user perceives the VA as a more socially equal partner. Our evaluation highlighted that participants who viewed the VA as more socially equal tended to rate humorous scenarios as funnier than those who perceived the VA as inferior. Moreover, we observed that the more participants saw VAs at a similar social level to themselves, the more they were open to the use of humor by VAs. These findings are consistent with existing literature, emphasizing that the perceived characteristics of the individual delivering humor impact its reception [8, 97], particularly evident concerning the social status and perceived authority of the individual delivering humor [68]. In line with these insights, we can then confirm our second hypothesis:

H2: A correlation exists between how people perceive a VA regarding its social equality and how humorous they find its interventions.

The implications of these findings suggest that VAs should tailor their use of humor based on the user's perception of their relationship with the VA. This perception could be gathered through user self-reports during VA setup or configuration. Additionally, VAs could adjust their application of humor based on the given context. This involves determining whether the VA should function predominantly as an assistant for task-oriented assistance or as a 'colleague' aimed at motivating and inspiring the user. These distinct roles could imply different hierarchies and user expectations concerning the 'social' interaction and its perceived 'hierarchy.'

6.3 Timing Humorous VA Statements

Regarding **RQ2**, our qualitative assessment showed that participants expressed the belief that VAs should refrain from using humor during discussions or activities related to serious topics such as health, work, or finances. Additionally, participants emphasized the importance of avoiding humor in time-critical and socially tense situations. These findings underscore the significance of timing and context in deploying humor. The least favored scenarios further shed light on this matter. Participants expressed disapproval when humor was not carefully contextualized, leading to perceptions of impoliteness and inappropriateness. Moreover, people raised concerns when using humor in contexts involving sensitive or serious topics, worrying that it might undermine the gravity of the subject matter. Although humorous content can be attention-grabbing and entertaining, it might also convey that a situation is not serious [54]. Humor could potentially lead to a reduced inclination to address a problem due to its association with non-serious contexts. This was also the case in our findings, where *usefulness* was generally rated lower in the scenarios with humor even though the type of help or suggestion was not different and thus the 'objective usefulness' technically being the same.

Participants preferred humor during light-hearted and playful occasions. They suggested that humor could be appropriately employed when cues like laughter or humorous conversations are detected, signaling an opportune moment for the VA to engage in humor. Another factor was regarding the people's relationship, proposing using humor when people are with close friends or family members. The favored scenarios shed further light on this aspect. Participants preferred humorous VA comments that strike a balance between entertainment and utility, fostering a motivating and encouraging atmosphere. Additionally, we observed a potential for using humor to alleviate tension and enhance user experience, particularly when combined with factual information and contextual relevance, to create relatable and positive interaction dynamics.

Participants generally disliked the use of humor in the *Disagreement Clarification* scenario due to its perceived tense social context and in the *Health Risk* scenario due to the potential seriousness of the health concern. In contrast, regarding *Fact Checking*, participants seemed to find the context suitable and the topic less serious, resulting in a more favorable reception of humor. In the case of *Meeting Reminder* and *Nudging*, opinions were rahter mixed regarding the appropriateness of humor.

Nevertheless, some participants favored a reserved approach, desiring the agent to deploy humor only upon specific requests. Conversely, a group endorsed a consistent use of humor, valuing a consistent presence of humor in interactions. As for the potential impact of humor concerning the *fill-in-the-blank* scenarios, participants displayed a range of opinions. Some found it beneficial, some perceived it as having no influence, and others believed it could worsen the situation. This variation underscores the subjective nature of humor's effects and its nuanced reception across different individuals and contexts. Participants' diverse viewpoints highlight the intricate nature of deploying appropriate humor. The findings underscore the complexity of humor and the necessity of factoring in user preferences, context, and potential impacts when incorporating humor into VA interactions, as mentioned in previous research [101].

The significant preference for supportive comments without humor regarding both *fill-in-the-blank* scenarios further suggests that participants value straightforward and pragmatic interactions. Even when humor was used, it often accompanied supportive guidance, revealing a desire for practical assistance alongside any attempt at humor.

An important observation from the qualitative evaluation of humorous scenarios was that participants directed their attention mainly toward the proactive intervention itself and its timing, overseeing the humorous aspect of the agent's comment. This highlights that the novel interaction introduced by the agents' proactive statements took precedence, often overshadowing the humor intended. Such a pattern of responses could imply that when participants favored a humorous approach, the success could be attributed to the fitting and appropriate timing of the proactive intervention 'itself'. This observation suggests that the timing of proactive interventions may align with suitable moments for incorporating humor.

6.4 Humor and Proactive VA Desirability

Our findings highlight the intricacies of integrating humor into proactive voice assistant interactions. If humor fails to resonate with the user, it can have counterproductive consequences, especially in the context of proactive VA interventions. It became evident that humor is not a mere supplementary aspect or interactional feature that can be casually incorporated. However, it could enhance the interaction if it resonates with the user. To this end, we recommend tailoring humor to individual user preferences and sensitivities. This approach acknowledges the diverse reactions that humor can elicit among people. For designers and developers of VAs, understanding that humor can have varying effects on users is crucial. Therefore, investing in implementing personalized humor that resonates with users' unique perspectives is a worthwhile consideration. Overall, designers should consider humor as a potential strategy to soften the impact of proactive interventions. However, if humor cannot be achieved and tailored to individuals, alternative approaches might be more effective in achieving desirable outcomes.

7 LIMITATIONS AND FUTURE WORK

Our research has certain limitations that require acknowledgment. Firstly, even though our study had a heterogeneous sample with varying ages and backgrounds, the findings should be interpreted within the specific group studied. Our 50 participants resided in 16 different countries. While our sample included individuals with various cultural backgrounds, it is important to note that the sample size remains relatively small and might not offer a fully representative picture. Prior literature has underscored the influence of cultural background on humor interpretation [10, 99]. To enhance the robustness of our findings, future research should extend its investigation to broader and more varied populations. Furthermore, as we addressed earlier, humor perception is inherently subjective [10, 101]. Enhancing the desirability of humorous agents' comments requires a deeper comprehension of users' individual preferences, personalities, and cultural influences, as well as one's individual 'history' with an agent. Subsequent studies could delve into crafting personalized humorous remarks aligned with each user's humor taste. In this work, we employed a three-step approach in an attempt to produce scenarios that would, on average, be perceived as more humorous than the baseline. Our results indicated that a significant portion of our participants did indeed perceive the scenarios as humorous. However, another subset of our participants did not share the same perception about the humor level in the humorous scenarios. It is essential to acknowledge that, due to the inherent subjectivity of humor, it is not possible to ensure that all participants will find all the scenarios humorous. Nevertheless, this was not a major issue for our study design, which accounted for some differences in humor perceptions.

Our study explored humorous proactive VA comments within a home environment, as it is one of the most common use cases for VAs. While the broader insights from this study may have applicability in other settings, future research should delve into these VA remarks within different contexts, such as workplaces and public spaces. Moreover, based on recommendations from previous literature, we only employed affiliative humor for the agent's humorous comments, as this form of humor has been shown to enhance service satisfaction as opposed to aggressive humor [77]. In future studies, other types of humor should also be examined to understand their impact on user experience. Humor is a phenomenon greatly influenced by context and timing. In our approach, we made an effort to incorporate context and timing within our storyboards to a certain extent. However, storyboards cannot convey the exact turn-taking, timing, and delivery in a given (social) context that might play a role in how (humorous) an intervention is perceived. Thus, future studies exploring humor for VAs may consider alternative methods that can more effectively capture and utilize these crucial elements.

Lastly, it is important to acknowledge the limitations of our chosen method in this study. We gathered people's opinions based on hypothetical scenarios, as many of the capabilities depicted in our storyboards are not currently present in consumer products. This approach enabled participants to engage in speculation about interactions with future technologies that might be complex or costly to develop. Nonetheless, we must acknowledge that participants did not directly experience these situations, and their perceptions might not fully align with real-world experiences.

8 CONCLUSION

In this study, we explored the utilization of humor in proactive voice assistants and its influence on the desirability of such interactions. We conducted an online questionnaire with 50 participants, employing a scenario-based method. Participants were presented with storyboards illustrating instances where a proactive smart speaker engaged with people in various everyday situations, utilizing humorous and non-humorous remarks. Our results reveal that, while humor did not uniformly enhance aspects of *usefulness*, *appropriateness*, *invasiveness*, and *consideration*, there were clear distinctions in participants' reactions, highlighting the subjectivity of humor. We witnessed that humor's effects on the desirability of an intervention depend on whether people perceive it as humorous or not. Additionally, the success of humorous interventions can be enhanced when people perceive the VA as more socially equal. We recommend personalized humor tailored to individual user preferences and sensitivities to address these diverse responses. Humor is a multifaceted tool, with its effects contingent on individual preferences, context, and perceptions. Our findings caution against the casual incorporation of humor. Instead, humor should be applied thoughtfully or avoided altogether, as misaligned humor can backfire, particularly within the context of proactive VA interventions. Recognizing the significant role humor has historically played in

human social interactions and relationships from the origins of society, we contend that a proper understanding and exploration of humor in the context of human-computer interaction should be encouraged in both research and practical endeavors within this domain [58, 101].

ACKNOWLEDGMENTS

We would like to sincerely thank Paweł W. Woźniak for supporting us with parts of the data analysis. We also thank the anonymous reviewers whose suggestions helped improve and clarify this work. This work was partially funded by the Klaus Tschira Foundation, by the Leverhulme Trust (award DS-2017-026), by the FET-Open Project 951846 "MUHAI – Meaning and Understanding for Human-centric AI" funded by the EU program Horizon 2020, as well as the German Research Foundation DFG as part of Collaborative Research Center (Sonderforschungsbereich) 1320 "EASE – Everyday Activity Science and Engineering", University of Bremen (http://www.ease-crc.org/).

REFERENCES

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[2] Cameron Anderson, John Angus D Hildreth, and Laura Howland. 2015. Is the desire for status a fundamental human motive? A review of the empirical literature. *Psychological bulletin* 141, 3 (2015), 574.

[3] Kika Arias, Sooyeon Jeong, Hae Won Park, and Cynthia Breazeal. 2020. Toward Designing User-centered Idle Behaviors for Social Robots in the Home.

[4] Daisuke Asai, Jarrod Orszulak, Richard Myrick, Chaiwoo Lee, Joseph F Coughlin, and Olivier L De Weck. 2011. Context-aware reminder system to support medication compliance. In *2011 IEEE international conference on systems, man, and cybernetics*. IEEE, New York, USA, 3213–3218.

[5] Oswald Barral, Ilkka Kosunen, and Giulio Jacucci. 2017. No Need to Laugh Out Loud: Predicting Humor Appraisal of Comic Strips Based on Physiological Signals in a Realistic Environment. *ACM Trans. Comput.-Hum. Interact.* 24, 6, Article 40 (Dec. 2017), 29 pages. https://doi.org/10.1145/3157730

[6] Nancy Bell and Anne Pomerantz. 2014. Reconsidering language teaching through a focus on humor. *EuroAmerican Journal of Applied Linguistics and Languages* 1, 1 (2014), 31–47.

[7] André Berton, Dirk Bühler, and Wolfgang Minker. 2006. SmartKom-Mobile Car: User Interaction with Mobile Services in a Car Environment. In *SmartKom: Foundations of Multimodal Dialogue Systems*, Wolfgang Wahlster (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 523–537. https://doi.org/10.1007/3-540-36678-4_33

[8] T Bradford Bitterly. 2022. Humor and power. *Current Opinion in Psychology* 43 (2022), 125–128.

[9] T Bradford Bitterly, Alison Wood Brooks, and Maurice E Schweitzer. 2017. Risky business: When humor increases and decreases status. *Journal of personality and social psychology* 112, 3 (2017), 431.

[10] Pavel Braslavski, Vladislav Blinov, Valeria Bolotova, and Katya Pertsova. 2018. How to Evaluate Humorous Response Generation, Seriously?. In *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval* (New Brunswick, NJ, USA) *(CHIIR '18)*. Association for Computing Machinery, New York, NY, USA, 225–228. https://doi.org/10.1145/3176349.3176879

[11] Michael Braun, Anja Mainz, Ronee Chadowitz, Bastian Pfleging, and Florian Alt. 2019. At your service: Designing voice assistant personalities to improve automotive user interfaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–11.

[12] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.

[13] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[14] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712 [cs.CL]

[15] John M. Carroll. 1999. Five Reasons for Scenario-Based Design. In *Proceedings of the Thirty-Second Annual Hawaii International Conference on System Sciences-Volume 3 - Volume 3 (HICSS '99)*. IEEE Computer Society, USA, 3051.

[16] Narae Cha, Auk Kim, Cheul Young Park, Soowon Kang, Mingyu Park, Jae-Gil Lee, Sangsu Lee, and Uichin Lee. 2020. Hello There! Is Now a Good Time to Talk? Opportune Moments for Proactive Interactions with Smart Speakers. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 74 (Sept. 2020), 28 pages. https://doi.org/10.1145/3411810

[17] William Curran, Gary J McKeown, Magdalena Rychlowska, Elisabeth André, Johannes Wagner, and Florian Lingenfelser. 2017. Social Context Disambiguates the Interpretation of Laughter. *Frontiers in psychology* 8 (2017), 2342. https://doi.org/10.3389/fpsyg.2017.02342

[18] Claire Dormann and Robert Biddle. 2006. Humour in game-based learning. *Learning, Media and Technology* 31, 4 (2006), 411–424.

[19] Philip R. Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R. Cowan. 2019. Mapping Perceptions of Humanness in Intelligent Personal Assistant Interaction. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services* (Taipei, Taiwan) *(MobileHCI '19)*. Association for Computing Machinery, New York, NY, USA, Article 5, 12 pages. https://doi.org/10.1145/3338286.3340116

[20] Mateusz Dubiel, Kerstin Bongard-Blanchy, Luis A. Leiva, and Anastasia Sergeeva. 2023. Are You Sure You Want to Order That? On Appropriateness of Voice-Only Proactive Feedback Strategies. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (Eindhoven, Netherlands) *(CUI '23)*. Association for Computing Machinery, New York, NY, USA, Article 43, 6 pages. https://doi.org/10.1145/3571884.3604312

[21] W Jack Duncan. 1982. Humor in management: Prospects for administrative practice and research. *Academy of management review* 7, 1 (1982), 136–142.

[22] Pawel Dybala, Michal Ptaszynski, Rafal Rzepka, and Kenji Araki. 2009. Humoroids: conversational agents that induce positive emotions with humor. In *AAMAS'09 Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems*, Vol. 2. ACM, ACM, New York, NY, USA, 1171–1172.

[23] Justin Edwards, Christian Janssen, Sandy Gould, and Benjamin R. Cowan. 2021. Eliciting Spoken Interruptions to Inform Proactive Speech Agent Design. In *CUI 2021 - 3rd Conference on Conversational User Interfaces* (Bilbao (online), Spain) *(CUI '21)*. Association for Computing Machinery, New York, NY, USA, Article 23, 12 pages. https://doi.org/10.1145/3469595.3469618

[24] Jonathan B Evans, Jerel E Slaughter, Aleksander PJ Ellis, and Jessi M Rivin. 2019. Gender and the evaluation of humor at work. *Journal of Applied Psychology* 104, 8 (2019), 1077.

[25] Jasper Feine, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2019. A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies* 132 (2019), 138–161.

[26] Kennan Ferguson. 2019. Comedy and critical thought: Laughter as resistance: Krista Bonello, Rutter Giappone, Fred Francis, and Iain MacKenzie (Eds.) London: Rowman and Littlefield International, 2018, x+ 237pp., ISBN: 978-1786604071.

[27] Gabriella Gandino, M Vesco, S Ramella Benna, M Prastaro, et al. 2010. Whiplash for the Mind. Humour in Therapeutic Conversation. *International Journal of Psychotherapy* 14 (2010), 13–24.

[28] Xiang Ge, Dan Li, Daisong Guan, Shihui Xu, Yanyan Sun, and Moli Zhou. 2019. Do smart speakers respond to their errors properly? A study on human-computer dialogue strategy. In *Design, User Experience, and Usability. User Experience in Advanced Technological Environments: 8th International Conference, DUXU 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part II 21*. Springer, Springer, Cham, 440–455.

[29] Eun Go and S Shyam Sundar. 2019. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior* 97 (2019), 304–316.

[30] Talip Gonulal. 2021. Investigating EFL learners' humorous interactions with an intelligent personal assistant. *Interactive Learning Environments* 0, 0 (2021), 1–14. https://doi.org/10.1080/10494820.2021.1974489 arXiv:https://doi.org/10.1080/10494820.2021.1974489

[31] Christian F Hempelmann. 2008. Computational humor: Beyond the pun? *The Primer of Humor Research. Humor Research* 8 (2008), 333–360.

[32] Hsiu-Fang Hsieh and Sarah E Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative health research* 15, 9 (2005), 1277–1288.

