INCLUSIVE INTERACTION DESIGN OF VOICE ASSISTANTS FOR OLDER ADULTS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Andrea Paola Cuadra

May 2022

INCLUSIVE INTERACTION DESIGN OF VOICE ASSISTANTS FOR OLDER ADULTS

Andrea Paola Cuadra, Ph.D.
Cornell University 2022

We do not all have equal voices in the design decisions that affect us. Often, the institutions that make these decisions are not committed to designing for members of marginalized groups who may particularly benefit from or be harmed by the outcomes of these decisions. Although voice assistants are a potentially life-altering technology for some, they currently exist principally as luxury goods. The needs and values of marginalized groups for whom voice assistants may be more impactful have received relatively little attention. Among these underrepresented voices in the design process of voice assistants are older adults (age 65 and older), who are at higher risks of experiencing cognitive, functional, visual, hearing, emotional, social, and other impairments, who are often not experienced technology users, and who may depend on caregivers.

This dissertation adapts existing concepts from inclusive design to the interaction design of voice assistants, contributing four studies that advocate for improving their inclusivity (and safety). Two of these are empirical studies that explore the use of voice assistants by older adults from two different angles: 1) an approximately two-month long intervention studying private-setting interactions with users who become experienced over the duration of the study, and 2) a close look (second by second) at public-setting interactions with mostly novice users.

Together, these two studies examine communication gaps that may lead to exclusion and inform design strategies to increase the inclusivity of voice assistants for older adults (e.g., by integrating into healthcare systems, and by reacting to audio-prosodic and visual cues). The third study explores building voice assistants that perform self-repair as a way to address communication breakdowns such as those seen in the first two studies. We demonstrate that voice assistants that perform self-repair in response to visual cues improve interactions, even after factoring in instances in which these visual cues are not appropriately interpreted. Finally, the fourth study grapples with ethical concerns by developing a framework to reflect on the role of empathy in these interactions and mitigate their potential risk of harm. My studies were carefully designed to complement each other—together they show the power and perils of voice assistants.

As a whole, I (i) examine how older adults currently interact with voice assistants; (ii) evaluate alternative designs that could accelerate the development of inclusive voice assistant features; and (iii) develop a framework to mitigate potential harms of empathetic conversational user interfaces. Taken together, this work aspires to relieve some of the burdens placed on underrepresented individuals to adapt to the impositions of digital technologies, serves to promote the inclusion of marginalized groups in mainstream technological activities that can increase justice and equity, and warns us about potential risk of harm.

BIOGRAPHICAL SKETCH

Andrea Paola Cuadra wrote this dissertation as an Information Science PhD student at Cornell Tech under the advisorship of Deborah Estrin, Nicola Dell, Malte Jung, and external committee member, Amon Millner. Previously, she earned a master's degree in product design from Stanford University, where she was advised by David Kelley and Bill Burnett, and her bachelor's degree in engineering with a concentration in interaction design from Olin College, where she was advised by Amon Millner.

Her work seeks to characterize the design decisions needed to maintain inclusive participation in this era of rapid technological advancements. She advocates for technologies to be designed in ways that take into account the needs of communities that may not have equal voices in the design process, and for these technologies to have paths for re-design that take into account unexpected effects.

In addition to pursuing her graduate education, she spent several summers working in industry at: Google on Chrome and Search, Sidewalk Labs on Delve and Mesa, and Yahoo on Flurry Analytics. She also spent a few exciting post-college years as an independent inventor of toys and games. During her invention journey, she enjoyed giving a TEDx talk on creativity, self-publishing an interactive children's book and a board game, and building talking potato robots. Before moving to the U.S. to attend college, Andrea grew up swimming and hunting for stones in the fresh waters of Lake Nicaragua.

To Alexa Lempel. I admire your kindness.

ACKNOWLEDGEMENTS

Before starting my Ph.D. program, I has the (mis)perception that pursuing a Ph.D. would be a lonely endeavor. In reality, it has been the opposite. Even though the name on this dissertation is mine, the work in this dissertation was made possible thanks to many others.

Deborah Estrin, Nicola Dell, Malte Jung, and Amon Millner, I will be forever grateful to you for having accompanied and guided me in this exciting journey of exploration and discovery as a stellar Ph.D. committee.

Lynn Stein, thank you for stepping in as a pseudo advisor during the last year of my Ph.D. and introducing me to ideas and possibilities that illuminated this work.

Dr. Armin Shahrokni, it was an honor to have worked with you to develop the voice electronic rapid fitness assessment. Thank you for opening up the doors to the medical world for me, and for your mentorship and support.

Cheryl Spector, I am so thankful for your expert editing skills and cheerleading. My communication skills (and confidence!) have improved greatly thanks to your coaching.

Jessica Bethune, I am so lucky to have a brilliant friend who gets excited about helping me with writing and coding. Your support has turned what could have been frustrating into fun.

Priscilla Sevilla, from the start to the end, you were there, consistently celebrating the good moments and accompanying me through the bad ones. ¡Gracias, amiga!

Paulina Cuadra, thanks for proofreading my papers before you even graduated from high school!

Ana Miranda and Carmen Deshon, thank you for being my models and role models.

Infinite thanks to my labmates, mentors, collaborators, and student advisees. Your support, effort, contributions, and generosity have meant so much to me.

I am also immensely grateful for the older adults who agreed to participate in my research because they knew it would help me write this dissertation.

Thanks to my parents for instilling the value of education in me, and to my extended family for your constant support.

I made the mistake of putting off writing the acknowledgements to the very last minute out of fear that I would miss mentioning someone. I most certainly have. If you are reading this, and you played a role, thank you!

It is important to note that this dissertation was written while being a student at Cornell Tech, which is located in Roosevelt Island, New York City. The land that is now Roosevelt Island used to belonged to the Canarsies before settler colonialism changed its identity and sovereignty. Moreover, Cornell Tech is part of Cornell University, which is located on the traditional homelands of the Gayogo̱hó:nǫ’ (the Cayuga Nation)[1]. The Gayogo̱hó:nǫ’ are members of the Haudenosaunee Confederacy, an alliance of six sovereign Nations with a historic presence on said land. The Confederacy precedes the establishment of Cornell University, New York State, and the United States of America. I acknowledge the painful history of Canarsie and Gayogo̱hó:nǫ’ dispossession, and honor the ongo-

[1]This is the closest spelling to the one in `https://gayogohono.org/` as I could achieve without creating PDF compilation errors. However, I hope that soon the LaTex compilers will be more inclusive of more spellings!

ing connection of Canarsie and Gayogo̱hó:nǫ’ people, past and present, to these lands and waters. I encourage all who engage with Cornell University to learn more about the Carnarsie and the Gayogo̱hó:nǫ’, their history, and people, and to take meaningful action to support indigenous scholars and their communities.

My research was supported by NSF Awards #2026577 and #1700832, the Mozilla Foundation, and a Digital Life Initiative Doctoral Fellowship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Inclusive design brings the needs of the people at the margins[1] to the design of mainstream products and services [104]. However, often, the institutions that make these decisions are not committed to representing the needs of people from marginalized groups, who may particularly benefit from or be harmed by the outcomes of these decisions, in the products they build. For example, voice assistants, which are predominantly used for music, hands-free search, and to control smart home devices [48], have tremendous potential for increasing access to healthcare for people with low literacy, and low digital literacy. Despite this potential, work in the space has received relatively little attention.

This dissertation focuses on voice assistants, which are becoming mainstream, but could be much more inclusive. Voice assistants rely on several types of AI[2], including natural language processing and generation, and machine learning. Siri, the first voice assistant available to the mass market, was introduced merely 10 years ago (2011) as an iPhone feature. Many voice assistants compete in today's market (e.g., Amazon Alexa—released in 2014, and Google Assistant—released in 2016), exist in an increasing number of devices (smart speakers, cars, televisions, watches, cameras, and so on), allow third-party developers to use their plat-

[1] I use a working definition of marginalization written by Alakhunova et al. [41]: **Marginalization** is both a condition and a process that prevents individuals and groups from full participation in social, economic, and political life enjoyed by the wider society.

[2] **AI** is intelligence demonstrated by machines, as opposed to natural intelligence displayed by animals including humans [308].

forms to build voice applications, and are projected to continue becoming more widespread in the years to come[3]. Their relatively early stage of diffusion and relatively unequal distribution[4] makes them an important technology to study through a lens of inclusivity. Designed with care, voice assistants could play an important role in integrating the needs of marginalized groups into mainstream products and services that would otherwise be unavailable to them [147, 114]. In the same vein, as conversational user interfaces (CUIs) become increasingly pervasive, we must keep in mind the challenges introduced by their humanlikeness and potential deception. While more humanlike CUIs can be more inclusive of people new to digital technologies by being more intuitive to use, they also evoke emotions that can make people act in ways that are not in their best interests. This increases risks of harm, in particular for people who may be in more vulnerable positions, creating an urgent need for a justice-oriented perspective.

As part of practicing and promoting inclusive design, I designed my studies to focus on understanding older adult interactions with voice assistants and on addressing older adult needs that arose. Note, I use the term "older adult" in keeping with the American Psychological Association guidelines[5], and consider anyone age 60 or older (the requirement for senior center membership in New York City[6]) an older adult. This said, though defining age chronologically is the standard in academic literature, there are other, potentially more appropriate,

[3]www.reportlinker.com/p06033265/Global-Voice-Assistant-Application-Industry.html
[4]https://review42.com/resources/voice-search-stats/
[5]https://apastyle.apa.org/style-grammar-guidelines/bias-free-language/age
[6]https://www1.nyc.gov/site/dfta/services/senior-centers.page

approaches for defining age, such as functional or performance-based age, psychosocial or subjective age, organizational age, and the life span concept of age [196]. I chose older adults for many reasons. First, they are at higher risk of being pushed to the margins as they begin to experience cognitive, functional, visual, hearing, emotional, social, and other impairments [119]. Addressing the needs of older adults experiencing impairments helps others as well. Most people are likely to experience impairments in one way or another temporarily or permanently at some point in their lives [104]. So too, many people who are not older adults can experience similar challenges for different reasons. For example, over forty percent of the world population does not actively use the internet [7]. Children, young adults, and middle-aged adults who do not actively use the internet may face similar challenges navigating unfamiliar digital interfaces as do older adults, who tend to have less experience with digital technologies [119]. Additionally, older adults, like other marginalized groups (e.g., Black people, Indigenous people, and women) are underrepresented in the higher ranks of powerful technology companies [108], which means that their needs are less likely to be understood and prioritized within those companies' products. Finally, older adults represent a large (and growing) demographic segment[8] at an increased risk of being marginalized by technology. If addressed well, inclusive mainstream solutions will benefit other users as well—for example, a hands-free device designed with people with motor impairments in mind can also benefit someone with oc-

[7] https://www.statista.com/statistics/617136/digital-population-worldwide/

[8] https://www.census.gov/newsroom/press-releases/2018/cb18-41-population-projections.html

cupied hands, such as someone who is driving.

In this dissertation, I explore voice assistants with older adults in two particular contexts: in a private setting with users who transition from novice to more-experienced users, and in a public setting with mostly novice users. Then, I propose two particular ways to move forward designing inclusively: by building voice assistants that perform self-repair[9], and by developing a framework to systematically and generatively reflect on role of empathy in human-CUI interactions. I contribute the following four studies towards increasing the inclusivity of voice assistants:

In the first study, I explore the promise of two prototype voice apps for older adults, one for health data reporting and the other for positive reminiscing, in the privacy of their homes. In this study, we deliver Amazon Echo Shows with Alexa to the homes of five older adults, provided appropriate training, tracked usage, and deployed the prototype voice apps (as design probes) to their devices. Participants learned to use voice assistants, and saw value in the prototype voice apps. Reminders and the display were helpful supports. We surface design challenges entailing the use of voice assistants for health data reporting, such as participants' perception that they would be bothering their doctors. We then provide design guidelines—such as the need to balance expectations of humanlike interac-

[9]Repair is the conversational analysis term for when interactants try to fix problems in speaking, hearing or understanding that come up during conversation [282, 100, 315, 309]. Self-repair is repair by the speaker of that which is being repaired instead of repair by another interactant [319]. Hofstetter provides a thorough video explanation of repair here: `http://emcawiki.net/Repair`.

tion affordances and the technology's limited capabilities—for developing voice apps tailored to older adults, and discuss how voice assistants might enable continuity of care in people's own homes.

In the second study, I explore the use of voice assistants through a study that looks at older adult first-time encounters with voice assistants in a public setting. In this study, we video record older adults (*n*=26) interacting with a multimodal voice assistants while waiting in line at food pantries, and use Interaction Analysis to draw insights from these recordings. We find that by being agnostic to body language, audio-prosodic features, and other contextual factors, voice assistants fail to capture and react to some important aspects of interactions. Voice assistants did not detect when participants were leaning forward to speak to the device or were not able to tell that the voice assistant was not listening to what they were saying. We discuss design (e.g, interpreting users' posture as a cue to wake the device when they are leaning towards the device, or using older design paradigms) and research (e.g., surveillance trade-offs) implications, and argue for the use of multi-modal inputs with attention to privacy.