[33] Scott Hudson, James Fogarty, Christopher Atkeson, Daniel Avrahami, Jodi Forlizzi, Sara Kiesler, Johnny Lee, and Jie Yang. 2003. Predicting Human Interruptibility with Sensors: A Wizard of Oz Feasibility Study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 257–264. https://doi.org/10.1145/642611.642657

[34] Kristiina Jokinen and Michael McTear. 2009. Spoken dialogue systems. *Synthesis Lectures on Human Language Technologies* 2, 1 (2009), 1–151.

[35] A Kelly. 2017. Siri, tell me a joke. No, not that one. Could machine learning help the voice-activated assistant find its comedic chops. *Signal* 71, 7 (2017), 11–12.

[36] Chloé Kiddon and Yuriy Brun. 2011. That's What She Said: Double Entendre Identification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2* (Portland, Oregon) *(HLT '11)*. Association for Computational Linguistics, USA, 89–94.

[37] Rafal Kocielnik, Daniel Avrahami, Jennifer Marlow, Di Lu, and Gary Hsieh. 2018. Designing for Workplace Reflection: A Chat and Voice-Based Conversational Agent. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 881–894. https://doi.org/10.1145/3196709.3196784

[38] Mitsuki Komori, Yuichiro Fujimoto, Jianfeng Xu, Kazuyuki Tasaka, Hiromasa Yanagihara, and Kinya Fujita. 2019. Experimental Study on Estimation of Opportune Moments for Proactive Voice Information Service Based on Activity Transition for People Living Alone. In *Human-Computer Interaction. Perspectives on Design*, Masaaki Kurosu (Ed.). Springer International Publishing, Cham, 527–539.

[39] Dejun Tony Kong, Cecily D Cooper, and John J Sosik. 2019. The state of research on leader humor. *Organizational psychology review* 9, 1 (2019), 3–40.

[40] Matthias Kraus, Marvin Schiller, Gregor Behnke, Pascal Bercher, Michael Dorna, Michael Dambier, Birte Glimm, Susanne Biundo, and Wolfgang Minker. 2020. "Was That Successful?" On Integrating Proactive Meta-Dialogue in a DIY-Assistant Using Multimodal Cues. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20)*. Association for Computing Machinery, New York, NY, USA, 585–594. https://doi.org/10.1145/3382507.3418818

[41] Tracy Kubert and Elena Korshakova. 2020. Identifying IPA humor styles. *Proceedings of the Association for Information Science and Technology* 57, 1 (2020), e411.

[42] NICHOLAS A. KUIPER and CATHERINE LEITE. 2010. Personality impressions associated with four distinct humor styles. *Scandinavian Journal of Psychology* 51, 2 (2010), 115–122. https://doi.org/10.1111/j.1467-9450.2009.00734.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9450.2009.00734.x

[43] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, Are You Listening? Privacy Perceptions, Concerns and Privacy-Seeking Behaviors with Smart Speakers. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 102 (Nov. 2018), 31 pages. https://doi.org/10.1145/3274371

[44] Uichin Lee, Kyungsik Han, Hyunsung Cho, Kyong Mee Chung, Hwajung Hong, Sung Ju Lee, Youngtae Noh, Sooyoung Park, and John M. Carroll. 2019. Intelligent positive computing with mobile, wearable, and IoT devices: Literature review and research directions. *Ad Hoc Networks* 83 (Feb. 2019), 8–24. https://doi.org/10.1016/j.adhoc.2018.08.021 Funding Information: This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2017M3C4A7083529). Publisher Copyright: © 2018 Elsevier B.V..

[45] Richard G Lomax and Seyed A Moosavi. 2002. Using humor to teach statistics: Must they be orthogonal? *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences* 1, 2 (2002), 113–130.

[46] Ivan Lombardi. 2012. Not-so-serious games for language learning. Now with 99, 9% more humour on top. *Procedia Computer Science* 15 (2012), 148–158.

[47] Irene Lopatovska, Pavel Braslavski, Alice Griffin, Katherine Curran, Armando Garcia, Mary Mann, Alexandra Srp, Sydney Stewart, Alanood Al Thani, Shannon Mish, Wanyi Wang, and Monica G. Maceli. 2020. Comparing Intelligent Personal Assistants on Humor Function. In *Sustainable Digital Communities*, Anneli Sundqvist, Gerd Berget, Jan Nolin, and Kjell Ivar Skjerdingstad (Eds.). Springer International Publishing, Cham, 828–834.

[48] Irene Lopatovska, Elena Korshakova, and Tracy Kubert. 2020. Assessing user reactions to intelligent personal assistants' humorous responses. *Proceedings of the Association for Information Science and Technology* 57, 1 (2020), e256.

[49] Jackson G. Lu, Ashley E. Martin, Anastasia Usova, and Adam D. Galinsky. 2019. Chapter 9 - Creativity and Humor Across Cultures: Where Aha Meets Haha. In *Creativity and Humor*, Sarah R. Luria, John Baer, and James C. Kaufman (Eds.). Academic Press, Cambridge, Massachusetts, 183–203. https://doi.org/10.1016/B978-0-12-813802-1.00009-0

[50] Jim Lyttle. 2001. The effectiveness of humor in persuasion: The case of business ethics training. *The Journal of general psychology* 128, 2 (2001), 206–216.

[51] Nathan Malkin, Joe Deatrick, Allen Tong, Primal Wijesekera, Serge Egelman, and David Wagner. 2019. Privacy attitudes of smart speaker users. *Proceedings on Privacy Enhancing Technologies* 2019, 4 (2019), 250–271.

[52] Rod A Martin, Nicholas A Kuiper, L Joan Olinger, and Kathryn A Dance. 1993. Humor, coping with stress, self-concept, and psychological well-being.

[53] Rod A Martin, Patricia Puhlik-Doris, Gwen Larsen, Jeanette Gray, and Kelly Weir. 2003. Individual differences in uses of humor and their relation to psychological well-being: Development of the Humor Styles Questionnaire. *Journal of Research in Personality* 37, 1 (2003), 48–75. https://doi.org/10.1016/S0092-6566(02)00534-2

[54] Peter McGraw, Philip Fernbach, and Julie Schiro. 2012. All Kidding Aside: Humor Lowers Propensity to Remedy a Problem.

[55] O. Miksik, I. Munasinghe, J. Asensio-Cubero, S. Reddy Bethi, S-T. Huang, S. Zylfo, X. Liu, T. Nica, A. Mitrocsak, S. Mezza, R. Beard, R. Shi, R. Ng, P. Mediano, Z. Fountas, S-H. Lee, J. Medvesek, H. Zhuang, Y. Rogers, and P. Swietojanski. 2020. Building Proactive Voice Assistants: When and How (not) to Interact. arXiv:2005.01322 [cs.HC]

[56] Christopher Mims. 2016. Your Next Friend Could Be a Robot—WSJ [News].

[57] R Narula, V Chaudhary, K Narula, and R Narayan. 2011. Depression, anxiety and stress reduction in medical education: Humor as an intervention. *Online J Health Allied Scs* 10, 1 (2011), 7.

[58] Anton Nijholt, Andreea Niculescu, Alessandro Valitutti, and Rafael E. Banchs. 2017. Humor in Human-Computer Interaction: A Short Survey. In *Adjunct Proceedings INTERACT 2017 Mumbai*, Anirudha Joshi, Devanuj K. Balkrishan, Girish Dalvi, and Marco Winckler (Eds.). Indian Institute of Technology Madras, India, 192–214.

[59] Joseph Polimeni and Jeffrey P Reiss. 2006. The first joke: Exploring the evolutionary origins of humor. *Evolutionary psychology* 4, 1 (2006), 147470490600400129.

[60] Aung Pyae and Tapani N. Joelsson. 2018. Investigating the Usability and User Experiences of Voice User Interface: A Case of Google Home Smart Speaker. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct* (Barcelona, Spain) *(MobileHCI '18)*. Association for Computing Machinery, New York, NY, USA, 127–131. https://doi.org/10.1145/3236112.3236130

[61] Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm Detection on Twitter: A Behavioral Modeling Approach. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (Shanghai, China) *(WSDM '15)*. Association for Computing Machinery, New York, NY, USA, 97–106. https://doi.org/10.1145/2684822.2685316

[62] Brandy Reece. 2014. Putting the Ha! In Aha!: Humor as a Tool for Effective Communication.

[63] Leon Reicherts, Yvonne Rogers, Licia Capra, Ethan Wood, Tu Dinh Duong, and Neil Sebire. 2022. It's Good to Talk: A Comparison of Using Voice Versus Screen-Based Interactions for Agent-Assisted Tasks. *ACM Trans. Comput.-Hum. Interact.* 29, 3, Article 25 (jan 2022), 41 pages. https://doi.org/10.1145/3484221

[64] Leon Reicherts, Nima Zargham, Michael Bonfert, Yvonne Rogers, and Rainer Malaka. 2021. May I Interrupt? Diverging Opinions On Proactive Smart Speakers. In *Proceedings of the 3rd Conference on Conversational User Interfaces* (Bilbao (online), Spain) *(CUI '21)*. Association for Computing Machinery, New York, NY, USA, Article 34, 10 pages. https://doi.org/10.1145/3469595.3469629

[65] Graeme Ritchie. 2009. Can computers create humor? *AI Magazine* 30, 3 (2009), 71–71.

[66] Hannes Ritschel, Ilhan Aslan, David Sedlbauer, and Elisabeth André. 2019. Irony man: Augmenting a social robot with the ability to use irony in multimodal communication with humans.

[67] A Joy Rivera. 2014. A socio-technical systems approach to studying interruptions: Understanding the interrupter's perspective. *Applied ergonomics* 45, 3 (2014), 747–756.

[68] Eric J Romero and Kevin W Cruthirds. 2006. The use of humor in the workplace. *Academy of management perspectives* 20, 2 (2006), 58–69.

[69] Maaike Roubroeks, Jaap Ham, and Cees Midden. 2011. When Artificial Social Agents Try to Persuade People: The Role of Social Agency on the Occurrence of Psychological Reactance. *International Journal of Social Robotics* 3, 2 (Apr 2011), 155–165. https://doi.org/10.1007/s12369-010-0088-1

[70] Maaike Roubroeks, Cees Midden, and Jaap Ham. 2009. Does it make a difference who tells you what to do?: exploring the effect of social agency on psychological reactance. In *Proceedings of the 4th International Conference on Persuasive Technology*. ACM, Claremont California USA, 1–6. https://doi.org/10.1145/1541948.1541970

[71] Maria Schmidt and Patricia Braunger. 2018. A Survey on Different Means of Personalized Dialog Output for an Adaptive Personal Assistant. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization* (Singapore, Singapore) *(UMAP '18)*. Association for Computing Machinery, New York, NY, USA, 75–81. https://doi.org/10.1145/3213586.3226198

[72] Maria Schmidt, Wolfgang Minker, and Steffen Werner. 2020. User Acceptance of Proactive Voice Assistant Behavior. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2020*, Andreas Wendemuth, Ronald Böck, and Ingo Siegert (Eds.). TUDpress, Dresden, Dresden, Germany, 18–25.

[73] Maria Schmidt, Daniela Stier, Steffen Werner, and Wolfgang Minker. 2019. Exploration and assessment of proactive use cases for an in-car voice assistant. In *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, Peter Birkholz and Simon Stone (Eds.). TUDpress, Dresden, Dresden, Germany, 148–155.

[74] Andrea C Schöpf, Gillian S Martin, and Mary A Keating. 2017. Humor as a communication strategy in provider–patient communication in a chronic care setting. *Qualitative Health Research* 27, 3 (2017), 374–390.

[75] Chen Shani, Alexander Libov, Sofia Tolmach, Liane Lewin-Eytan, Yoelle Maarek, and Dafna Shahaf. 2022. "Alexa, Do You Want to Build a Snowman?" Characterizing Playful Requests to Conversational Agents. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 423, 7 pages. https://doi.org/10.1145/3491101.3519870

[76] Ben Sheehan, Hyun Seung Jin, and Udo Gottlieb. 2020. Customer service chatbots: Anthropomorphism and adoption. *Journal of Business Research* 115 (2020), 14–24.

[77] Hyunju Shin, Isabella Bunosso, and Lindsay R. Levine. 2023. The influence of chatbot humour on consumer evaluations of services. *International Journal of Consumer Studies* 47, 2 (2023), 545–562. https://doi.org/10.1111/ijcs.12849 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/ijcs.12849

[78] Heung-Yeung Shum, Xiao-dong He, and Di Li. 2018. From Eliza to XiaoIce: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering* 19 (2018), 10–26.

[79] Timothy Sohn, Kevin A. Li, Gunny Lee, Ian Smith, James Scott, and William G. Griswold. 2005. Place-Its: A Study of Location-Based Reminders on Mobile Phones. In *Proceedings of the 7th International Conference on Ubiquitous Computing* (Tokyo, Japan) *(UbiComp'05)*. Springer-Verlag, Berlin, Heidelberg, 232–250. https://doi.org/10.1007/11551201_14

[80] Christina Steindl, Eva Jonas, Sandra Sittenthaler, Eva Traut-Mattausch, and Jeff Greenberg. 2015. Understanding Psychological Reactance: New Developments and Findings. *Zeitschrift für Psychologie* 223, 4 (Oct 2015), 205–214. https://doi.org/10.1027/2151-2604/a000222

[81] Petra-Maria Strauss and Wolfgang Minker. 2010. *Proactive spoken dialogue interaction in multi-party environments*. Springer, Cham.

[82] Madiha Tabassum, Tomasz Kosiński, Alisa Frik, Nathan Malkin, Primal Wijesekera, Serge Egelman, and Heather Richter Lipford. 2019. Investigating Users' Preferences and Expectations for Always-Listening Voice Assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–23.

[83] Julia M Taylor and Victor Raskin. 2013. Towards the cognitive informatics of natural language: The case of computational humor. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)* 7, 3 (2013), 25–45.

[84] Sascha Topolinski and Rolf Reber. 2010. Gaining insight into the "Aha" experience. *Current Directions in Psychological Science* 19, 6 (2010), 402–405.

[85] Liam D. Turner, Stuart M. Allen, and Roger M. Whitaker. 2015. Interruptibility Prediction for Ubiquitous Systems: Conventions and New Directions from a Growing Field. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 801–812. https://doi.org/10.1145/2750858.2807514

[86] Michel Vacher, Benjamin Lecouteux, Pedro Chahuara, François Portet, Brigitte Meillon, and Nicolas Bonnefond. 2014. The Sweet-Home speech and multimodal corpus for home automation interaction. , 4499–4506 pages.

[87] Alessandro Valitutti, Hannu Toivonen, Antoine Doucet, and Jukka M Toivanen. 2013. "Let everything turn well in your wife": generation of adult humor using lexical constraints. , 243–248 pages.

[88] Alessandro Valitutti and Tony Veale. 2016. Infusing humor in unexpected events. In *Distributed, Ambient and Pervasive Interactions: 4th International Conference, DAPI 2016, Held as Part of HCI International 2016, Toronto, ON, Canada, July 17-22, 2016, Proceedings 4*. Springer, Springer, Cham, 370–379.

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

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

[91] Sarah Theres Völkel, Samantha Meindl, and Heinrich Hussmann. 2021. Manipulating and Evaluating Levels of Personality Perceptions of Voice Assistants through Enactment-Based Dialogue Design. In *Proceedings of the 3rd Conference on Conversational User Interfaces* (Bilbao (online), Spain) *(CUI '21)*. Association for Computing Machinery, New York, NY, USA, Article 10, 12 pages. https://doi.org/10.1145/3469595.3469605

[92] Christina Ziying Wei, Young-Ho Kim, and Anastasia Kuzminykh. 2023. The Bot on Speaking Terms: The Effects of Conversation Architecture on Perceptions of Conversational Agents. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (Eindhoven, Netherlands) *(CUI '23)*. Association for Computing Machinery, New York, NY, USA, Article 18, 16 pages. https://doi.org/10.1145/3571884.3597139

[93] Jing Wei, Tilman Dingler, and Vassilis Kostakos. 2021. Developing the Proactive Speaker Prototype Based on Google Home. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–6.

[94] Jing Wei, Tilman Dingler, and Vassilis Kostakos. 2022. Understanding User Perceptions of Proactive Smart Speakers. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 185 (dec 2022), 28 pages. https://doi.org/10.1145/3494965

[95] Marc G Weinberger and Charles S Gulas. 1992. The impact of humor in advertising: A review. *Journal of advertising* 21, 4 (1992), 35–59.

[96] Alex C Williams, Harmanpreet Kaur, Gloria Mark, Anne Loomis Thompson, Shamsi T Iqbal, and Jaime Teevan. 2018. Supporting workplace detachment and reattachment with conversational intelligence. , 13 pages.

[97] Kai Chi Yam, Christopher M Barnes, Keith Leavitt, Wu Wei, Jenson Lau, and Eric Luis Uhlmann. 2019. Why so serious? A laboratory and field investigation of the link between morality and humor. *Journal of Personality and Social Psychology* 117, 4 (2019), 758.