In the third study, I propose building voice assistants that perform self-repair as a way to address communication breakdowns such as those seen in the first two studies. We investigate how the self-repair of errors by voice assistants affects user interaction. To do so, we conduct a laboratory between-participants experiment (*N*=101), using four conditions in a 2 (presence of mistake: no mistakes made vs. mistakes made) $\times$ 2 (presence of repair: no repair performed vs. repair performed) matrix. We measure the impact of self-repair on the participant's perception of the interaction. We find that self-repair greatly improves people's assessment of an intelligent voice assistant if a mistake has been made, but can degrade assessment if no correction is needed. However, we find that the positive impact of self-repair in the wake of an error outweighs the negative impact of overcorrection. In addition, participants who recently experienced an error saw increased value in self-repair as a feature, regardless of whether they experienced a repair themselves. Based on participants' responses to our interview questions, we also raise ethical concerns surrounding continuing to normalize surveillance and the implications of human-mimetic interaction patterns. Note, we explored this through a different population (university students in their late teens or twenties), because doing so with older adults, who may be more vulnerable, is fraught at this early stage of experimentation and development. Future work should investigate whether self-repair also improves older adult interactions with voice assistants.

In the fourth study, I propose grappling with ethical concerns by developing a framework to systematically and generatively reflect on the role of empathy in human-CUI interactions. In this contribution, we argue that although humanlike empathy can help technology better serve human needs, it can also be deceptive and potentially exploitative. We introduce the notion of the *Illusion of Empathy* to distinguish evocations of empathy between two humans from ones between a human and a CUI. We introduce a framework to reflect on these evocations from an ethical perspective for harm mitigation. Our aim is to solidify and clarify current understanding of evocations of empathy in interactions with CUIs, and to develop mitigation strategies that allow us to benefit from the promise of empathetic CUIs while mitigating their risk of harm.

1.1 Contributions

My dissertation adapts existing concepts from inclusive design to the interaction design of voice assistants, contributing four studies that advocate for improving their inclusivity (and safety). In addition to the individual contributions described in the four studies, this dissertation also makes the following high-level contributions:

Contribution 1: **(A)** It examines communication gaps between older adults and voice assistants that might lead to exclusion. **(B)** In addressing communication gaps, it demonstrates that a feature needed by older adults (i.e., voice assistant self-repair) was shown to improve interactions for younger adults.

Contribution 2: **(A)** It draws on in-depth descriptions from three sets of empirical data to suggest human-centered design strategies—such as employing human-mimetic interaction patterns, improving explainability, and reducing the ambiguity of information flow—that could improve inclusion. **(B)** Towards improving inclusion, I also build and evaluate two prototype voice apps, one for health data reporting and the other for positive reminiscing, specifically tailored to meet older adults' needs and preferences.

Contribution 3: It identifies an area of tension—entailing the humanlikeness of CUIs and the risk associated with it—in the inclusive design of CUIs that I call the Illusion of Empathy. I develop a framework to navigate this tension, opening a new area for future research.

In this dissertation, I strive to delineate a path towards design that helps improve the inclusivity of voice assistants. This said, achieving inclusion is a "wicked" problem [303], which is difficult or impossible to solve because of incomplete, contradictory, and changing requirements that are often difficult to recognize. The approaches I propose have many limitations, including the limitations highlighted in each of the contributed studies (e.g., small scale, qualitative, and/or conducted in an urban setting), and biases stemming from my own positionality as a middle-aged graduate student living in the U.S. Despite these limitations, this work provides a new justice-oriented lens from which to look at humanlikeness and empathy in human-CUI interactions that will help us realize the promises of CUIs while mitigating their risk. I conclude by discussing these contributions in terms of my findings and future work that could add more breadth and depth.

CHAPTER 2

RELATED WORK

In this chapter, I first describe what inclusive interaction design is. Then I explain the current state of voice assistant technology (covering some of its past, present, and potential future). I conclude by arguing that the use of voice assistants by older adults is an area of research that urgently needs more attention.

2.1 Inclusive Interaction Design

In this work, I will be referring to inclusive interaction design as the design of mainstream products that includes the voices, needs, and requirements of those currently underrepresented. Two widely referenced textbooks on inclusive design are a European one titled *Inclusive design: Design for the whole population* [104], and an American one titled *Universal design: Creating inclusive environments* [333]. These textbooks broadly describe inclusive design as design that is aimed at including as many people as possible without the need for specialized systems. They both describe inclusive design as an evolving process that expands as our understanding of the needs of a diverse population grows. Older adults are frequently used as a canonical target population for inclusive design [102, 333, 356, 178, 366], as our human abilities tend to change with age. However, inclusive design also accounts for multiple demographies at the margins [333, 128, 216, 286, 171]. In 2020, two films were released that are relevant to inclusive design, *Crip Camp: A Disability Revolution* [10], and *Coded Bias* [8]. *Crip Camp: A Disability Revolution* tells the story of many activists from the disability rights movement from which inclusive design emerged, and *Coded Bias* foreshadows the move of inclusive design from the physical to the digital space by narrating the story of M.I.T. Media Lab researcher Joy Buolamwini as she uncovered flaws in facial recognition technology. In this dissertation, I will principally focus on the inclusive interaction design of voice assistants for older adults.

2.2 What Are Voice Assistants?

Voice assistants can be considered a form of AI, AI-based agents, AI agents, autonomous agents, and so on. AI is considered to be intelligence demonstrated by machines, as opposed to natural intelligence displayed by animals including humans [308]. There are some research efforts aimed at general AI to approach the sophistication of natural intelligence [357, 99]. However, recent trends have advanced narrow AI systems. These focus on one type of task in ways that can be perceived to be intelligent under certain circumstances [103, 160, 72, 343].

Voice assistants can perform tasks upon request—they tend to (by default) come with voices that sound like women, use natural language processing to derive intent from requests made by their users, and respond to those requests using built-in functionality or features built by third parties [372, 328]. Voice assistants first transcribe spoken words to text, then derive meaning from the text, and last respond using speech, and/or another modality. The most popular voice assistants in the United States today are Amazon's Alexa, the Google Assistant, and Apple's Siri. Some of the other, less-popular, voice assistants include Microsoft's Cortana, Samsung's Bixby, IBM's Watson Assistant, and Mycroft AI's Mycroft.

This chapter will show how the design of voice assistants do not always reflect the values and perspectives of marginalized groups, possibly to the detriment of their true potential. In the following sections, I will cover voice assistant embodiment, and multi-modalities. Subsequently, I will provide more background on the origins and state-of-the-art of voice assistants. Then, I will discuss the illusion of empathy. Finally, I will discuss why this work focuses on voice assistants for older adults.

2.2.1 Voice Assistant Embodiment

Because voice assistants are software agents, they must be embodied via hardware. The physical devices that house voice assistants may be principally designed for the voice assistant, like the Amazon Echo or Google Home smart speakers (which are frequently referred to by the name of the voice assistant they house—e.g. in reference to the Amazon Echo, people tend to say "ask the Alexa for the weather"), or multi-use devices such as laptop computers or smartphones [2, 12, 27]. Additionally, there are many devices, such as light bulbs, electrical plugs, locks, or vacuum cleaners, that can be connected to a device that houses a voice assistant via supporting software applications, many times requiring another device such as a smartphone or tablet [36, 15, 1]. Note, this need for an additional software application may create a barrier for users with modest digital literacy, and for users without the device necessary to download such an application (these devices tend to be much more expensive than the cheapest voice assistant speakers). Figure 2.1 shows embodiments and integrations by each of the leading voice assistants in today's market. Amazon, Google, and Apple all have screenless smart speakers that house their respective voice assistant. Amazon and Google also have speakers with screens (and cameras) that house Alexa and the Google Assistant. Apple and Google have assistants that come built-in with their mobile devices. All three companies have mobile applications that can be used to connect their respective voice assistants to compatible devices that do not house a voice assistant themselves. In addition, an increasing number of products are being released with built-in voice assistants. For example, all three voice assistants previously mentioned can come pre-installed in TVs and cars. Additionally, Apple and Google sell watches with Siri and the Google Assistant, respectively. Moreover, Facebook's Portal devices have a built-in Alexa [28], and Google's Nest Cam IQ comes with the Google Assistant [22]. Note, this is not a comprehensive list of the different embodiments and integrations of voice assistants. The embodiment of a voice assistant can dictate its role and its importance in a particular setting. A voice assistant speaker with a screen in a kitchen will likely be principally used for recipes or timers, and a screenless voice assistant in a car will likely be used to ask for directions, to make song requests, or to text while driving. In the

Figure 2.1: This table depicts the three popular commercial voice assistants and some of their different embodiments—some have screens, some come built-in in smartphones, all of them have designated mobile applications that can be used to connect to other devices, and all of them come built in to a variety of other products permeating the market.

future, voice assistants specifically designed for the home health of older adults may be connected to health sensors and may have bigger screens for larger font sizes.

2.2.2 Voice Assistant Multi-Modality

Many voice assistants are multi-modal, giving designers many possibilities to adapt them to the needs of specific user groups. The principal modality of voice assistants is speech, they are listen to and respond in spoken language; however, this modality rarely exists in isolation. Even the simplest, screenless smart speakers (i.e. the Google Home Mini [13] or the Amazon Echo Dot [4]) have visual outputs such as lights—to serve as conversational visual cues—, and tactile inputs such as buttons—to mute, control the volume, or reset the device. More sophisticated devices, such as the Apple Homepod [18], can sense the space where it is located to optimize sound output, taking space as input. Devices with screens can sense the amount of light in a room, and adjust their screen brightness accordingly, and by virtue of having a screen also have the modalities associated with such [14, 5]. Some of these modalities include text, or images. Similarly, devices with cameras can take images as input [5]. In the case of being connected to a robotic vacuum cleaner [35] or a lock [23], the voice assistant can take speech as input and produce mechanical movement as its output. Any hardware device (i.e., pet robots, health monitors, wearable devices...) could be designed to have a voice assistant built in or connected to it. For example, there are several commercialized social robots, such as Jibo and Cozmo, that were built with voice user interfaces similar to voice assistants [275]. From an interaction perspective, this means that the voice assistant would also have its hardware counterpart's characteristics, opening up a vast array of design possibilities for creating voice-enabled, multi-modal objects.

2.3 Background on Voice Assistants

Voice assistants are the synthesis of many technological advances that have occurred over many decades. Even though the start of the commercial proliferation of voice assistants was marked by the release of Siri as a built-in feature on the iPhone 4s in 2011 [37], voice assistants have existed in science fiction for decades [92, 140], a strong indicator of their potential for impact. This imagined utility of voice assistants in our potential future is extremely useful for understanding the current ways people comprehend the values of voice assistants, and hence understanding voice assistants in science fiction helps understand the values behind voice assistants in real life.

2.3.1 Voice Assistants in Science Fiction

Science fiction can foreshadow the future of technology, and voice assistants fall into this category of foreshadowed technology as they have indeed been imagined in our popular media for far longer than they have existed in real life (see Figure 2.2). The 1962 futuristic cartoon show *The Jetsons* [164] featured many technological innovations that are close to present-day products—one of those was Rosey the Robotic Maid who is a humanoid robot able to recognize and produce speech, perform household tasks, and snappily respond in social situations (e.g. when she replies to the robotic made salesman—who had been spelling out words to insult her, presumably thinking that she would not understand if he did not say

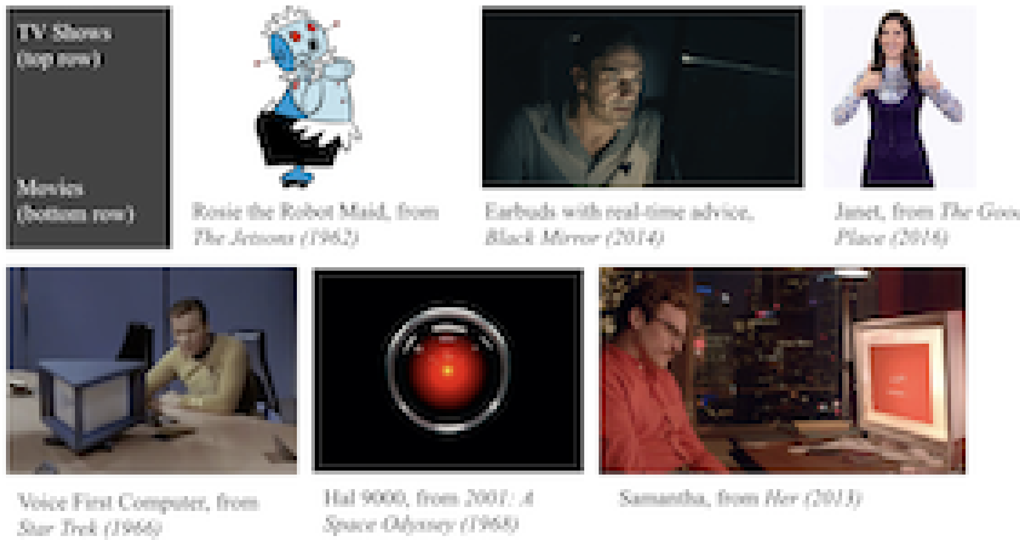

Figure 2.2: Science fiction can foreshadow the future of technology. This image portrays some voice assistant-like characters from science fiction TV shows (top row) and movies (bottom row). It is worthy to note what this is foreshadowing about voice assistant gender roles. The female voice assistants are Rosey: a house maid; Janet, a perfect assistant; and Samantha, a love interest. Meanwhile, the male voice assistants are depicted as more powerful, sometimes threatening, decision-making or manipulation agents.

the full word—by saying, "I may be homely, buster, but I'm smart. S.M.A.R.T"). Today's voice assistants perform similar functions to Rosey, but most of them are embodied in speakers about the size of tissue boxes, not in human-size, human-like robots. One can ask a voice assistant to activate the robotic vacuum which will autonomously clean the floor, and a voice assistant is likely to make snarky remarks if one accidentally calls it by its competitor's name.