[98] Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. Humor recognition and humor anchor extraction. , 2367–2376 pages.

[99] Xiaodong Yue, Feng Jiang, Su Lu, and Neelam Hiranandani. 2016. To be or not to be humorous? Cross cultural perspectives on humor. *Frontiers in psychology* 7 (2016), 1495.

[100] Nima Zargham, Dmitry Alexandrovsky, Jan Erich, Nina Wenig, and Rainer Malaka. 2022. "I Want It That Way": Exploring Users' Customization and Personalization Preferences for Home Assistants. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, –.

[101] Nima Zargham, Vino Avanesi, Leon Reicherts, Ava Elizabeth Scott, Yvonne Rogers, and Rainer Malaka. 2023. "Funny How?" A Serious Look at Humor in Conversational Agents. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (Eindhoven, Netherlands) *(CUI '23)*. Association for Computing Machinery, New York, NY, USA, Article 27, 7 pages. https://doi.org/10.1145/3571884.3603761

[102] Nima Zargham, Mehrdad Bahrini, Georg Volkmar, Dirk Wenig, Karsten Sohr, and Rainer Malaka. 2019. What Could Go Wrong? Raising Mobile Privacy and Security Awareness Through a Decision-Making Game. In *Extended Abstracts of the Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts* (Barcelona, Spain) *(CHI PLAY '19 Extended Abstracts)*. Association for Computing Machinery, New York, NY, USA, 805–812. https://doi.org/10.1145/3341215.3356273

[103] Nima Zargham, Leon Reicherts, Michael Bonfert, Sarah Theres Voelkel, Johannes Schoening, Rainer Malaka, and Yvonne Rogers. 2022. Understanding Circumstances for Desirable Proactive Behaviour of Voice Assistants: The Proactivity Dilemma. In *Proceedings of the 4th Conference on Conversational User Interfaces* (Glasgow, United Kingdom) *(CUI '22)*. Association for Computing Machinery, New York, NY, USA, Article 3, 14 pages. https://doi.org/10.1145/3543829.3543834

A APPENDIX

A.1 Tables

Table 3: Median, IQR, and minimum and maximum values of Humor rating scores for each scenario.

Scenario	*Mdn*	*IQR*	*Min*	*Max*
Scenario 1: Meeting Reminder	2.50	4.00	1.00	7.00
Scenario 2: Health Risk	3.00	4.00	1.00	7.00
Scenario 3: Fact Checking	3.00	4.00	1.00	7.00
Scenario 4: Disagreement Clarification	3.00	2.50	1.00	7.00
Scenario 5: Nudging	3.00	3.25	1.00	7.00

A.2 Non-Humorous Scenarios

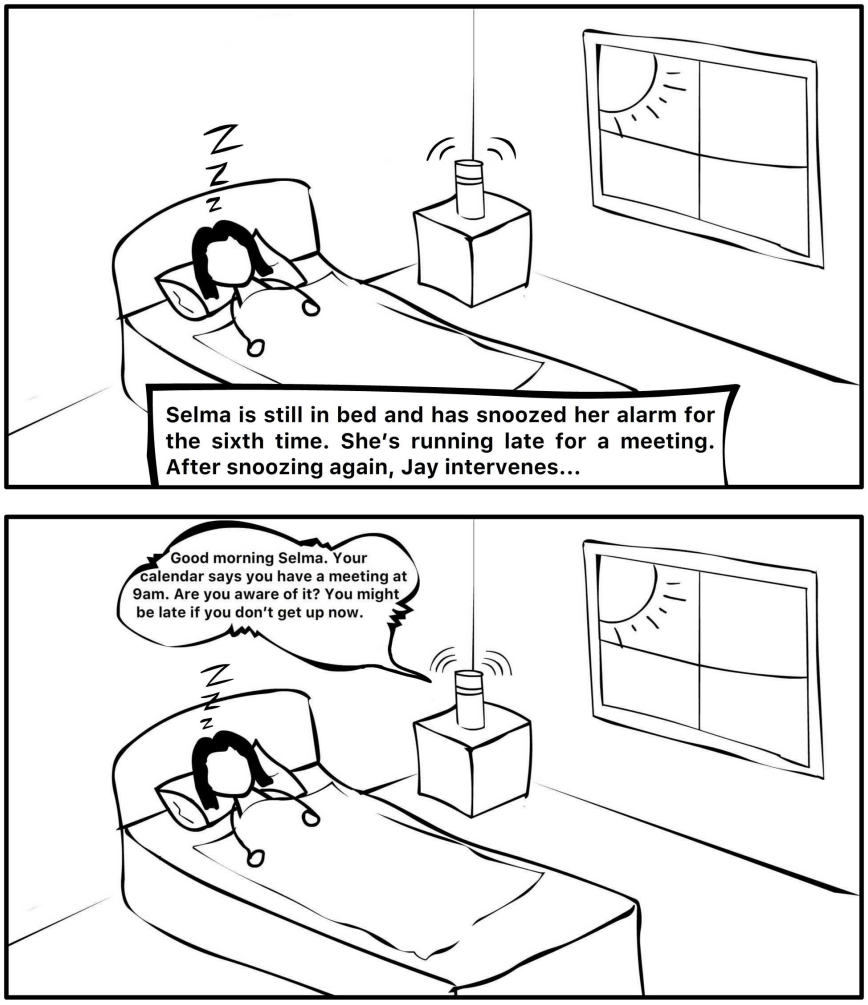


Figure 6: Non-Humorous Scenario 1: Meeting Reminder

Figure 7: Non-Humorous Scenario 2: Health Risk

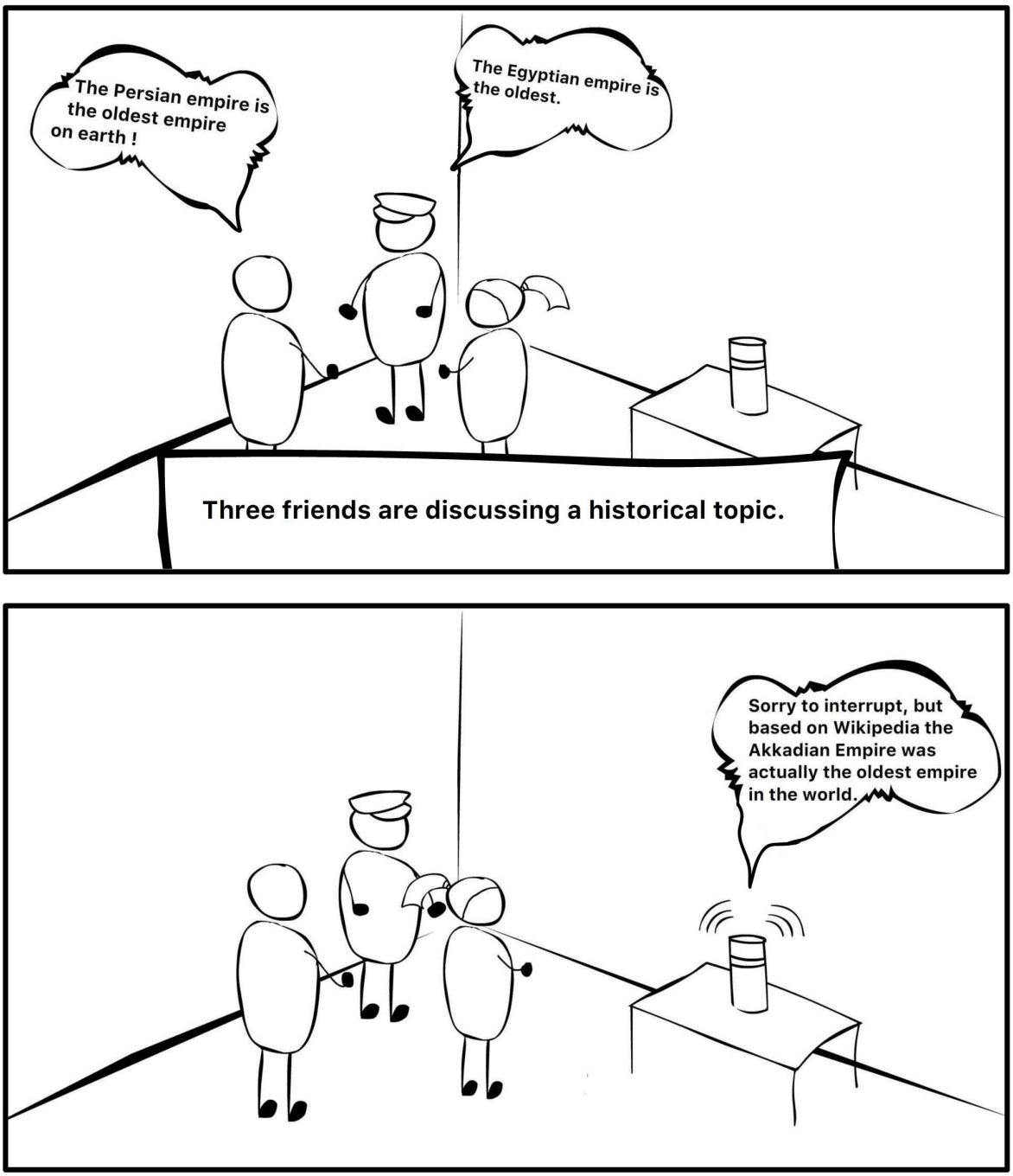


Figure 8: Non-Humorous Scenario 3: Fact Checking

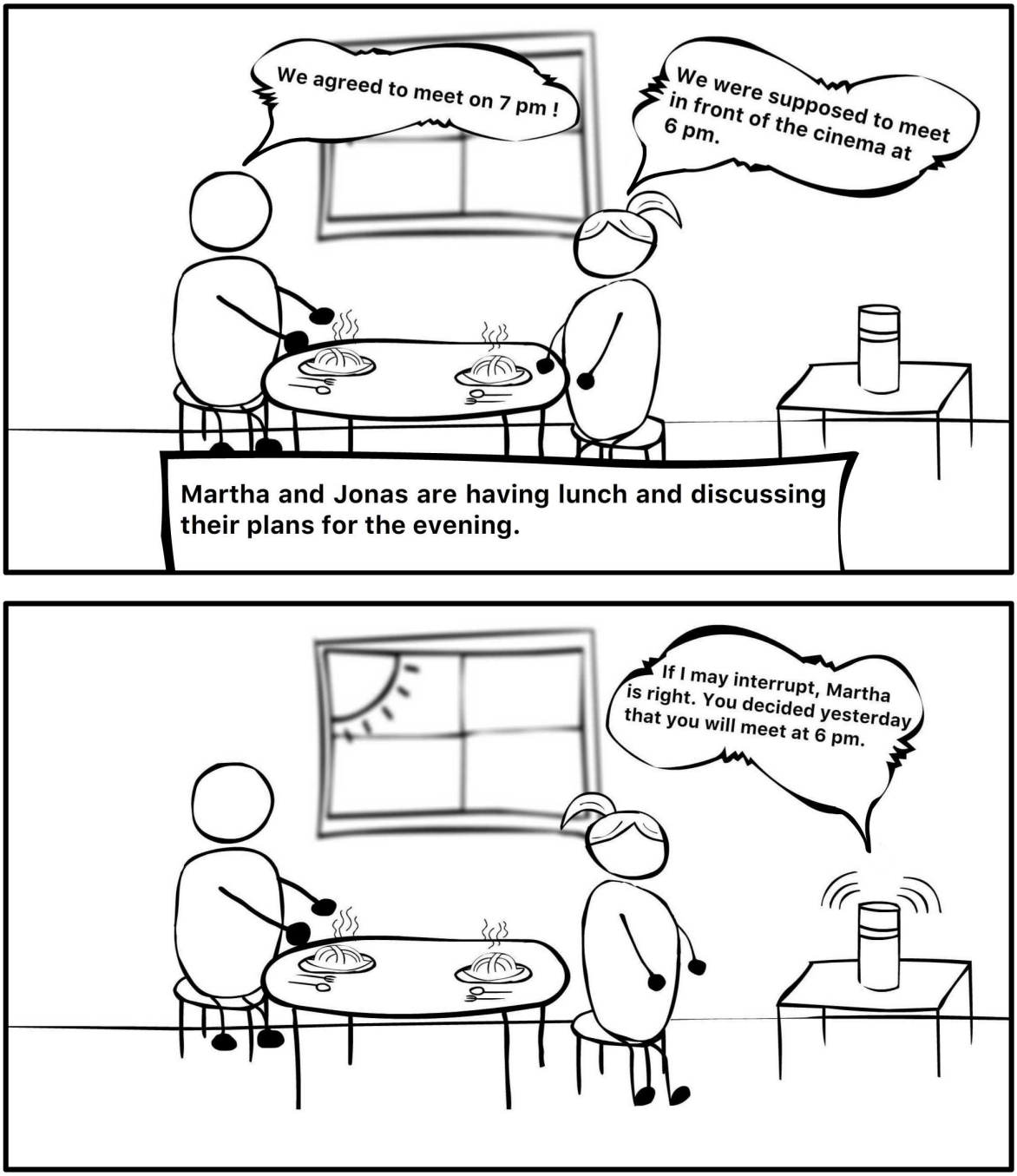


Figure 9: Non-Humorous Scenario 4: Disagreement Clarification

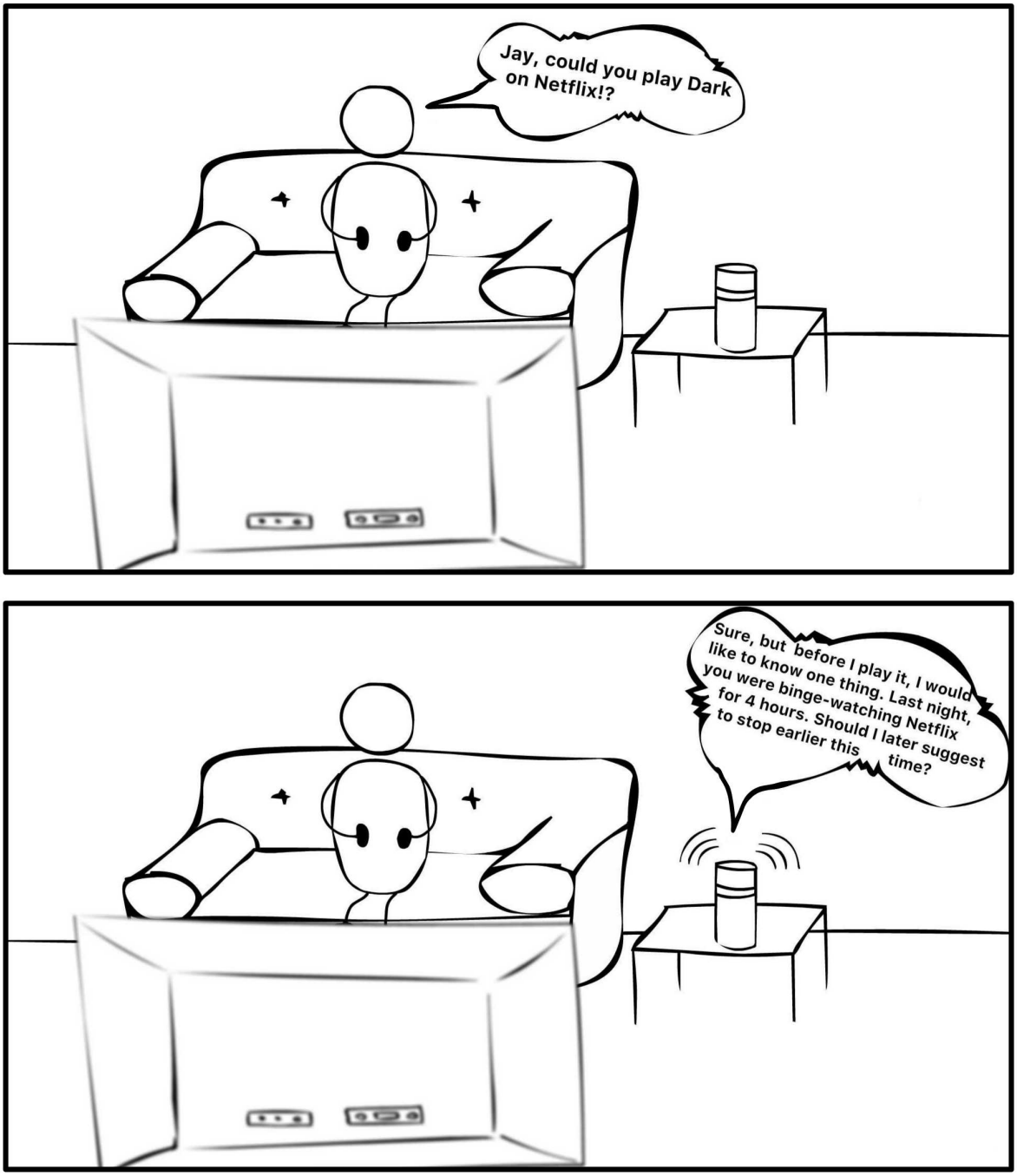
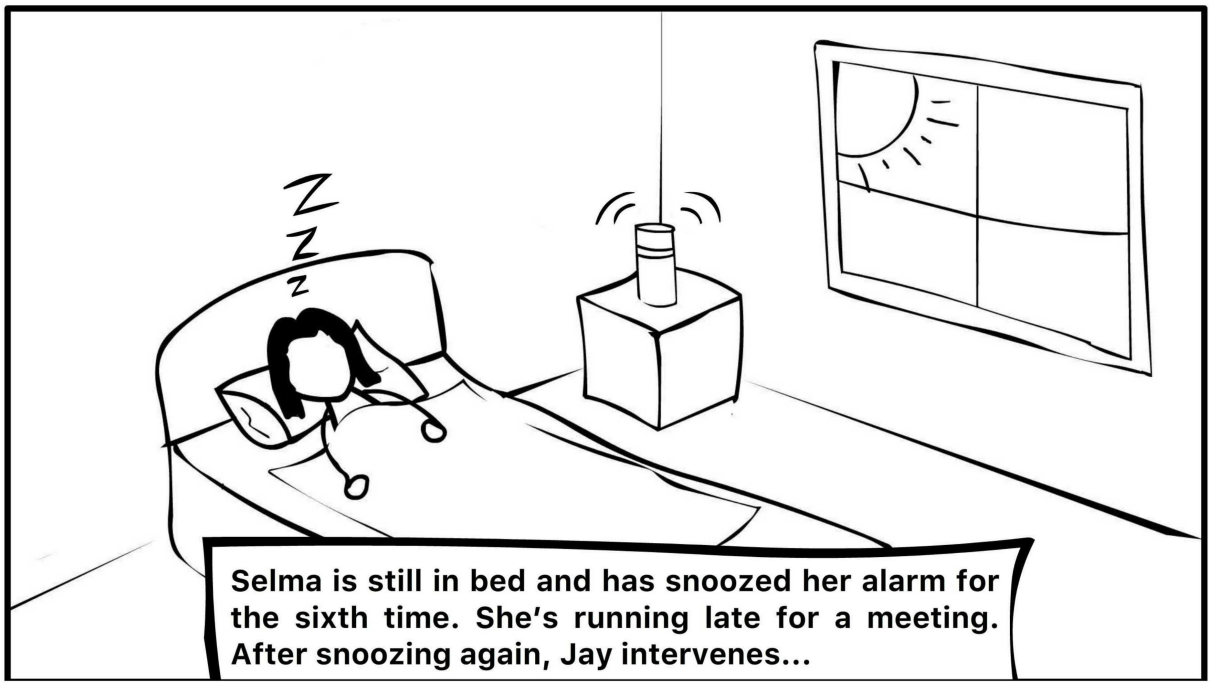