There are many examples of aspects of voice assistants, or voice assistants themselves, in science fiction. In the "White Christmas" episode (2014) of the contemporary TV show *Black Mirror*, men remotely watch the lives of others while whispering advice into their ear—if one substitutes the human for an automated machine, one ends up with a voice assistant-powered earbud giving real-time life advice. This substitution is not too far off given humans are currently being used to impersonate and thus train voice assistants [92]. Voice assistants are also alluded to in the lighthearted TV show *The Good Place*, which features Janet, a cheery character that can be summoned on-demand, just like Siri, Alexa, and the Google Assistant, and assists other characters (although never an older adults with disabilities) by providing requested information and objects. In the show, there is also a version of Janet that is bad, Bad Janet who lives in The Bad Place. By creating "good" Janet's foil, Bad Janet, the creators of the show recognized both the power and perils of an assistant like Janet. Finally, the Voice-First Computer in *Star Trek (1966)*, HAL 9000 from *2001: A Space Odyssey (1968)* [17], and Samantha, the name of the operating system that the main character, a man, falls in love with in the movie *Her (2013)* [184], are all fictional voice assistants featured in movies that more directly match the voice assistants described in this work. Notice the values reflected in the role of voice assistants portrayed in popular media, and what these values are foreshadowing about the future. For example, drawing from a feminist HCI perspective [58], and analyzing the gender roles portrayed, we may observe that the female voice assistants are Rosey: a house maid; Janet, a perfect assistant; and Samantha, a love interest. Similarly, all of these assistants are predominantly designed to ultimately serve the goals of men. Rosey supports a relatively wealthy woman by doing house chores for a family headed by a man, Janet's principal user is a male character who designs the neighborhood where the story takes place, and Samantha becomes the love interest of a man. Meanwhile, the male voice assistants are depicted as more powerful, sometimes threatening, decision-making or manipulation agents. Moreover, the main users of the male voice assistants are rarely female, and even less likely to be serving the goals of women. These stories are lacking empowering representations of the women. Gender is used here as a straightforward (yet limited as gender is not binary) example, but similar analyses could be made based on other underrepresented identities. The messages being amplified by these characters underrepresent the voices and perspectives of marginalized groups, and signal towards a future where technology replicates harmful inequalities.

This is merely a small selection of voice assistants in science fiction, in a late-breaking work piece presented at CHI in 2019, Cheng et al. systematically analyzed the role of voice assistants in science fiction by consulting an online movie club and conducting a workshop where they identified 31 movies featuring 43 voice assistants [92]. In this systematic review that considers each voice assistant in detail there is no mention of highly important identity markers representative of large portions of our global population, such as race, disability, or older age. Also, not unlike most research on new technologies, the participant pool did not include older adults. This example represents one of many under-addressed opportunities towards mitigating disparities. A method to do so is by applying more-inclusive theoretical lenses, such as critical race theory [271], post-colonial

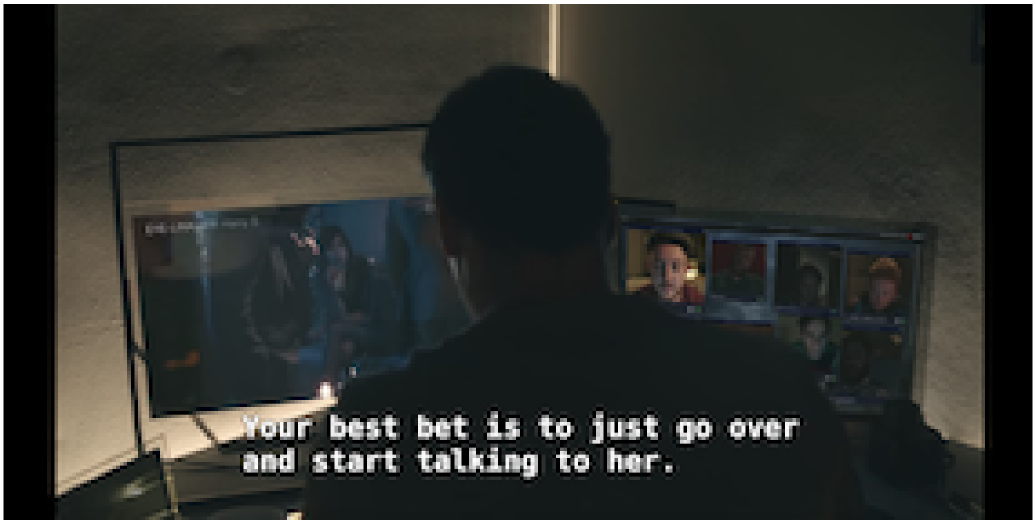

Figure 2.3: Screenshot from *Black Mirror's* "White Christmas" episode. The back of the main character is shown as he and other men (depicted in the screen on the right) are advising another man whose every move they are watching (on the left screen) regarding how to approach a woman in order to have her become romantically interested.

computing [176], queer theory [212],or social justice interaction design [127], as we study these new technologies. It is imperative that the voices of underrepresented groups are amplified in design of voice assistants, as voice assistants' potential for positive impact depends on it.

Moreover, scholars are also engaging in science-fiction-like speculation about the future of voice assistants; for example, in 2017, Mozilla commissioned a group of researchers to imagine long-term potential futures for voice assistants, and they created three visions for three different voice assistant devices. In these visions, they attempted to break the typical depictions of family engendered in the media, in particular in Amazon's Echo commercials. However, their attempt at doing so is superficial by "showing more than one language being spoken, or challenging gender stereotypes." One of the devices they envisioned asks "why" before fulfilling a request and gains context-awareness and intelligence by doing so. In a vignette portraying an interaction, after its user asked it to turn on the lights (to go to the bathroom) a few times, the voice assistant identified these requests as a potential symptom for prostate cancer and automatically made an medical appointment. Another of their envisioned devices is able to impersonate its user. In the vignette for this device, the user uses the voice assistant device to make a service call, and as the user realizes how unhelpful the agent on the other side is being, she turns down the politeness knob and turns up the confidence knob of the voice assistant impersonating her. Finally, the third vision is for a voice assistant device that is able to synthesize information from literature and help the user argue for what he wants by citing historical events [304]. These ideas are short descriptions that portray what may happen as technological advancements approach scenarios foreshadowed by science fiction. However, even in these scenarios where the creators attempted to be more inclusive than the status quo, the lack of reflection based on positionality leaves significant issues unaddressed and limits their potential for positive impact. For example, how might the effects of using decreasing a "politeness" knob be interpreted differently when the user sounds local versus foreign? As they have been envisioned, these voice assistants meet the needs of a person who does not feel like going to work, and of someone else who is annoyed by the customer service representative. However, how could a voice assistant be designed to make the online world more inclusive to someone with limited technological literacy? What interactions would benefit individuals with disabilities for whom a voice assistant could grant more autonomy? How would these design speculations be different if they had employed social-justice oriented frameworks to promote public interest and to amplify underrepresented needs?