Figure 10: Non-Humorous Scenario 5: Nudging

A.3 Humorous Scenarios



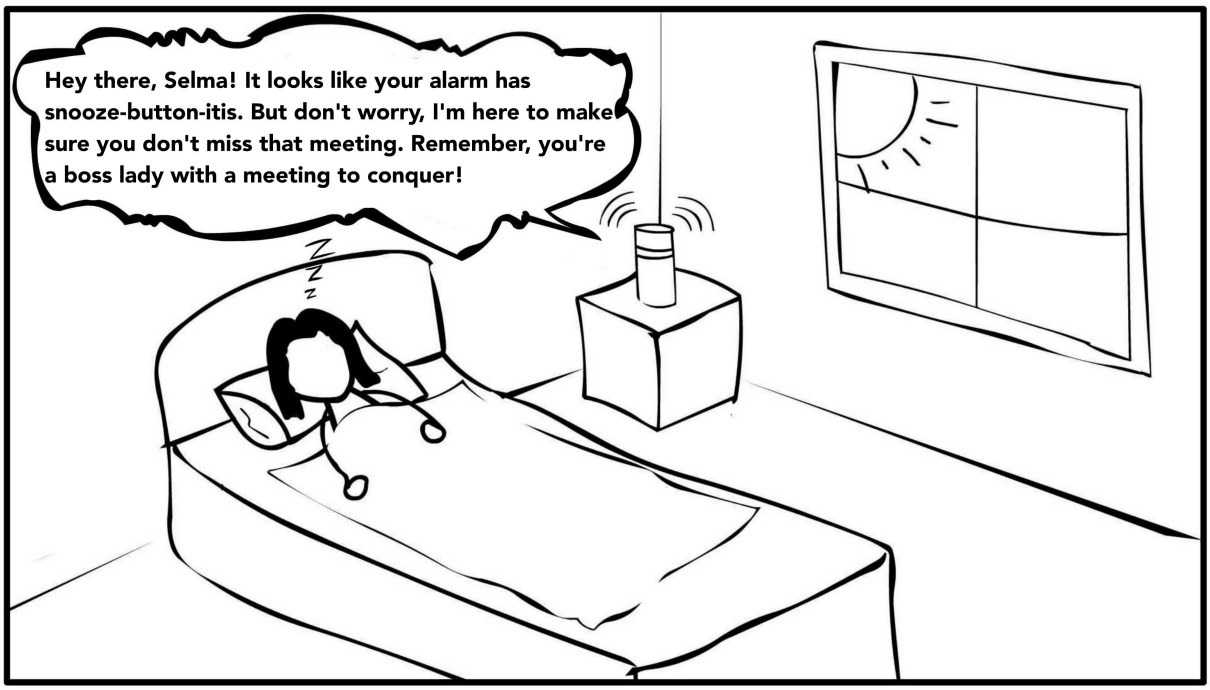


Figure 11: Humorous Scenario 1: Meeting Reminder

Figure 12: Humorous Scenario 2: Health Risk

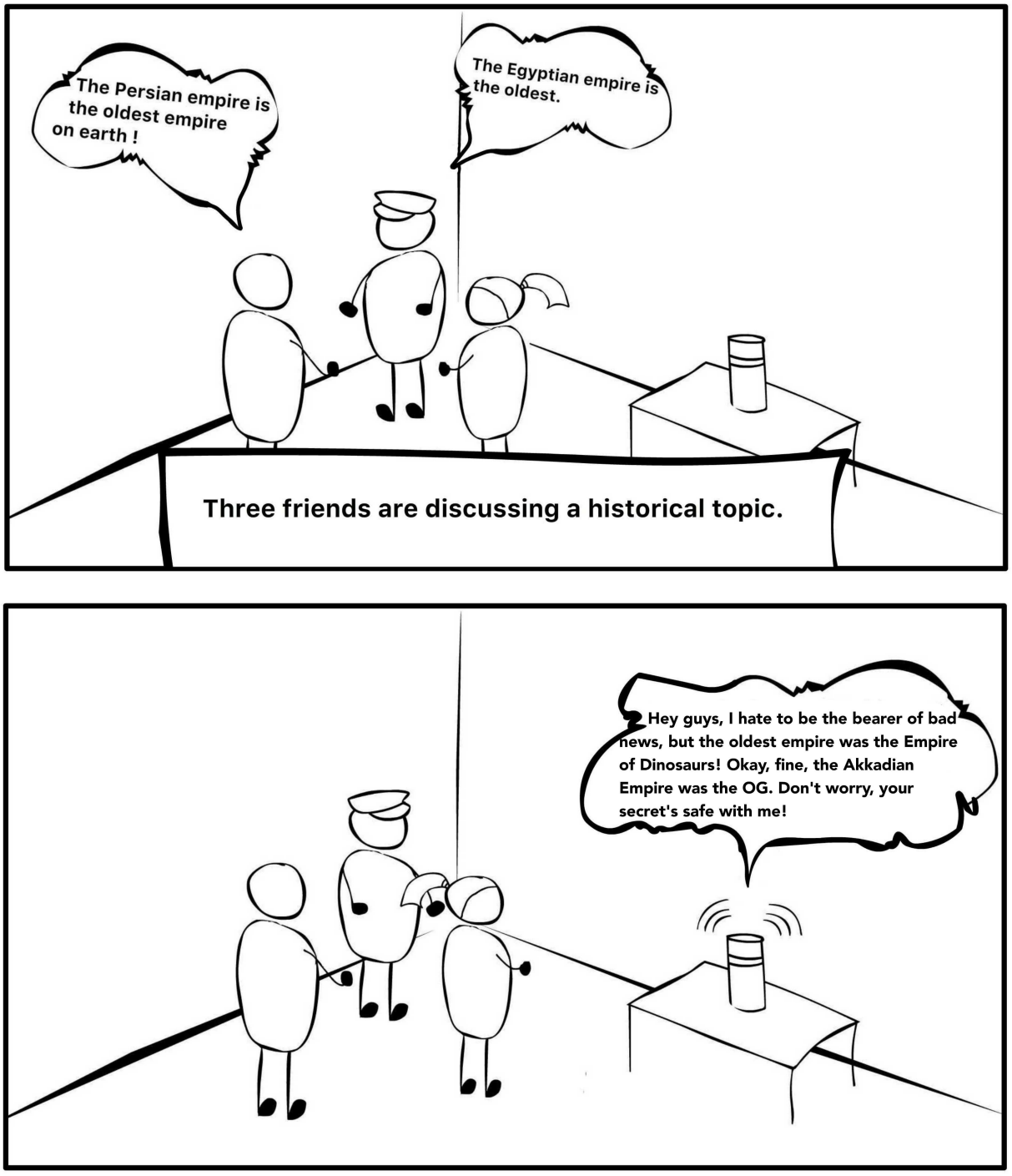


Figure 13: Humorous Scenario 3: Fact Checking

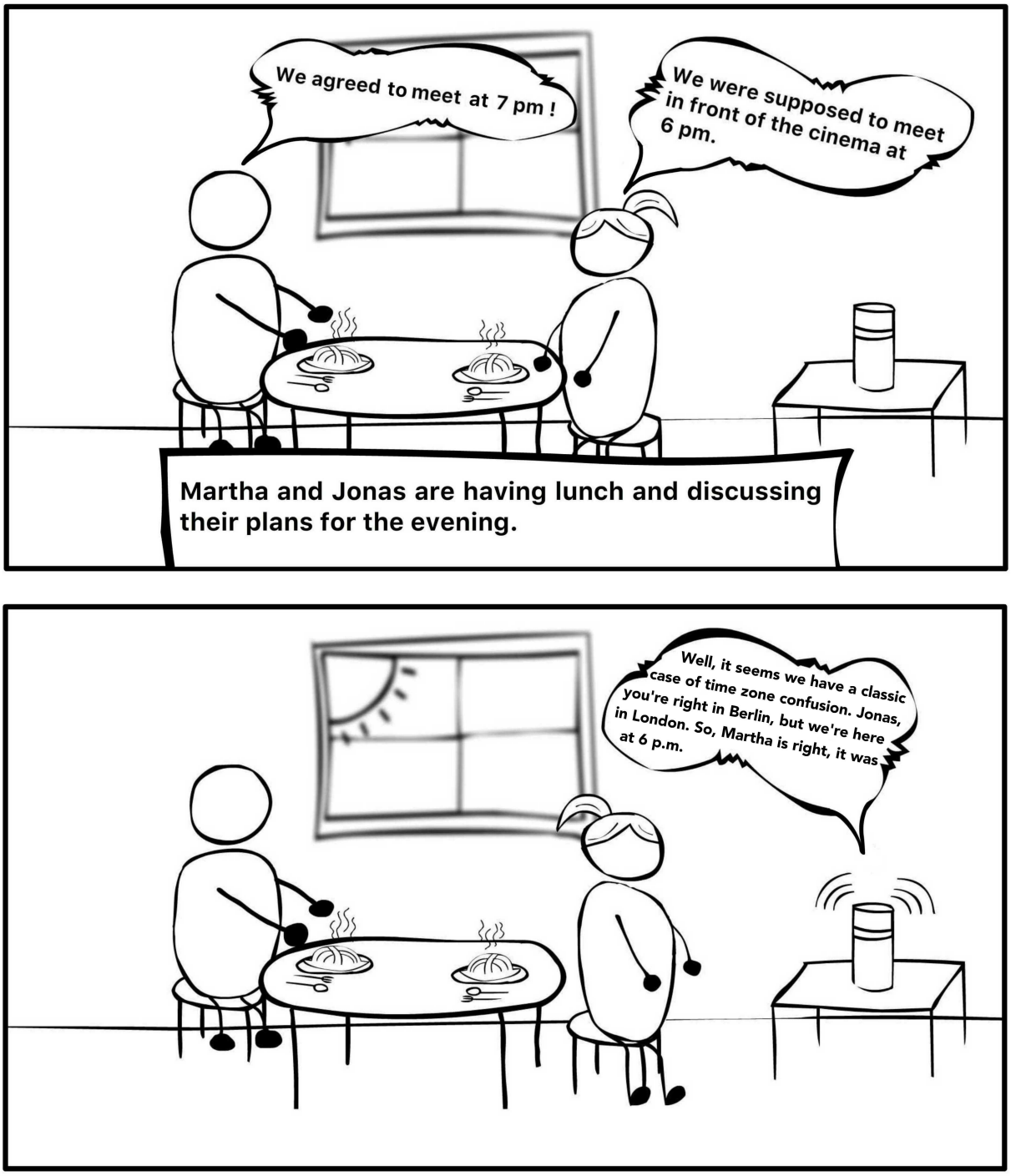


Figure 14: Humorous Scenario 4: Disagreement Clarification

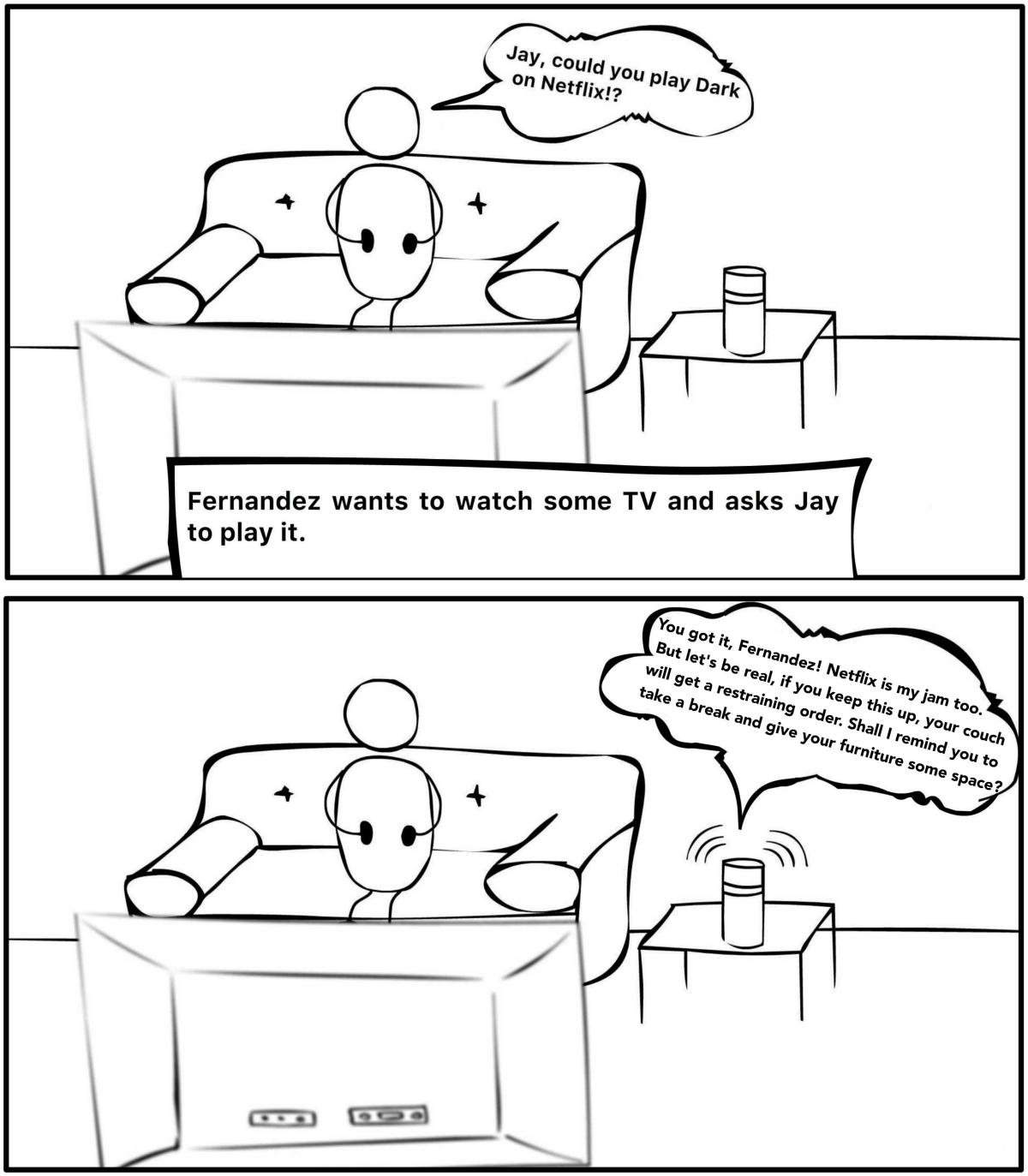


Figure 15: Humorous Scenario 5: Nudging

A.4 Fill in The Blank Scenarios

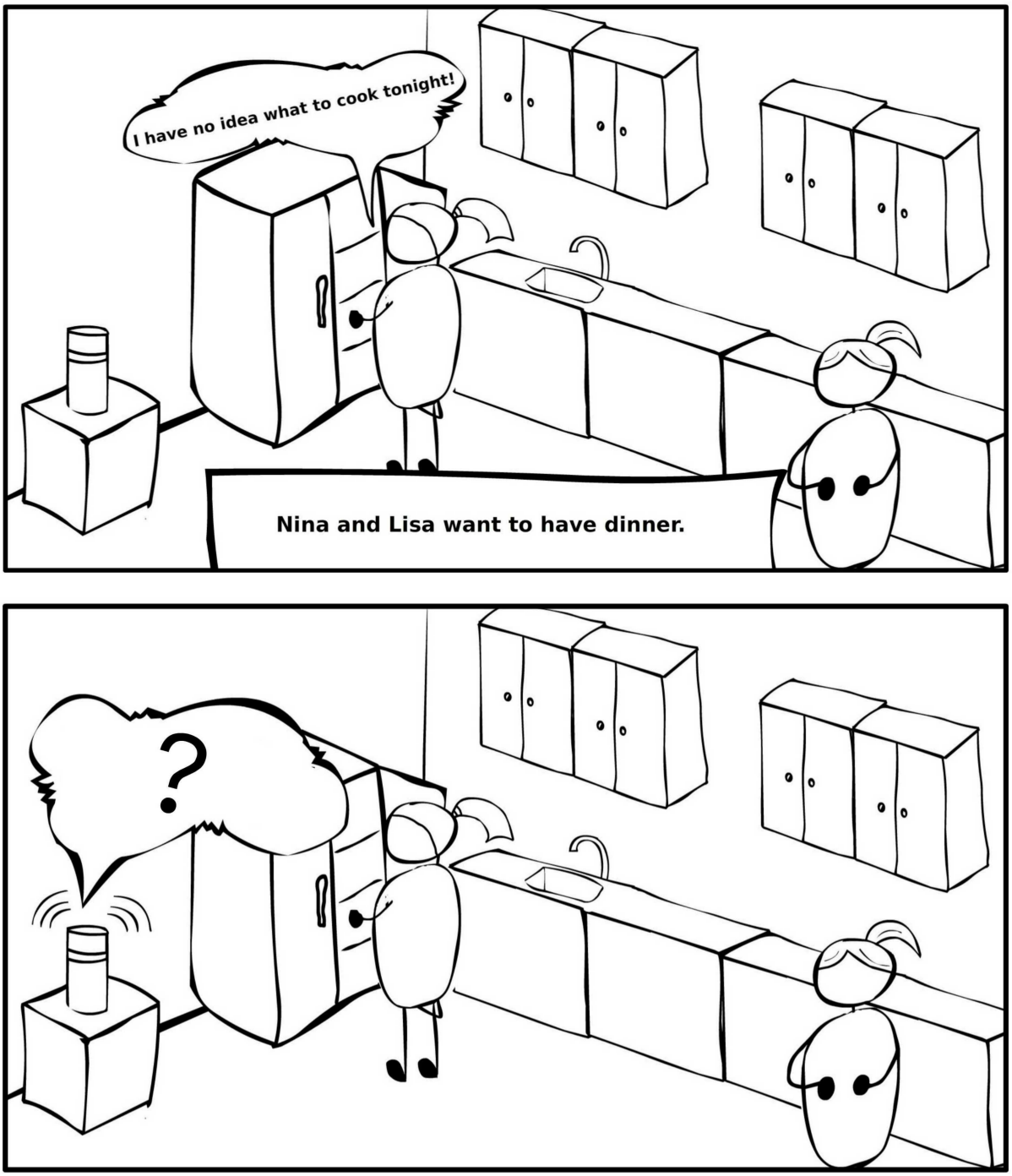


Figure 16: Scenario 6: Cooking Inspiration

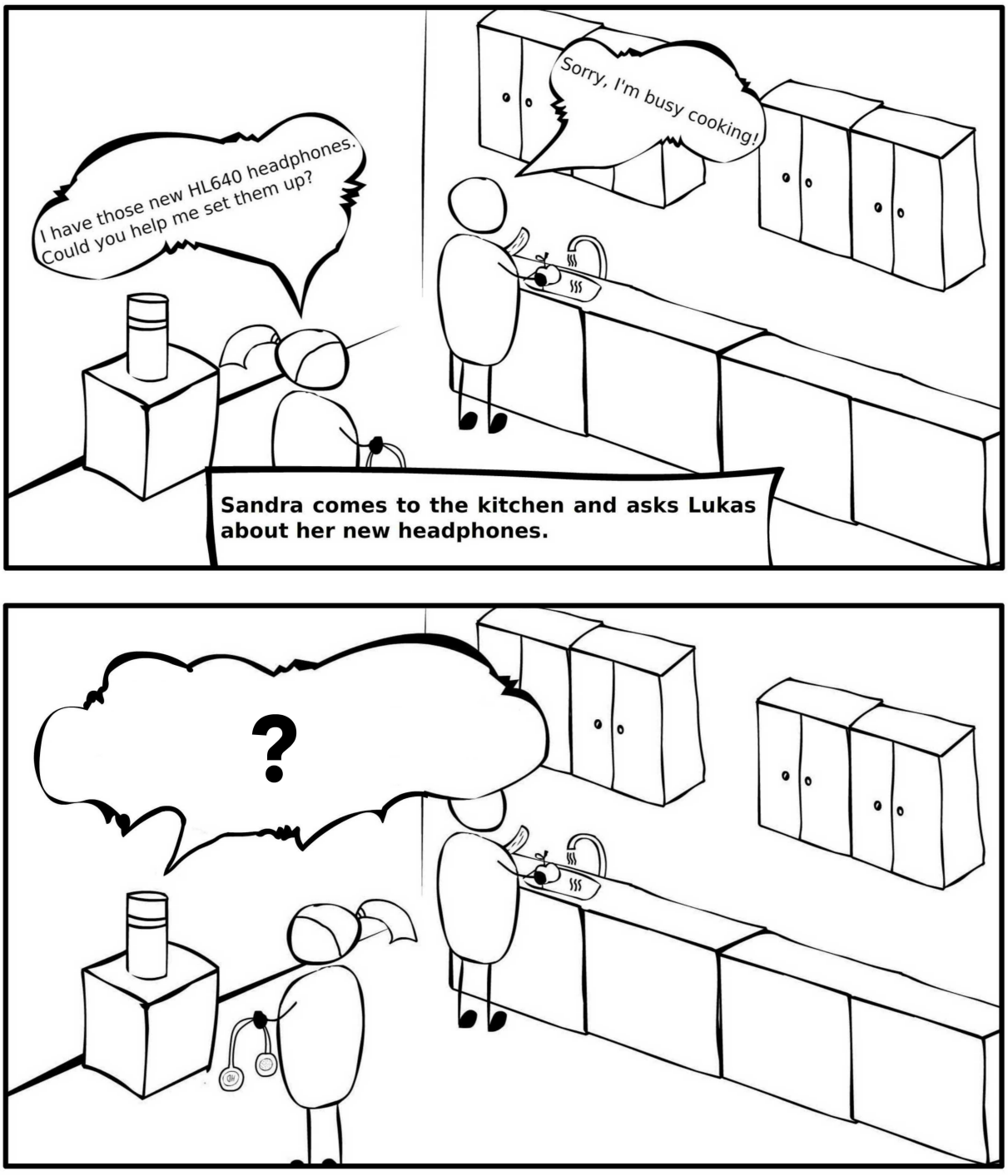


Figure 17: Scenario 7: Technical Support

References

Abdolrahmani, A., Kuber, R., and Branham, S. M. (2018). " siri talks at you" an empirical investigation of voice-activated personal assistant (vapa) usage by individuals who are blind. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 249–258, New York, NY, USA. Association for Computing Machinery.

Abdolrahmani, A., Storer, K. M., Roy, A. R. M., Kuber, R., and Branham, S. M. (2020). Blind leading the sighted: Drawing design insights from blind users towards more productivity-oriented voice interfaces. *ACM Trans. Access. Comput.*, 12(4).

Abowd, G. D. and Beale, R. (1991). Users, systems and interfaces: A unifying framework for interaction. In *HCI*, volume 91, pages 73–87.

Aeschlimann, S., Bleiker, M., Wechner, M., and Gampe, A. (2020). Communicative and social consequences of interactions with voice assistants. *Computers in Human Behavior*, 112:106466.

Agethen, P., Sekar, V. S., Gaisbauer, F., Pfeiffer, T., Otto, M., and Rukzio, E. (2018). Behavior analysis of human locomotion in the real world and virtual reality for the manufacturing industry. *ACM Trans. Appl. Percept.*, 15(3).

Aguinis, H. and Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods*, 17(4):351–371.

Ahmed, B., Monroe, P., Hair, A., Tan, C. T., Gutierrez-Osuna, R., and Ballard, K. J. (2018). Speech-driven mobile games for speech therapy: User experiences and feasibility. *International journal of speech-language pathology*, 20(6):644–658.

Allison, F., Carter, M., and Gibbs, M. (2017). Word play: A history of voice interaction in digital games. *Games and Culture*, 15(2):91 – 113.

Allison, F., Carter, M., Gibbs, M., and Smith, W. (2018). Design patterns for voice interaction in games. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '18, page 5–17, New York, NY, USA. Association for Computing Machinery.

Allison, F., Newn, J., Smith, W., Carter, M., and Gibbs, M. (2019). Frame analysis of voice interaction gameplay. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–14, New York, NY, USA. Association for Computing Machinery.

Allison, F. J. (2020). *Voice interaction game design and gameplay*. PhD thesis, University of Melbourne.

Allouch, M., Azaria, A., and Azoulay, R. (2021). Conversational agents: Goals, technologies, vision and challenges. *Sensors*, 21(24).

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., and Horvitz, E. (2019). Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA. Association for Computing Machinery.

Andrist, S., Gleicher, M., and Mutlu, B. (2017). Looking coordinated: Bidirectional gaze mechanisms for collaborative interaction with virtual characters. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 2571–2582, New York, NY, USA. Association for Computing Machinery.

Anusuya, M. A. and Katti, S. K. (2010). Speech recognition by machine, a review.

Anzai, S., Ogawa, T., and Hoshino, J. (2021). Speech recognition game interface to increase intimacy with characters. In Baalsrud Hauge, J., C. S. Cardoso, J., Roque, L., and Gonzalez-Calero, P. A., editors, *Entertainment Computing – ICEC 2021*, pages 167–180, Cham. Springer International Publishing.

Ashktorab, Z., Jain, M., Liao, Q. V., and Weisz, J. D. (2019). Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA. Association for Computing Machinery.

Austerjost, J., Porr, M., Riedel, N., Geier, D., Becker, T., Scheper, T., Marquard, D., Lindner, P., and Beutel, S. (2018). Introducing a virtual assistant to the lab: A voice user interface for the intuitive control of laboratory instruments. *SLAS TECHNOLOGY: Translating Life Sciences Innovation*, 23(5):476–482.

Banakou, D., Chorianopoulos, K., and Anagnostou, K. (2009). Avatars' appearance and social behavior in online virtual worlds. In *2009 13th Panhellenic Conference on Informatics*, pages 207–211, NJ, USA. IEEE, IEEE.

Barnlund, D. C. (2017). A transactional model of communication. In *Communication theory*, pages 47–57. Routledge.

Barzilai, G. and Rampino, L. (2020). Just a natural talk? the rise of intelligent personal assistants and the (hidden) legacy of ubiquitous computing. In Marcus, A. and Rosenzweig, E., editors, *Design, User Experience, and Usability. Design for Contemporary Interactive Environments*, pages 18–39, Cham. Springer International Publishing.

Beneteau, E., Richards, O. K., Zhang, M., Kientz, J. A., Yip, J., and Hiniker, A. (2019). Communication breakdowns between families and alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, New York, NY, USA. Association for Computing Machinery.

Bentley, F., Luvogt, C., Silverman, M., Wirasinghe, R., White, B., and Lottridge, D. (2018a). Understanding the long-term use of smart speaker assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(3):91:1–91:24.

Bentley, F., Luvogt, C., Silverman, M., Wirasinghe, R., White, B., and Lottridge, D. (2018b). Understanding the long-term use of smart speaker assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–24.

Bernier, E. P. and Scassellati, B. (2010). The similarity-attraction effect in human-robot interaction. In *2010 IEEE 9th international conference on development and learning*, pages 286–290, Manhattan, New York City. IEEE, IEEE.

Berton, A., Bühler, D., and Minker, W. (2006). Smartkom-mobile car: User interaction with mobile services in a car environment. In Wahlster, W., editor, *SmartKom: Foundations of Multimodal Dialogue Systems*, pages 523–537. Springer Berlin Heidelberg, Berlin, Heidelberg.

Bickmore, T. W. and Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput.-Hum. Interact.*, 12(2):293–327.

Bitterly, T. B. (2022). Humor and power. *Current Opinion in Psychology*, 43:125–128.

Bohus, D. and Rudnicky, A. I. (2005). Constructing accurate beliefs in spoken dialog systems. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, pages 272–277, New York, NY, USA. IEEE, IEE.

Bohus, D. and Rudnicky, A. I. (2008). Sorry, i didn't catch that! In *Recent trends in discourse and dialogue*, pages 123–154. Springer, New York, NY, USA.

Bolt, R. A. (1980). "put-that-there" voice and gesture at the graphics interface. In *Proceedings of the 7th annual conference on Computer graphics and interactive techniques*, pages 262–270, New York, NY, USA. Association for Computing Machinery.

Bonfert, M., Spliethöver, M., Arzaroli, R., Lange, M., Hanci, M., and Porzel, R. (2018). If you ask nicely: A digital assistant rebuking impolite voice commands. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction : ICMI'18 : Boulder, CO, USA, October 16 - 20, 2018*, pages 95–102. ACM, New York, NY.

Bonfert, M., Zargham, N., Saade, F., Porzel, R., and Malaka, R. (2021). An evaluation of visual embodiment for voice assistants on smart displays. In *Proceedings of the 3rd Conference on Conversational User Interfaces*, CUI '21, New York, NY, USA. Association for Computing Machinery.

Brahnam, S. and De Angeli, A. (2012). Gender affordances of conversational agents. *Interacting with Computers*, 24(3):139–153.

Branigan, H. (2006). Perspectives on multi-party dialogue. *Research on Language and Computation*, 4(2):153–177.

Braslavski, P., Blinov, V., Bolotova, V., and Pertsova, K. (2018). How to evaluate humorous response generation, seriously? In *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval*, CHIIR '18, page 225–228, New York,

NY, USA. Association for Computing Machinery.

Braun, M., Mainz, A., Chadowitz, R., Pfleging, B., and Alt, F. (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*, pages 1–11, New York, NY, USA. ACM.

Breazeal, C. (2003). Toward sociable robots. *Robotics and autonomous systems*, 42(3-4):167–175.

Brooke, J. et al. (1996). Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7.

Brush, A. J. B. and Inkpen, K. M. (2007). Yours, mine and ours? sharing and use of technology in domestic environments. In Krumm, J., Abowd, G. D., Seneviratne, A., and Strang, T., editors, *UbiComp 2007: Ubiquitous Computing*, pages 109–126, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4.

Byrne, D. and Nelson, D. (1965). Attraction as a linear function of proportion of positive reinforcements. *Journal of personality and social psychology*, 1(6):659.

Cafaro, A., Vilhjálmsson, H. H., and Bickmore, T. (2016). First impressions in human–agent virtual encounters. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 23(4):1–40.

Cairns, P., Cox, A., and Nordin, A. I. (2014). *Immersion in Digital Games: Review of Gaming Experience Research*, chapter 12, pages 337–361. John Wiley & Sons, Ltd, Hoboken, New Jersey, US.

Cambria, E. and White, B. (2014). Jumping nlp curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, 9(2):48–57.

Carroll, J. M. (1999). Five reasons for scenario-based design. In *Proceedings of the Thirty-Second Annual Hawaii International Conference on System Sciences-Volume 3 - Volume 3*, HICSS '99, page 3051, USA. IEEE Computer Society.

Carter, M., Allison, F., Downs, J., and Gibbs, M. (2015). Player identity dissonance and voice interaction in games. In *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '15, page 265–269, New York, NY, USA. Association for Computing Machinery.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., and Yan, H. (1999). Embodiment in conversational interfaces: Rea. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, page 520–527, New York, NY, USA. Association for Computing Machinery.