Humans have been anticipating the arrival of voice assistant technology for decades, and now that the technology is finally here, we continue speculating about where it is headed. As is evident from the multiple scenarios portrayed in science fiction, and from the examples that were alluded to, voice assistant technology can evolve in many ways. These ways can be good and bad, respectful or intrusive, genuinely helpful or manipulative, and inclusive or divisive. This is why it is so important to innovate with caution and thoughtfulness.

2.3.2 How Voice Assistants Work

There has been a enormous amount of work invested in making voice assistants technically possible. Efforts towards this goal include: learning from biology (i. e., how does human speech work in the first place) [43, 314], creating hardware systems to record and make sense of sound [241], building mathematical and computational models to interpret recorded sound [186, 155], and bringing these and elements from other fields together into a cohesive, connected system that can

Figure 2.4: Voice assistant devices envisioned by Rogers et al. The one on the left has knobs for humor, politeness, mood, personality, and confidence. Image reproduced from [304].

be packaged in computers smaller than a matchbox (e.g. the Apple watch [31]). The subsequent sections break down how voice assistants work into two subcategories: speech recognition and response (see Figure 2.5). Speech recognition includes transcribing speech to text, and deriving intent from the text [328]. The voice assistant response includes transforming the intent into an action, which could include a speech response, requiring text to speech, and/or another action, such as fulfilling a command to turn on a light or send a text message [328]. Voice assistants perform these computations in a way that is essentially invisible to the user.

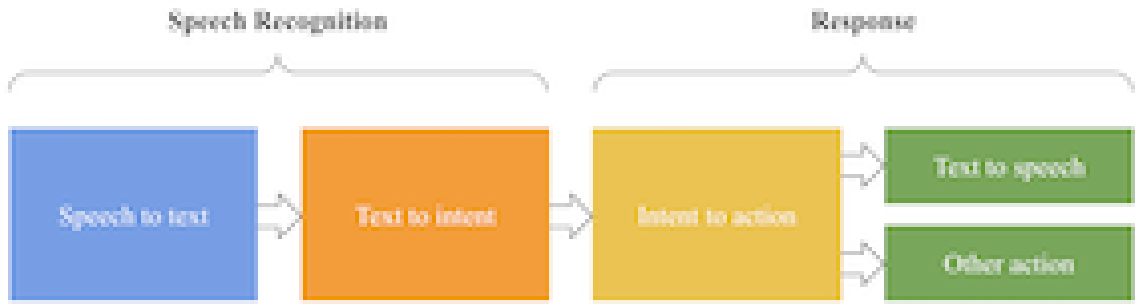

Figure 2.5: Diagram breaking down how voice assistants work into two subcategories: speech recognition and response. Speech recognition includes transcribing speech to text, and deriving intent from the text. The voice assistant response includes transforming the intent into an action, which could include a speech response, requiring text to speech, and/or another action.

Speech to text

An important step towards building machines that can recognize speech was learning how human speech works. The human mouth has different cavities, with which we can generate sound at different frequencies to make vowels [269]. The frequencies generated from these different cavities are called formants, and they vary based on how the articulators' (such as the tongue) position and shape affect the cavities in our resonator tract [269]. This directly maps to how spoken sounds generate different signals that machines can register and categorize. In 1952, a controlled study was ran to measure the frequency, in terms of cycles per second (cps), and amplitude of the formants making up vowels [283]. The first formant (principally generated in the pharynx) of the vowels in the word "peak" is relatively low (~ 310 cps), meanwhile the first formant in the vowel of the word "clock" is relatively high (~ 850 cps); on the other hand, the second formant (mostly created from in the oral cavity) of the vowels in the word "peak" is relatively high (~ 2790 cps), meanwhile the second formant in the vowel of the word "clock" is relatively low (~ 1220 cps); the frequencies selected here are measurements of adult female speech, as they tend to be in the middle, lower than children but higher than men [269, 283]. Being able to classify vowels was an important milestone for speech recognition.

Another part of understanding speech recognition is understanding how to represent sound from speech in ways that are understandable to humans and machines alike. Electrical signals from speech can be can be interpreted by machines, and turned into visualizations that make it easier for humans to interpret the data. There a few ways to represent sound via visualizations, including waveforms, frequency spectra, and spectograms. Waveforms depict sound intensity over time, and are common in many video and sound editing software tools such as Adobe Premier Pro or GarageBand. Jürgen, a phonetics expert, explains how formants can be visualized using a frequency spectrum (the range of frequencies in an electrical signal that can be generated by speech), or spectograms (sound frequencies and amplitudes over time) [163]. We can identify vowels from a single snapshot of their frequencies; however, identifying consonants requires incorporating time-based elements such as silences and formant transitions into the analysis [163]. These representations and techniques were used to further the knowledge necessary to build speech recognition machines.

To briefly illustrate the evolution of speech recognition technology, I rely on a book chapter on the history of automated speech recognition by Gold et al. [155], and add some more current context. The first speech recognition machines relied on knowledge about how speech works to program rules that would parse the electrical signals generated by speech [155]. For example, because we can represent speech in terms of formants, a rule was created to parse speech based on a cutoff point, 900 cps, which was an appropriate cutoff approximation between formants for adult male speech [155]. In other words, the frequencies below 900cps made up the first formant, the ones between 900cps and 1800cps made up the second formant, and so on. Once parsed, the frequency for each formant present in one moment of sound could be approximated, and the corresponding speech sound could be inferred [155]. A less sophisticated technique, that nevertheless relied on similar knowledge, was used for creating the Radio Rex, a speech-activated toy patented in 1916, that would push Rex out of its house when its name (~ 500 cps) was called out [65, 155]. In this case, the toy skipped the "text to intent" step by only having one possible intent, triggering the dog to come out of the house. Another important technique for more robust speech recognition involves using far-field microphone arrays, which are used to measure the direction speech is coming from [199]. This section briefly explains some elements of speech to text recognition; however, for voice assistants, this is only the first component of how they work.