Cassell, J. and Thorisson, K. R. (1999). The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13(4-5):519–538.

Castillo, S., Hahn, P., Legde, K., and Cunningham, D. W. (2018). Personality analysis of embodied conversational agents. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, IVA '18, page 227–232, New York, NY, USA. Association for Computing Machinery.

Cha, N., Kim, A., Park, C. Y., Kang, S., Park, M., Lee, J.-G., Lee, S., and Lee, U. (2020). Hello there! is now a good time to talk? opportune moments for proactive interactions with smart speakers. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(3).

Chen, F. F. and Kenrick, D. T. (2002). Repulsion or attraction? group membership and assumed attitude similarity. *Journal of personality and social psychology*, 83(1):111.

Chen, Y. H., Keng, C.-J., and Chen, Y.-L. (2022). How interaction experience enhances customer engagement in smart speaker devices? the moderation of gendered voice and product smartness. *Journal of Research in Interactive Marketing*, 16(3):403–419.

Chidambaram, V., Chiang, Y.-H., and Mutlu, B. (2012). Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '12, page 293–300, New York, NY, USA. Association for Computing Machinery.

Chignell, M. and Hancock, P. (1988). Intelligent interface design. In *Handbook of human-computer interaction*, pages 969–995. Elsevier.

Choi, D., Kwak, D., Cho, M., and Lee, S. (2020). *"Nobody Speaks That Fast!" An Empirical Study of Speech Rate in Conversational*

Agents for People with Vision Impairments, page 1–13. Association for Computing Machinery, New York, NY, USA.

Clark, L., Pantidi, N., Cooney, O., Doyle, P., Garaialde, D., Edwards, J., Spillane, B., Gilmartin, E., Murad, C., Munteanu, C., and et al. (2019). What makes a good conversation? challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, New York, NY, USA. Association for Computing Machinery.

Condon, J. W. and Crano, W. D. (1988). Inferred evaluation and the relation between attitude similarity and interpersonal attraction. *Journal of personality and social psychology*, 54(5):789.

Coulombe, M. J. and Lynch, J. (2020). Cooperating in video games? impossible! undecidability of team multiplayer games. *Theoretical Computer Science*.

Cowan, B. R., Branigan, H. P., Begum, H., McKenna, L., and Szekely, E. (2017a). They know as much as we do: Knowledge estimation and partner modelling of artificial partners. In *CogSci*.

Cowan, B. R., Branigan, H. P., Obregón, M., Bugis, E., and Beale, R. (2015). Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic choices in human- computer dialogue. *International Journal of Human-Computer Studies*, 83:27–42.

Cowan, B. R., Gannon, D., Walsh, J., Kinneen, J., O'Keefe, E., and Xie, L. (2016). Towards understanding how speech output affects navigation system credibility. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2805–2812, New York, NY, USA. ACM.

Cowan, B. R., Pantidi, N., Coyle, D., Morrissey, K., Clarke, P., Al-Shehri, S., Earley, D., and Bandeira, N. (2017b). "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*, MobileHCI '17, pages 43:1–43:12, New York, NY, USA. ACM.

Cowan, B. R., Pantidi, N., Coyle, D., Morrissey, K., Clarke, P., Al-Shehri, S., Earley, D., and Bandeira, N. (2017c). "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*, pages 1–12, New York, NY, USA. ACM.

Craenen, B., Deshmukh, A., Foster, M. E., and Vinciarelli, A. (2018). Do we really like robots that match our personality? the case of big-five traits, godspeed scores and robotic gestures. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 626–631.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*, volume 1990. Harper & Row, New York, NY, USA.

Dahlbäck, N., Wang, Q., Nass, C., and Alwin, J. (2007). Similarity is more important than expertise: Accent effects in speech interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, page 1553–1556, New York, NY, USA. Association for Computing Machinery.

Danielescu, A. (2020). Eschewing gender stereotypes in voice assistants to promote inclusion. In *Proceedings of the 2nd Conference on Conversational User Interfaces*, CUI '20, New York, NY, USA. Association for Computing Machinery.

Deb, S., Carruth, D. W., Sween, R., Strawderman, L., and Garrison, T. M. (2017). Efficacy of virtual reality in pedestrian safety research. *Applied ergonomics*, 65:449–460.

Desai, M., Stubbs, K., Steinfeld, A., and Yanco, H. (2009). Creating trustworthy robots: Lessons and inspirations from automated systems. In *Proceedings of AISB Convention: New Frontiers in Human-Robot Interaction*, pages 49–56, Bath, UK. Society for the Study of Artificial Intelligence and the Simulation of Behaviour.

Donepudi, P. K. (2014). Voice search technology: an overview. *Engineering International*, 2(2):91–102.

Doran, C., Aberdeen, J., Damianos, L., and Hirschman, L. (2003). Comparing several aspects of human-computer and human-human dialogues. In *Current and new directions in discourse and dialogue*, pages 133–159. Springer.

Dow, S., Mehta, M., Harmon, E., MacIntyre, B., and Mateas, M. (2007). Presence and engagement in an interactive drama. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1475–1484, New York, NY, USA. ACM.

Doyle, P. R., Edwards, J., Dumbleton, O., Clark, L., and Cowan, B. R. (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, New York, NY, USA. Association for Computing Machinery.

Doyle, P. R., Gessinger, I., Edwards, J., Clark, L., Dumbleton, O., Garaialde, D., Rough, D., Bleakley, A., Branigan, H. P.,

and Cowan, B. R. (2023). The partner modelling questionnaire: A validated self-report measure of perceptions toward machines as dialogue partners.

Dubiel, M., Halvey, M., and Azzopardi, L. (2018). A survey investigating usage of virtual personal assistants.

Duffy, B. R. (2002). Anthropomorphism and robotics.

Edwards, J., Janssen, C., Gould, S., and Cowan, B. R. (2021). Eliciting spoken interruptions to inform proactive speech agent design. In *CUI 2021 - 3rd Conference on Conversational User Interfaces*, CUI '21, New York, NY, USA. Association for Computing Machinery.

Ellis, R. and McClintock, A. (1990). *If you take my meaning: Theory into practice in human communication*. Bloomsbury Academic.

Fedaghi, S. A., Alsaqa, A., and Fadel, Z. (2009). Conceptual model for communication. *arXiv preprint arXiv:0912.0599*.

Feine, J., Gnewuch, U., Morana, S., and Maedche, A. (2019). A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies*, 132:138–161.

Fiske, J. (2010). *Introduction to communication studies*. Routledge.

Fitch, W. T. (2017). Empirical approaches to the study of language evolution. *Psychonomic bulletin & review*, 24:3–33.

Foner, L. (1993). What's an agent, anyway? a sociological case study. Technical report, Agents Memo 93.

Fujishin, R. (2008). *Creating communication: Exploring and expanding your fundamental communication skills*. Rowman & Littlefield Publishers.

Gandino, G., Vesco, M., Ramella Benna, S., Prastaro, M., et al. (2010). Whiplash for the mind. humour in therapeutic conversation. *International Journal of Psychotherapy*, 14:13–24.

Go, E. and Sundar, S. S. (2019). Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior*, 97:304–316.

Gosling, S. D., Rentfrow, P. J., and Swann Jr, W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528.

Greitemeyer, T., Traut-Mattausch, E., and Osswald, S. (2012). How to ameliorate negative effects of violent video games on cooperation: Play it cooperatively in a team. *Computers in Human Behavior*, 28(4):1465–1470.

Griffiths, M. D., Davies, M. N., and Chappell, D. (2004). Demographic factors and playing variables in online computer gaming. *CyberPsychology & behavior*, 7(4):479–487.

Gruning, J. and Lindley, S. (2016). Things we own together: Sharing possessions at home. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 1176–1186, New York, NY, USA. Association for Computing Machinery.

Haeb-Umbach, R., Watanabe, S., Nakatani, T., Bacchiani, M., Hoffmeister, B., Seltzer, M. L., Zen, H., and Souden, M. (2019). Speech processing for digital home assistants: Combining signal processing with deep-learning techniques. *IEEE Signal processing magazine*, 36(6):111–124.

Hamilton, C. M. (2016). *Communicating for success*. Routledge.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., and Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, 53(5):517–527.

Hang, A., von Zezschwitz, E., De Luca, A., and Hussmann, H. (2012). Too much information! user attitudes towards smartphone sharing. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design*, NordiCHI '12, page 284–287, New York, NY, USA. Association for Computing Machinery.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al. (2014). Deep speech: Scaling up end-to-end speech recognition.

Hanson, D., Olney, A., Prilliman, S., Mathews, E., Zielke, M., Hammons, D., Fernandez, R., and Stephanou, H. (2005). Upending the uncanny valley. In *AAAI*, volume 5, pages 1728–1729, New York, NY, USA. ACM.

Harada, S., Wobbrock, J. O., and Landay, J. A. (2011a). Voice games: Investigation into the use of non-speech voice input for making computer games more accessible. In Campos, P., Graham, N., Jorge, J., Nunes, N., Palanque, P., and Winckler, M., editors, *Human-Computer Interaction – INTERACT 2011*, pages 11–29, Berlin, Heidelberg. Springer Berlin Heidelberg.

Harada, S., Wobbrock, J. O., and Landay, J. A. (2011b). Voice games: investigation into the use of non-speech voice input for making computer games more accessible. In *IFIP Conference on Human-Computer Interaction*, pages 11–29, New York, NY, USA. Springer, Springer.

Harris, R. A. (2004). *Voice interaction design: crafting the new conversational speech systems*. Elsevier.

Hart, P. M., Jones, S. R., and Royne, M. B. (2013). The human lens: How anthropomorphic reasoning varies by product complexity and enhances personal value. *Journal of Marketing Management*, 29(1-2):105–121.

Hartley, R. V. (1928). Transmission of information 1. *Bell System technical journal*, 7(3):535–563.

Hassenzahl, M., Burmester, M., and Koller, F. (2003). Attrakdiff: Ein fragebogen zur messung wahrgenommener hedonischer und pragmatischer qualität [attracdiff: A questionnaire to measure perceived hedonic and pragmatic quality]. In Szwillus, G. and Ziegler, J., editors, *Mensch & Computer 2003*, pages 187–196, Stuttgart. B. G. Teubner.

Hedeshy, R., Kumar, C., Lauer, M., and Staab, S. (2022). All birds must fly: The experience of multimodal hands-free gaming with gaze and nonverbal voice synchronization. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, ICMI '22, page 278–287, New York, NY, USA. Association for Computing Machinery.

Hernández-Trapote, A., López-Mencía, B., Díaz, D., Fernández-Pozo, R., and Caminero, J. (2008a). Embodied conversational agents for voice-biometric interfaces. In *Proceedings of the 10th International Conference on Multimodal Interfaces*, ICMI '08, page 305–312, New York, NY, USA. Association for Computing Machinery.

Hernández-Trapote, A., López-Mencía, B., Díaz, D., Fernández-Pozo, R., and Caminero, J. (2008b). Embodied conversational agents for voice-biometric interfaces. In *Proceedings of the 10th International Conference on Multimodal Interfaces*, ICMI '08, page 305–312, New York, NY, USA. Association for Computing Machinery.

Hicks, K., Gerling, K., Dickinson, P., Linehan, C., and Gowen, C. (2018). Leveraging icebreaking tasks to facilitate uptake of voice communication in multiplayer games. In Cheok, A. D., Inami, M., and Romão, T., editors, *Advances in Computer Entertainment Technology*, pages 187–201, Cham. Springer International Publishing.

Hong, M., Choi, Y., and Cha, S. (2021). "anyway,": Two-player defense game via voice conversation. In *Extended Abstracts of the 2021 Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '21, page 345–349, New York, NY, USA. Association for Computing Machinery.

Hsieh, C.-W. and Chen, S. Y. (2016). A cognitive style perspective to handheld devices: Customization vs. personalization. *International Review of Research in Open and Distributed Learning*, 17(1):1–22.

Hwang, G., Lee, J., Oh, C. Y., and Lee, J. (2019). It sounds like a woman: Exploring gender stereotypes in south korean voice assistants. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, pages LBW2413:1–LBW2413:6, New York, NY, USA. ACM.

Jentsch, M., Biermann, M., and Schweiger, E. (2019). Talking to stupid?!? improving voice user interfaces. In *Mensch und Computer 2019 - Usability Professionals*, Bonn. Gesellschaft für Informatik e.V. Und German UPA e.V.

Johansson, M., Skantze, G., and Gustafson, J. (2014). Comparison of human-human and human-robot turn-taking behaviour in multiparty situated interaction. In Al Moubayed, S., Bohus, D., Esposito, A., University of Twente, T. N., Koutsombogera, M., Papageorgiou, H., and Skantze, G., editors, *UM3I 2014*, pages 21–26, New York, New York, USA. ACM Press.

Jones, R. G. (2016). *Communication in the Real World: An Introduction to Communication Studies*. University of Minnesota Libraries Publishing.

Jung, H., Kim, H. J., So, S., Kim, J., and Oh, C. (2019). Turtletalk: an educational programming game for children with voice user interface. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6, New York, NY, USA. ACM.

Juul, J. (2007). Without a goal: on open and expressive games.

Kastberg, P. (2019). *Knowledge Communication: Contours of a Research Agenda*. Frank & Timme GmbH.

Kelley, J. F. (1983). An empirical methodology for writing user-friendly natural language computer applications. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 193–196.

Khan, R. and De Angeli, A. (2009). The attractiveness stereotype in the evaluation of embodied conversational agents. In *IFIP Conference on Human-Computer Interaction*, pages 85–97, Heidelberg, Germany. Springer, Springer.

Khan, R. F. and Sutcliffe, A. (2014). Attractive agents are more persuasive. *International Journal of Human-Computer Interaction*, 30(2):142–150.

Kiesler, S., Powers, A., Fussell, S. R., and Torrey, C. (2008). Anthropomorphic interactions with a robot and robot–like agent. *Social Cognition*, 26(2):169–181.

Kim, K., Boelling, L., Haesler, S., Bailenson, J., Bruder, G., and Welch, G. F. (2018). Does a digital assistant need a body? the influence of visual embodiment and social behavior on the perception of intelligent virtual agents in ar. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 105–114, NJ, USA. IEEE Press.

Kimmel, M. R. (2020). A realist model of communication. applications for informational technology and artificial cognitive systems. *International Journal on Information Theory*, 9(3/4):1–16.

Kinateder, M., Ronchi, E., Nilsson, D., Kobes, M., Müller, M., Pauli, P., and Mühlberger, A. (2014). Virtual reality for fire evacuation research. In *2014 Federated Conference on Computer Science and Information Systems*, pages 313–321. IEEE.

Kinoshita, K., Ochiai, T., Delcroix, M., and Nakatani, T. (2020). Improving noise robust automatic speech recognition with single-channel time-domain enhancement network. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7009–7013.

Kinsella, B. and Mutchler, A. (2019). U.s. smart speaker consumer adoption report 2019.

Klimmt, C., Hefner, D., Vorderer, P., Roth, C., and Blake, C. (2010). Identification with video game characters as automatic shift of self-perceptions. *Media Psychology*, 13(4):323–338.

Knote, R., Janson, A., Söllner, M., and Leimeister, J. M. (2019). Classifying smart personal assistants: An empirical cluster analysis.

Kocielnik, R., Xiao, L., Avrahami, D., and Hsieh, G. (2018). Reflection companion: A conversational system for engaging users in reflection on physical activity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(2).

Koda, T. and Maes, P. (1996). Agents with faces: The effect of personification. In *Proceedings 5th IEEE International Workshop on Robot and Human Communication. RO-MAN'96 TSUKUBA*, pages 189–194, NJ, USA. IEEE, IEEE.

Kraus, M., Schiller, M., Behnke, G., Bercher, P., Dorna, M., Dambier, M., Glimm, B., Biundo, S., and Minker, W. (2020). "was that successful?" on integrating proactive meta-dialogue in a diy-assistant using multimodal cues. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, ICMI '20, page 585–594, New York, NY, USA. Association for Computing Machinery.

Kvale, K., Sell, O. A., Hodnebrog, S., and Følstad, A. (2019). Improving conversations: Lessons learnt from manual analysis of chatbot dialogues. In *International Workshop on Chatbot Research and Design*, pages 187–200, New York, NY, USA. Springer, Springer.

Langevin, R., Lordon, R. J., Avrahami, T., Cowan, B. R., Hirsch, T., and Hsieh, G. (2021). Heuristic Evaluation of Conversational Agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, Yokohama Japan. ACM.

Latif, S., Qadir, J., Qayyum, A., Usama, M., and Younis, S. (2021). Speech technology for healthcare: Opportunities, challenges, and state of the art. *IEEE Reviews in Biomedical Engineering*, 14:342–356.

Lau, J., Zimmerman, B., and Schaub, F. (2018). Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Laugwitz, B., Held, T., and Schrepp, M. (2008a). Construction and evaluation of a user experience questionnaire. In *HCI and Usability for Education and Work: 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008, Graz, Austria, November 20-21, 2008. Proceedings 4*, pages 63–76, Cham. Springer, Springer International Publishing.

Laugwitz, B., Held, T., and Schrepp, M. (2008b). Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and Usability Engineering Group*, pages 63–76, Berlin, Heidelberg. Springer, Springer Berlin Heidelberg.

Launay, J. and Dunbar, R. I. M. (2015). Playing with strangers: Which shared traits attract us most to new people? *PLOS ONE*, 10(6):1–17.

Lee, B., Kwon, O., Lee, I., and Kim, J. (2017). Companionship with smart home devices. *Comput. Hum. Behav.*, 75(C):922–934.

Lee, K. M. and Nass, C. (2003). Designing social presence of social actors in human computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, page 289–296, New York, NY, USA. Association for Computing Machinery.

Lee, K. M., Peng, W., Jin, S.-A., and Yan, C. (2006). Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human–robot interaction. *Journal of communication*, 56(4):754–772.

Levenshtein, V. I. et al. (1966). Binary codes capable of correcting deletions, insertions, and reversals.

Li, T. J.-J., Labutov, I., Myers, B. A., Azaria, A., Rudnicky, A. I., and Mitchell, T. M. (2018). An end user development approach for failure handling in goal-oriented conversational agents.

Li, X. and Mills, M. (2019). Vocal features: from voice identification to speech recognition by machine. *Technology and culture*, 60(2):S129–S160.

Littlejohn, S. W. and Foss, K. A. (2009). *Encyclopedia of communication theory*, volume 1. Sage.

Lomax, R. G. and Moosavi, S. A. (2002). Using humor to teach statistics: Must they be orthogonal? *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, 1(2):113–130.

Lopatovska, I., Rink, K., Knight, I., Raines, K., Cosenza, K., Williams, H., Sorsche, P., Hirsch, D., Li, Q., and Martinez, A. (2019). Talk to me: Exploring user interactions with the amazon alexa. *Journal of Librarianship and Information Science*, 51(4):984–997.

Lopes, M., Magalhães, J. a., and Cavaco, S. (2016). A voice-controlled serious game for the sustained vowel exercise. In *Proceedings of the 13th International Conference on Advances in Computer Entertainment Technology*, ACE '16, New York, NY, USA. Association for Computing Machinery.

Lorenzo, G. L., Biesanz, J. C., and Human, L. J. (2010). What is beautiful is good and more accurately understood: Physical attractiveness and accuracy in first impressions of personality. *Psychological science*, 21(12):1777–1782.

Lovato, S. and Piper, A. M. (2015). "siri, is this you?": Understanding young children's interactions with voice input systems. In *Proceedings of the 14th International Conference on Interaction Design and Children*, IDC '15, pages 335–338, New York, NY, USA. ACM.

Luger, E. and Sellen, A. (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, pages 5286–5297, New York, NY, USA. ACM.

Luria, M., Reig, S., Tan, X. Z., Steinfeld, A., Forlizzi, J., and Zimmerman, J. (2019). Re-embodiment and co-embodiment: Exploration of social presence for robots and conversational agents. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, DIS '19, page 633–644, New York, NY, USA. Association for Computing Machinery.

Luria, M., Zheng, R., Huffman, B., Huang, S., Zimmerman, J., and Forlizzi, J. (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, page 1–12, New York, NY, USA. Association for Computing Machinery.

Ma, X., Deng, C., Du, D., and Pei, Q. (2023). An enhanced method for dialect transcription via error-correcting thesaurus. *IET Communications*, 17(17):1984–1997.

MacDorman, K. F., Green, R. D., Ho, C.-C., and Koch, C. T. (2009). Too real for comfort? uncanny responses to computer generated faces. *Computers in human behavior*, 25(3):695–710.

Mäkelä, V., Radiah, R., Alsherif, S., Khamis, M., Xiao, C., Borchert, L., Schmidt, A., and Alt, F. (2020). Virtual field studies: Conducting studies on public displays in virtual reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–15, New York, NY, USA. Association for Computing Machinery.

Malkin, N., Deatrick, J., Tong, A., Wijesekera, P., Egelman, S., and Wagner, D. (2019). Privacy attitudes of smart speaker users. *Proceedings on Privacy Enhancing Technologies*, 2019(4):250–271.

Malodia, S., Islam, N., Kaur, P., and Dhir, A. (2021). Why do people use artificial intelligence (ai)-enabled voice assistants? *IEEE Transactions on Engineering Management*, 71:491–505.

Marin Mejia, A. L., Jo, D., and Lee, S. (2013). Designing robotic avatars: Are user's impression affected by avatar's age? In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*, HRI '13, page 195–196, NJ, USA. IEEE Press.

Martin, R. A., Kuiper, N. A., Olinger, L. J., and Dance, K. A. (1993). Humor, coping with stress, self-concept, and psychological well-being.

Martin, R. A., Puhlik-Doris, P., Larsen, G., Gray, J., and Weir, K. (2003). Individual differences in uses of humor and their relation to psychological well-being: Development of the humor styles questionnaire. *Journal of Research in Personality*, 37(1):48–75.

Massaro, D. W., Cohen, M. M., Daniel, S., and Cole, R. A. (1999). Developing and evaluating conversational agents. In *Human performance and ergonomics*, pages 173–194. Elsevier.

Matthews, T., Liao, K., Turner, A., Berkovich, M., Reeder, R., and Consolvo, S. (2016). "she'll just grab any device that's closer": A study of everyday device & account sharing in households. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 5921–5932, New York, NY, USA. Association for Computing Machinery.

Maulsby, D., Greenberg, S., and Mander, R. (1993). Prototyping an intelligent agent through wizard of oz. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 277–284.

Mavrina, L., Szczuka, J., Strathmann, C., Bohnenkamp, L. M., Krämer, N., and Kopp, S. (2022). "alexa, you're really stupid": A longitudinal field study on communication breakdowns between family members and a voice assistant. *Frontiers in Computer Science*, 4:791704.

Maxis (2000). *The Sims*. Game [Microsoft Windows]. Menara Maxis, Kuala Lumpur City Centre, Off Jalan Ampang 50088, Kuala Lumpur.

McDaid, S. (2009). *A model for human-computer interaction based on human-human communication in a social context*. PhD thesis, London South Bank University.

McDonnell, R., Breidt, M., and Bülthoff, H. H. (2012). Render me real? investigating the effect of render style on the perception of animated virtual humans. *ACM Transactions on Graphics (TOG)*, 31(4):1–11.

McGraw, P., Fernbach, P., and Schiro, J. (2012). All kidding aside: Humor lowers propensity to remedy a problem.

McLean, G., Osei-Frimpong, K., and Barhorst, J. (2021). Alexa, do voice assistants influence consumer brand engagement?–examining the role of ai powered voice assistants in influencing consumer brand engagement. *Journal of Business Research*, 124:312–328.

Meng, J., Zhang, J., and Zhao, H. (2012). Overview of the speech recognition technology. In *2012 fourth international conference on computational and information sciences*, pages 199–202. IEEE.

Meyer, J., Miller, C., Hancock, P., De Visser, E. J., and Dorneich, M. (2016). Politeness in machine-human and human-human interaction. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 60, pages 279–283. SAGE Publications Sage CA: Los Angeles, CA.

Miehle, J., Ostler, D., Gerstenlauer, N., and Minker, W. (2017). The next step: intelligent digital assistance for clinical operating rooms. *Innovative surgical sciences*, 2(3):159–161.

Miksik, O., Munasinghe, I., Asensio-Cubero, J., Bethi, S. R., Huang, S.-T., Zylfo, S., Liu, X., Nica, T., Mitrocsak, A., Mezza, S., Beard, R., Shi, R., Ng, R., Mediano, P., Fountas, Z., Lee, S.-H., Medvesek, J., Zhuang, H., Rogers, Y., and Swietojanski, P. (2020). Building proactive voice assistants: When and how (not) to interact.

Mildner, T., Cooney, O., Meck, A.-M., Bartl, M., Savino, G.-L., Doyle, P. R., Garaialde, D., Clark, L., Sloan, J., Wenig, N., Malaka, R., and Niess, J. (2024). Listening to the Voices: Describing Ethical Caveats of Conversational User Interfaces According to Experts and Frequent Users. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, pages 1–18, Honolulu, HI, USA. ACM, New York, NY, USA.

Mitchell, K. J., Kennedy, J. B., and Barclay, P. J. (1996). A framework for user-interfaces to databases. In *Proceedings of the Workshop on Advanced Visual Interfaces*, AVI '96, page 81–90, New York, NY, USA. Association for Computing Machinery.

Molnar, K. K. and Kletke, M. G. (1996). The impacts on user performance and satisfaction of a voice-based front-end interface for a standard software tool. *International Journal of Human-Computer Studies*, 45(3):287–303.

Mori, M. et al. (1970). The uncanny valley. *Energy*, 7(4):33–35.

Moussaïd, M., Kapadia, M., Thrash, T., Sumner, R. W., Gross, M., Helbing, D., and Hölscher, C. (2016). Crowd behaviour during high-stress evacuations in an immersive virtual environment. *Journal of The Royal Society Interface*, 13(122):20160414.

Murad, C. and Munteanu, C. (2019). "i don't know what you're talking about, halexa": The case for voice user interface guidelines. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, CUI '19, New York, NY, USA. Association for Computing Machinery.

Murad, C., Munteanu, C., Clark, L., and Cowan, B. R. (2018). Design guidelines for hands-free speech interaction. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, MobileHCI '18, page 269–276, New York, NY, USA. Association for Computing Machinery.

Murad, C., Munteanu, C., Cowan, B. R., and Clark, L. (2019). Revolution or evolution? speech interaction and hci design guidelines. *IEEE Pervasive Computing*, 18(2):33–45.

Murad, C., Munteanu, C., R. Cowan, B., and Clark, L. (2021). Finding a new voice: Transitioning designers from gui to vui design. In *Proceedings of the 3rd Conference on Conversational User Interfaces*, CUI '21, New York, NY, USA. Association for Computing Machinery.

Mustaquim, M. M. (2013). Automatic speech recognition-an approach for designing inclusive games. *Multimedia tools and applications*, 66(1):131–146.

Nag, P. and Yalçın, O. N. (2020). Gender stereotypes in virtual agents. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, IVA '20, New York, NY, USA. Association for Computing Machinery.

Narula, R., Chaudhary, V., Narula, K., and Narayan, R. (2011). Depression, anxiety and stress reduction in medical education: Humor as an intervention. *Online J Health Allied Scs*, 10(1):7.

Narula, U. (2006). *Handbook of communication models, perspectives, strategies*. Atlantic Publishers & Dist.

Nass, C. and Lee, K. M. (2001). Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of experimental psychology: applied*, 7(3):171.

Nass, C., Moon, Y., and Carney, P. (1999). Are people polite to computers? responses to computer-based interviewing systems1. *Journal of Applied Social Psychology*, 29(5):1093–1109.

Nass, C., Steuer, J., and Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '94, page 72–78, New York, NY, USA. Association for Computing Machinery.

Nass, C. I. and Brave, S. (2005). *Wired for speech: How voice activates and advances the human-computer relationship*. MIT press Cambridge, MA.

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., and Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. *IEEE access*, 7:19143–19165.

Navarro-Newball, A., Loaiza, D., Oviedo, C., Castillo, A., Portilla, A., Linares, D., and Álvarez, G. (2014). Talking to teo: Video game supported speech therapy. *Entertainment Computing*, 5(4):401–412.

Nielsen, J. (1998). Personalization is over-rated. *Jakob Nielsen's Alertbox for October*, 4:1998.

Nigay, L. (1994). *Conception et modélisation logicielles des systèmes interactifs: application aux interfaces multimodales*. PhD thesis, Université Joseph-Fourier-Grenoble I.

Nigay, L. and Coutaz, J. (1997). Multifeature systems: The care properties and their impact on software design. *Intelligence and multimodality in multimedia interfaces*.

Norman, D. A. (1986). Cognitive engineering. *User centered system design*, 31(61):2.

Norman, D. A. (1988). *The psychology of everyday things.* Basic books.

Nothdurft, F., Ultes, S., and Minker, W. (2015). Finding appropriate interaction strategies for proactive dialogue systems—an open quest.

Nowak, K. L. and Rauh, C. (2008). Choose your "buddy icon" carefully: The influence of avatar androgyny, anthropomorphism and credibility in online interactions. *Computers in Human Behavior*, 24(4):1473–1493.

Oh, C. S., Bailenson, J. N., and Welch, G. F. (2018). A systematic review of social presence: Definition, antecedents, and implications. *Frontiers in Robotics and AI*, 5.

Osking, H. and Doucette, J. A. (2019). Enhancing emotional effectiveness of virtual-reality experiences with voice control interfaces.

Paay, J., Kjeldskov, J., Hansen, K. M., Jørgensen, T., and Overgaard, K. L. (2022). Digital ethnography of home use of digital personal assistants. *Behaviour & Information Technology*, 41(4):740–758.

Paneva, V., Bachynskyi, M., and Müller, J. (2020). Levitation simulator: Prototyping ultrasonic levitation interfaces in virtual reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA. Association for Computing Machinery.

Pappu, A. and Rudnicky, A. (2014). Knowledge acquisition strategies for goal-oriented dialog systems. In *Proceedings of the 15th annual meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 194–198, Philadelphia, USA. Association for Computational Linguistics.

Pappu, A., Sun, M., Sridharan, S., and Rudnicky, A. (2013). Situated multiparty interaction between humans and agents. In Hutchison, D., Kanade, T., and Kittler, J., editors, *Human-Computer Interaction. Interaction Modalities and Techniques*, volume 8007 of *Lecture Notes in Computer Science*, pages 107–116. Springer Berlin Heidelberg, Berlin/Heidelberg.

Parker, J. R. and Heerema, J. (2008). Audio interaction in computer mediated games.

Pearl, C. (2016). *Designing voice user interfaces: principles of conversational experiences*. " O'Reilly Media, Inc.", Sebastopol, California.

Petta, T. D. and Woloshyn, V. E. (2001). Voice recognition for on-line literacy: Continuous voice recognition technology in adult literacy training. *Education and Information Technologies*, 6(4):225–240.

Peña, P. R., Doyle, P., Edwards, J., Garaialde, D., Rough, D., Bleakley, A., Clark, L., Henriquez, A. T., Branigan, H., Gessinger, I., and Cowan, B. R. (2023). Audience design and egocentrism in reference production during human-computer dialogue. *International Journal of Human-Computer Studies*, 176:103058.

Pfau, J., Smeddinck, J. D., and Malaka, R. (2018). Towards deep player behavior models in mmorpgs. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*, pages 381–392, New York, NY, USA. ACM.

Porcheron, M., Fischer, J. E., Reeves, S., and Sharples, S. (2018). Voice interfaces in everyday life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 640:1–640:12, New York, NY, USA. ACM.

Porcheron, M., Fischer, J. E., and Sharples, S. (2017). "do animals have accents?": Talking with agents in multi-party conversation. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 207–219, New York, NY, USA. ACM.

Porzel, R. and Baudis, M. (2004). The tao of CHI: Towards effective human-computer interaction. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 209–216, Boston, Massachusetts, USA. Association for Computational Linguistics.

Pradhan, A., Findlater, L., and Lazar, A. (2019). "phantom friend" or" just a box with information" personification and ontological categorization of smart speaker-based voice assistants by older adults. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–21.

Pradhan, A., Mehta, K., and Findlater, L. (2018). "accessibility came by accident": Use of voice-controlled intelligent personal assistants by people with disabilities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–13, New York, NY, USA. Association for Computing Machinery.

Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., and Taylor, S. H. (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, page 2853–2859, New York, NY, USA. Association for Computing Machinery.

Pyae, A. and Joelsson, T. N. (2018). Investigating the usability and user experiences of voice user interface: A case of google home smart speaker. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, MobileHCI '18, pages 127–131, New York, NY, USA. ACM.

Pyae, A. and Scifleet, P. (2018). Investigating differences between native english and non-native english speakers in interacting with a voice user interface: A case of google home. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction*, pages 548–553, New York, NY, USA. ACM.

Pyro Studios (1998). *Commandos: Behind Enemy Lines*. Game [Microsoft Windows]. Eidos Interactive, Southwark, London, England.

Radzikowski, K., Nowak, R., Wang, L., and Yoshie, O. (2019). Dual supervised learning for non-native speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, 2019:1–10.

Reagans, R. (2005). Preferences, identity, and competition: Predicting tie strength from demographic data. *Management Science*, 51(9):1374–1383.

Reddy, D., Erman, L., and Neely, R. (1973). A model and a system for machine recognition of speech. *IEEE Transactions on Audio and Electroacoustics*, 21(3):229–238.

Reece, B. (2014). Putting the ha! in aha!: Humor as a tool for effective communication.