Text to intent

The Radio Rex and the first speech recognition machines were limited in many ways, and relied on hard-coded rules to work. However, subsequent innovations and methodologies became increasingly sophisticated, allowing the recognition of more words, phrases, and subsequently intents. Advances in various techniques such as in the use of Hidden Markov Models (use statistics to predict future states based on current state) [186], the fast fourier transform (an algorithm that allowed the efficient filtering of sound into separate waves of different frequencies) [267], and linear predictive coding (a form of compressing sound information) [274] allowed machines to process and interpret a wider array of sounds. For example, the use of dynamic programs enabled pattern recognition, which helped with identifying utterances that were said at different speeds [310]. These concepts form part of a field called natural language processing (commonly referred to as NLP). One of the most successful, commercially available speech recognition tools is Dragon Home, a software tool produced by a company called Nuance that can be installed in computers to allow computer use via dictation. Dragon was first developed in 1975 using probabilistic function of a Markov process [56], and now a useful speech-to-text tool for many, especially useful for individuals with motor or other impediments that prevent them from being able to type. In order to prevent mistakes, and to accurately type (in this case this would be the action) Dragon relies on NLP models to make predictions about what a person is trying to communicate [11]. These models work by learning from large datasets, which have labeled data to "train" models and "test" models. The creation of standardized datasets was also a key component of moving speech recognition forward [155]. The first standard data corpus was TIMIT [150], which was also used as a way to unify and compare research emerging from different places. At first the datasets were used for mostly for detecting whether speech had been correctly transcribed. There are many NLP breakthroughs that make deriving meaning from text possible, an important one for deriving meaning is a recurrent neural net method developed in the late 1990s called long short-term memory, capable of processing sequences of data, such as speech. [169]. More recent developments include large transformer models for synthesizing speech from text [210], and for question answering and language inference [122]. In summary, using various methods to record and extract different features of speech, the implementation of different theoretical models, and the creation of centralized datasets to train and test different methods have resulted in extremely capable speech recognition technology that is becoming increasingly more robust [155].

Intent to action

Once an intent has been identified, an action can be assigned. An iconic example for a machine able to carry out a conversation on its own is ELIZA (1966), which was a hard-coded program that allowed its users to carry out conversations similar to that of some psychotherapists. ELIZA worked by identifying a keyword in what a user typed and then applying rules to compute a response based on the keyword (or lack thereof) [370]. Many years later, Richard Wallace, influenced by ELIZA, developed A.L.I.C.E. with the intention of passing the Turing Test, a test to see if a human can tell a machine and another human apart [350]; however, even though much more capable than ELIZA, A.L.I.C.E. was not able to pass this test [364]. More discussion of distinguishing human from machine is included in Chapter 6. When programming an Alexa "skill" [44] or a Google Assistant "action" [123], a developer can select from a menu of options how to respond, and apply rules to determine when to respond how.

Text to speech and other actions

A speech response is developed by creating synthetic speech, which requires the automatic reproduction of sound. Automatic sound reproduction dates back to around 875 A.D., when a pair of brothers created a “Banu Musa", which was a water-powered musical instrument able to automatically generate sound [145]. If we fast-forward several centuries, we can find YouTube videos demonstrating how to trim audio clips to represent the main sounds in a given language, and then use computer programs to stitch these clips together to form speech [32]. Susan Bennet’s website contains her testimonies regarding having her voice professionally recorded to use for the original voice for Siri [38] suggests that a similar technique to the one in the YouTube videos was used by Apple to create the first commercial version of Siri. Today, we can use deep learning to train a synthetic voice to speak like specific humans with just seconds of recorded speech [180, 179].

The ease and flexibility enabled by the emerging technological developments in speech synthesis will give room for a lot of control over the types of voices we create. This control can be beneficial in helping personalize voices, in particular for individuals who have difficulty hearing certain conversational styles or accents.

Given the multi-modality of voice assistants, a response can be in the form of speech alone, or it can also be paired with or take other forms. For example, to respond the the command, “Alexa, set a timer for 10 minutes," Alexa can respond, “your 10 minute timer is starting now." In doing so, Alexa has generated a speech response, and also performed the action of setting a timer, which could also include a visual display if Alexa is in a device with a screen and the (non-speech) sound of the timer when it ends. There are many actions a voice assistant can respond with; some of these include: turning on a light, or an electrical appliance [1, 15]; making a robotic vacuum return to its base [35]; playing an audiobook [6], music [64], or a video [25]; providing sound effects as one reads a book [26]; or making a call [20] or sending a text to a contact [19]. In short, knowing how to respond involves first understanding what was asked, and then following rules to create a response, which can be a speech and/or other type of response.

The commercialization of voice assistants

Voice assistants capable of robustly developing responses to speech started to appear on the market in the 2010s. In 2011, Siri launched the iPhone 4s [37], marking an important milestone in bringing voice assistants to millions of consumers. Three years later, in 2014, Amazon announced the launch of its first Alexa-powered speaker, the Echo [33]. And Google announced the Google Assistant and the Google Home in 2016 [34]. The introduction of voice assistants into the market is creating an ecological change. I make this argument borrowing from Postman's fourth idea about technological change, which argues that "[t]echnological change is not additive; it is ecological." [290] Postman describes what he means by ecological change with an analogy: "What happens if we place a drop of red dye into a beaker of clear water? Do we have clear water plus a spot of red dye? Obviously not. We have a new coloration to every molecule of water." [290] This change affects all of us. For example, upon entering any room with a voice assistant device, if a person says something that the voice assistant interprets as its wake word then that person will likely be recorded regardless of having consented and without necessarily being notified. Subsequently, these recordings may be saved and shared without the person's permission. In addition, and also borrowing from Postman, "there are always winners and losers in technological change." [290] In the case of voice assistants, the winners are most likely to be those whose voices and needs are represented when voice assistants are being designed. Today, voice assistants are ubiquitous (news reports indicate that smart speakers sales alone reached new record of 146.9M in 2019 [29]) and ever more capable, creating an urgent need for more equitable representation in their design process.

2.3.3 How Possible Is It for Anyone to Build Voice Assistant Interactions?

There are some options for third-party developers to build voice apps using developer consoles. For example, Amazon has the Alexa developer console, which allows people to develop voice apps, or “skills”, for Alexa[1]. Similarly, Google has a developer console for people to build voice apps, or “actions”, on the Google Assistant[2]. It is worth noting that platforms controlled by the bigger technological companies allow these companies to benefit from having data associated with using them and may make it more difficult to compete against them. Off-the-shelf software frameworks include Amazon’s Alexa Skill Kit [44], Google’s Dialogflow [123], and Mycroft AI’s Skills Kit [40]. Mycroft claims to be the most "hacker"-friendly platform, and claims to give users and developers more control over their data, giving people the option to share everything (and be put towards the bigger dataset to increase their system’s intelligence) or nothing; however, Mycroft still relies on tools provided by bigger companies, and it is difficult to assess how these companies use that data. All these platforms provide software development kits, including APIs, resources, and documentation that enable developers to benefit from the platforms’ machine-learning and natural language processing capabilities. If looking to develop voice-based interaction in multi-modal systems (e.g., by including video), some of these platforms already offer ready-made APIs for

[1] https://developer.amazon.com/en-US/alexa
[2] https://developers.google.com/assistant/console

this purpose, such as Alexa's Video Skill API [46] or Google's Interactive Canvas API [156]. Additionally, there are voice assistant tools that allow developers to combine voice-based and text-based AI, such Amazon Lex [45] and Dialogflow [123]. These options allow a wide range of people to be able to contribute towards building the future of voice assistants; however, it is much more difficult to build anything comparable in functionality without depending on tools provided by companies leading the market. Again, the needs of marginalized groups are underrepresented in the design and development decisions these companies often make, exacerbating inequality.