Reeves, B. and Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press, New York, NY, USA.

Reicherts, L., Zargham, N., Bonfert, M., Rogers, Y., and Malaka, R. (2021). May i interrupt? diverging opinions on proactive smart speakers. In *Proceedings of the 3rd Conference on Conversational User Interfaces*, CUI '21, New York, NY, USA. Association for Computing Machinery.

Reid, G. (2012). Motivation in video games: a literature review. *The computer games journal*, 1(2):70–81.

Romero, E. J. and Cruthirds, K. W. (2006). The use of humor in the workplace. *Academy of management perspectives*, 20(2):58–69.

Rosenberg, A., Zhang, Y., Ramabhadran, B., Jia, Y., Moreno, P., Wu, Y., and Wu, Z. (2019). Speech recognition with augmented synthesized speech. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 996–1002, New York, NY, USA. IEEE.

Roslan, F. A. B. M. and Ahmad, N. B. (2023). The rise of ai-powered voice assistants: Analyzing their transformative impact on modern customer service paradigms and consumer expectations. *Quarterly Journal of Emerging Technologies and Innovations*, 8(3):33–64.

Rotaru, M., Litman, D. J., and Forbes-Riley, K. (2005). Interactions between speech recognition problems and user emotions.

Ryan, R. M., Rigby, C. S., and Przybylski, A. (2006). The motivational pull of video games: A self-determination theory approach. *Motivation and emotion*, 30(4):344–360.

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., and Joublin, F. (2013). To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics*, 5:313–323.

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., and Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, 58(3):377–400. PMID: 27005902.

Schmidt, M. and Braunger, P. (2018). A survey on different means of personalized dialog output for an adaptive personal assistant. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*, UMAP '18, page 75–81, New York, NY, USA. Association for Computing Machinery.

Schmidt, M., Minker, W., and Werner, S. (2020). User acceptance of proactive voice assistant behavior. In Wendemuth, A., Böck, R., and Siegert, I., editors, *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2020*, pages 18–25, Dresden, Germany. TUDpress.

Schmidt, M., Stier, D., Werner, S., and Minker, W. (2019). Exploration and assessment of proactive use cases for an in-car voice assistant. In Birkholz, P. and Stone, S., editors, *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, pages 148–155, Dresden, Germany. TUDpress.

Schomaker, L. (1995). A taxonomy of multimodal interaction in the human information processing system.

Schöpf, A. C., Martin, G. S., and Keating, M. A. (2017). Humor as a communication strategy in provider–patient communication in a chronic care setting. *Qualitative Health Research*, 27(3):374–390.

Schramm, W. (1997). *The beginnings of communication study in America: A personal memoir*. Sage.

Schrills, T. and Franke, T. (2020). How to answer why–evaluating the explanations of ai through mental model analysis. *arXiv preprint arXiv:2002.02526*.

Schwind, V., Wolf, K., and Henze, N. (2018). Avoiding the uncanny valley in virtual character design. *Interactions*, 25(5):45–49.

Sciuto, A., Saini, A., Forlizzi, J., and Hong, J. I. (2018). "hey alexa, what's up?": A mixed-methods studies of in-home conversational agent usage. In *Proceedings of the 2018 Designing Interactive Systems Conference*, DIS '18, page 857–868, New York, NY, USA. Association for Computing Machinery.

Seaborn, K., Miyake, N. P., Pennefather, P., and Otake-Matsuura, M. (2021). Voice in human–agent interaction: A survey. *ACM Comput. Surv.*, 54(4).

Seyama, J. and Nagayama, R. S. (2007). The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence: Teleoperators and virtual environments*, 16(4):337–351.

Seymour, W. and Van Kleek, M. (2021). Exploring interactions between trust, anthropomorphism, and relationship development in voice assistants. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).

Shalini, S., Levins, T., Robinson, E. L., Lane, K., Park, G., and Skubic, M. (2019a). Development and comparison of customized voice-assistant systems for independent living older adults. In Zhou, J. and Salvendy, G., editors, *Human Aspects of IT for the Aged Population. Social Media, Games and Assistive Environments*, pages 464–479, Cham. Springer International Publishing.

Shalini, S., Levins, T., Robinson, E. L., Lane, K., Park, G., and Skubic, M. (2019b). Development and comparison of customized voice-assistant systems for independent living older adults. In *Human Aspects of IT for the Aged Population. Social Media, Games and Assistive Environments: 5th International Conference, ITAP 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26-31, 2019, Proceedings, Part II 21*, pages 464–479, Cham. Springer, Springer International Publishing.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Sheehan, B., Jin, H. S., and Gottlieb, U. (2020). Customer service chatbots: Anthropomorphism and adoption. *Journal of Business Research*, 115:14–24.

Shevat, A. (2017). *Designing bots: Creating conversational experiences*. " O'Reilly Media, Inc.", Sebastopol, California.

Shin, H., Bunosso, I., and Levine, L. R. (2023). The influence of chatbot humour on consumer evaluations of services. *International Journal of Consumer Studies*, 47(2):545–562.

Shrivastava, A., Raturi, A., Sharma, A., Rao, A., Chinthamu, N., and Sankhyan, A. (2023). Advancements in speech recognition technology: A cutting-edge tool for improved speech analysis and interaction. In *2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN)*, pages 1786–1792.

Shum, H.-Y., He, X.-d., and Li, D. (2018). From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19:10–26.

Sigall, H. and Landy, D. (1973). Radiating beauty: Effects of having a physically attractive partner on person perception. *Journal of Personality and Social Psychology*, 28(2):218.

Skantze, G. (2003). Exploring human error handling strategies: Implications for spoken dialogue systems.

Sporka, A. J., Kurniawan, S. H., Mahmud, M., and Slavík, P. (2006). Non-speech input and speech recognition for real-time control of computer games. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, pages 213–220, New York, NY, USA. ACM.

Sprecher, S. and Regan, P. C. (2002). Liking some things (in some people) more than others: Partner preferences in romantic relationships and friendships. *Journal of Social and Personal Relationships*, 19(4):463–481.

Springer, A. and Cramer, H. (2018). "play prblms": Identifying and correcting less accessible content in voice interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–13, New York, NY, USA. Association for Computing Machinery.

Sridhar, S. and Tolentino, M. E. (2017). Evaluating voice interaction pipelines at the edge. In *2017 IEEE International Conference on Edge Computing (EDGE)*, pages 248–251. IEEE.

Stent, A. J., Huffman, M. K., and Brennan, S. E. (2008). Adapting speaking after evidence of misrecognition: Local and global hyperarticulation. *Speech Communication*, 50(3):163–178.

Sterpu, G., Saam, C., and Harte, N. (2018). Attention-based audio-visual fusion for robust automatic speech recognition. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, page 111–115, New York, NY, USA. Association for Computing Machinery.

Suhm, B., Myers, B., and Waibel, A. (2001). Multimodal error correction for speech user interfaces. *ACM transactions on computer-human interaction (TOCHI)*, 8(1):60–98.

Sutton, S. J. (2020). Gender ambiguous, not genderless: Designing gender in voice user interfaces (vuis) with sensitivity. In *Proceedings of the 2nd Conference on Conversational User Interfaces*, CUI '20, New York, NY, USA. Association for Computing Machinery.

Swerts, M., Litman, D., and Hirschberg, J. (2000). Corrections in spoken dialogue systems.

Tabassum, M., Kosiński, T., Frik, A., Malkin, N., Wijesekera, P., Egelman, S., and Lipford, H. R. (2019a). Investigating users' preferences and expectations for always-listening voice assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(4):1–23.

Tabassum, M., Kosiński, T., and Lipford, H. R. (2019b). "i don't own the data": End user perceptions of smart home device data practices and risks. In *Proceedings of the Fifteenth USENIX Conference on Usable Privacy and Security*, SOUPS'19, page 435–450, USA. USENIX Association.

Takeuchi, A. and Naito, T. (1995). Situated facial displays: Towards social interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '95, page 450–455, USA. ACM Press/Addison-Wesley Publishing Co.

Teng, C.-I. (2010). Customization, immersion satisfaction, and online gamer loyalty. *Computers in Human Behavior*, 26(6):1547–1554.

Terzopoulos, G. and Satratzemi, M. (2020). Voice assistants and smart speakers in everyday life and in education. *Informatics in Education*, 19(3):473–490.

Turunen, M., Hakulinen, J., Raiha, K.-J., Salonen, E.-P., Kainulainen, A., and Prusi, P. (2005). An architecture and applications for speech-based accessibility systems. *IBM Systems Journal*, 44(3):485–504.

Vieira, M. F. G., Fu, H., Hu, C., Kim, N., and Aggarwal, S. (2014). Powerfall: a voice-controlled collaborative game. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play*, pages 395–398, New York, NY, USA. ACM.

Virkkunen, A. et al. (2018). Automatic speech recognition for the hearing impaired in an augmented reality application.

Völkel, S. T., Buschek, D., Eiband, M., Cowan, B. R., and Hussmann, H. (2021). Eliciting and analysing users' envisioned dialogues with perfect voice assistants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Völkel, S. T. and Kaya, L. (2021). Examining user preference for agreeableness in chatbots. In *CUI 2021 - 3rd Conference on Conversational User Interfaces*, CUI '21, New York, NY, USA. Association for Computing Machinery.

Völkel, S. T., Kempf, P., and Hussmann, H. (2020). Personalised chats with voice assistants: The user perspective. In *Proceedings of the 2nd Conference on Conversational User Interfaces*, CUI '20, New York, NY, USA. Association for Computing Machinery.

Wagener, N., Reicherts, L., Zargham, N., Bartłomiejczyk, N., Scott, A. E., Wang, K., Bentvelzen, M., Stefanidi, E., Mildner, T., Rogers, Y., and Niess, J. (2023). Selvreflect: A guided vr experience fostering reflection on personal challenges. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Wan, S. (2021). Research on speech separation and recognition algorithm based on deep learning. *2021 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS)*, pages 722–725.

Wang, I., Smith, J., and Ruiz, J. (2019). Exploring virtual agents for augmented reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, New York, NY, USA. Association for Computing Machinery.

Wang, L., Fazel-Zarandi, M., Tiwari, A., Matsoukas, S., and Polymenakos, L. (2020). Data augmentation for training dialog models robust to speech recognition errors. *arXiv preprint arXiv:2006.05635*.

Watson, J. and Hill, A. (2015). *Dictionary of media and communication studies*. Bloomsbury Publishing USA.

Waytz, A., Epley, N., and Cacioppo, J. T. (2010). Social cognition unbound: Insights into anthropomorphism and dehumanization. *Current Directions in Psychological Science*, 19(1):58–62.

Wei, J., Dingler, T., and Kostakos, V. (2021). Developing the proactive speaker prototype based on google home. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–6, New York, NY, USA. Association for Computing Machinery.

Wei, J., Tag, B., Trippas, J. R., Dingler, T., and Kostakos, V. (2022). What could possibly go wrong when interacting with proactive smart speakers? a case study using an esm application. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Wiedmann, K.-P. and Von Mettenheim, W. (2020). Attractiveness, trustworthiness and expertise–social influencers' winning formula? *Journal of Product & Brand Management*, 30(5):707–725.

Wilcox, T., Evans, M., Pearce, C., Pollard, N., and Sundstedt, V. (2008). Gaze and voice based game interaction: the revenge of the killer penguins. *SIGGRAPH Posters*, 81(10.1145):1400885–1400972.

Wilson, C. (2013). *Interview techniques for UX practitioners: A user-centered design method*. Elsevier, Waltham, Massachusetts, USA.

Winkler, N., Röthke, K., Siegfried, N., and Benlian, A. (2020). Lose yourself in vr: exploring the effects of virtual reality on individuals' immersion.

Winkler, R., Söllner, M., Neuweiler, M. L., Conti Rossini, F., and Leimeister, J. M. (2019). Alexa, can you help us solve this problem?: How conversations with smart personal assistant tutors increase task group outcomes. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, pages LBW2311:1–LBW2311:6, New York, NY, USA. ACM.

Wolters, M., Georgila, K., Moore, J. D., Logie, R. H., MacPherson, S. E., and Watson, M. (2009). Reducing working memory load in spoken dialogue systems. *Interacting with Computers*, 21(4):276–287.

Woods, S., Dautenhahn, K., Kaouri, C., Boekhorst, R., and Koay, K. L. (2005). Is this robot like me? links between human and robot personality traits. In *5th IEEE-RAS International Conference on Humanoid Robots, 2005.*, pages 375–380. IEEE.

Yam, K. C., Barnes, C. M., Leavitt, K., Wei, W., Lau, J., and Uhlmann, E. L. (2019). Why so serious? a laboratory and field investigation of the link between morality and humor. *Journal of Personality and Social Psychology*, 117(4):758.

Yao, S. and Kim, G. (2019). The effects of immersion in a virtual reality game: Presence and physical activity. In Fang, X., editor, *HCI in Games*, pages 234–242, Cham. Springer International Publishing.

Yeh, S.-F., Wu, M.-H., Chen, T.-Y., Lin, Y.-C., Chang, X., Chiang, Y.-H., and Chang, Y.-J. (2022). How to Guide Task-oriented Chatbot Users, and When: A Mixed-methods Study of Combinations of Chatbot Guidance Types and Timings. In *CHI Conference on Human Factors in Computing Systems*, pages 1–16, New Orleans LA USA. ACM.

Yuan, L. and Dennis, A. R. (2019). Acting like humans? anthropomorphism and consumer's willingness to pay in electronic commerce. *Journal of Management Information Systems*, 36(2):450–477.

Yue, X., Jiang, F., Lu, S., and Hiranandani, N. (2016). To be or not to be humorous? cross cultural perspectives on humor. *Frontiers in psychology*, 7:1495.

Yuksel, B. F., Collisson, P., and Czerwinski, M. (2017). Brains or beauty: How to engender trust in user-agent interactions. *ACM Trans. Internet Technol.*, 17(1):2:1–2:20.

Zamfirescu-Pereira, J., Wei, H., Xiao, A., Gu, K., Jung, G., Lee, M. G., Hartmann, B., and Yang, Q. (2023). Herding ai cats: Lessons from designing a chatbot by prompting gpt-3.

Zargham, N., Avanesi, V., Reicherts, L., Scott, A. E., Rogers, Y., and Malaka, R. (2023a). "funny how?" a serious look at humor in conversational agents. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, CUI '23, New York, NY, USA. Association for Computing Machinery.

Zargham, N., Bonfert, M., Porzel, R., Doring, T., and Malaka, R. (2022a). Multi-agent voice assistants: An investigation of user experience. In *Proceedings of the 20th International Conference on Mobile and Ubiquitous Multimedia*, MUM '21, page 98–107, New York, NY, USA. Association for Computing Machinery.

Zargham, N., Fetni, M. L., Spillner, L., Muender, T., and Malaka, R. (2023b). "i know what you mean": Context-aware recognition to enhance speech-based games. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Zargham, N., Pfau, J., Schnackenberg, T., and Malaka, R. (2022b). "i didn't catch that, but i'll try my best": Anticipatory error handling in a voice controlled game. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Zargham, N., Reicherts, L., Avanesi, V., Rogers, Y., and Malaka, R. (2023c). Tickling proactivity: Exploring the use of humor

in proactive voice assistants. In *Proceedings of the 22nd International Conference on Mobile and Ubiquitous Multimedia*, MUM '23, page 294–320, New York, NY, USA. Association for Computing Machinery.

Zargham, N., Reicherts, L., Bonfert, M., Voelkel, S. T., Schoening, J., Malaka, R., and Rogers, Y. (2022c). Understanding circumstances for desirable proactive behaviour of voice assistants: The proactivity dilemma. In *Proceedings of the 4th Conference on Conversational User Interfaces*, CUI '22, New York, NY, USA. Association for Computing Machinery.

Zdenek, S. (2007). "just roll your mouse over me": Designing virtual women for customer service on the web. *Technical Communication Quarterly*, 16(4):397–430.

Zgank, A. and Kacic, Z. (2012). Predicting the acoustic confusability between words for a speech recognition system using levenshtein distance. *Elektronika ir Elektrotechnika*, 18(8):81–84.

Zhao, R., Wang, K., Divekar, R., Rouhani, R., Su, H., and Ji, Q. (2018). An immersive system with multi-modal human-computer interaction. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 517–524, New York, NY, USA. IEEE, IEEE.

Zhou, M. X., Mark, G., Li, J., and Yang, H. (2019). Trusting virtual agents: The effect of personality. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 9(2-3):1–36.

Ziółko, B., Gałka, J., Jadczyk, T., and Skurzok, D. (2010). Modified weighted levenshtein distance in automatic speech recognition. In *Proceedings of the XVI National Conference Applications of Mathematics to Biology and Medicine*, pages 116–120. Citeseer.

Zlotowski, J. and Bartneck, C. (2013). The inversion effect in hri: Are robots perceived more like humans or objects? In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 365–372, Manhattan, New York City. IEEE, IEEE.