2.3.4 How Is Voice Assistant Technology Advancing in The Near Future?

Deep learning has accelerated the development of voice-based technology, and it will continue to improve in many fronts. For example, currently, commercial voice assistants have a difficult time understanding speech with different accents or conversational style from the ones that were used to train them [251, 49, 280]. However, voice assistants will become increasingly better at understanding all sorts of speech. For example, Google AI research has successfully trained a system to recognize impaired speech [158]. There is a growing list of datasets that can be used to train speech applications, a highly popular one is Mozilla's Common Voice dataset, which is part of an effort to bridge the digital speech divide, allowing people all over the world to contribute to it and to download the dataset to train speech-enabled applications [9]. Synthetic voices were initially trained using many hours of professionally recorded speech (for example, see Susan Bennett's testimonial as the original voice of Siri [38]), but today one can download pre-trained models and create a voice that sounds like any other voice with just seconds of training data [180, 179]. The technological advances foreshadow many opportunities for voice assistants to be more inclusive by being able to both understand and sound like a variety of voices.

2.4 Related Work on Voice Assistants

A significant amount of research has emerged related to voice assistant technology. There is research surrounding technical advances for the technology [120, 379, 199, 122, 210], associated risks [247, 201, 97, 384, 144], and platform comparisons [245, 218, 300]. Tangentially, some are researching voice assistant applications in other fields, such as mental health [244, 242, 245], or medicine [243, 324, 68]. More specifically within the human-computer interaction (HCI) field, researcher surrounding voice assistants' social role [295, 89, 94], conversational (or not) nature [62, 103], ability to help young children read and learn [376], and promise as a tool to encourage self-disclosure [207, 381] has emerged. At a higher level, researchers are also endeavoring to understand voice assistants' current use and impact by analyzing usage logs [64], or product reviews [268, 295].

Though more scarcely and more preliminary, some research has been specifically focused on underrepresented communities. For example, Amugongo found a need for more African accents and languages to be incorporated in voice assistants [49]. Shiller's research argues that "Alexa or Google Home elide and reproduce the gendered and racialized dimensions of domestic labor, streamline this labor for capture by capital, and heighten the very affective dynamics they promise to ameliorate" [320]. Relatedly, Hannon [165] and Ni Loideain and Adams [261] examine the role of gender in voice assistants. Finally, Pradhan et al. [292], Balasuriya et al. [57], and Friedman et al. [147] discuss the use of voice assistants by people with disabilities. Amongst the underrepresented groups in the design and research of voice assistants are older adults [334]. Even though much research has emerged surrounding voice assistants, the representation of groups that could highly benefit from voice assistant technology is lacking.

2.5 Voice Assistants for Older Adults

The proportion of older adults in the US population is larger than ever before and the growth trend is expected to continue [82], widening the need to support aging in place [175]. Many older adults desire to live independently at home for as long as possible, suggesting possible opportunities for technological innovation [138, 291]. As we age, our bodies begin to function differently: while we may not remember things like we used to, have shaky hands, and/or need glasses to read text [119], we also become better at regulating our emotions (i.e. by implementing by cognitive control mechanisms that enhance positive and diminish negative information) [233]. In the face of these changes, new technologies, such as voice assistants, could be tailored to older adults' needs and preference to support happily and independently aging in place [373, 39, 214].

Older adults are frequently excluded from research and design of mainstream technologies, such as voice assistants [313, 361, 64], resulting in technological products that do not address their needs [117, 118]. Despite the unique promise voice assistants present for older adults, voice assistant research focused on older adults is scarce [313]. A 2019 systematic literature review of research published in the Association of Computing Machinery Digital Library only found 16 articles addressing the perception and use of voice assistants by older adults as opposed to 1581 records when age-related keywords were removed [334]. This is a missed opportunity, as an analysis of voice assistant usage logs by Oh et al. [272] found that older adults used the voice assistant more when compared to younger adults. In 2020, Pradhan et al. [294] conducted a general-use study with similar methods to the ones we employ in Chapter 3, a 3-week field deployment of the Amazon Echo Dot in the homes of seven older adults, and found consistent usage for finding online information, including health related information. In another study, also with similar methods to ours, Kim and Choudhury [192] found that over time older adults felt less worried about making mistakes and enjoyed the digital companionship as they got used to using voice assistants. In contrast to the work presented in Chapter 3 of this dissertation, these two studies did not use design provocations to dive deeply on the specific topics of home health and wellbeing.

In their study, O'Brien et al. [268] identified five major themes for older adults' use of voice assistants: (1) entertainment; (2) companionship; (3) home control; (4) reminders; and (5) emergency communication. Out of these, though not explicitly stated in the paper, the first two may be related to wellbeing, and the last two to home health. Ennis et al. [138] developed a voice assistant-based innovation to support ambient assistive living and received promising feedback. However, despite their promise, many older adults abandon the use of these devices, citing difficulty in finding valuable uses, beliefs about the lack of essential benefits provided, or challenges with use in shared spaces [348].

2.5.1 Combating stereotypes surrounding older adult technology use

In response to harmful stereotypes made about older adults, Rogers and Marsden [305] called for us to move beyond the rhetoric of compassion to one of empowerment. Furthermore, Vines et al. [360] found that discourse is often framed in problematic ways that represent common stereotypes around old age, and call for more-individualized and more-contextualized approaches. Higher prevalence of technology use in older adults (65+) has been associated with younger age, being male, white race, higher education level, and being married [151]. Moreover, older adults who adopt information communication technology have been found to value their technology activities as important in their lives Vroman et al. [362].

In an effort to address potential inequalities, we 1) gave participants in our long-term study appropriate training for them to develop expertise in the use of voice assistants, and 2) sought to recruit participants with fewer of the characteristics associated with higher prevalence of technology use by Gell et al. [151] for both empirical studies involving older adults.

Chapters 3 and 4 of this dissertation, add to this early literature by deeply examining potential uses of voice assistants for older adults' home health and wellbeing, by identifying challenges and barriers that older adults may face when interacting with voice assistants, and by proposing design strategies to improve voice assistants' inclusivity.

CHAPTER 3

A FIELD STUDY OF TWO PROTOTYPE VOICE APPS TO SUPPORT AGING IN PLACE

This chapter focuses on the first goal of this dissertation—to examine how older adults currently interact with voice assistants through a lens of inclusivity—by exploring the use of voice assistants for supporting aging in place. I present a research-through-design field study of two prototype voice apps to support aging in place, one for health data reporting and the other for positive reminiscing. The content in this chapter is currently (at the time of writing this section of my dissertation) under submission to be published as an academic article elsewhere. It was written in collaboration with Jessica Bethune, Alexa Lempel, Rony Krell, Katrin Hänsel, Armin Shahrokni, Deborah Estrin, and Nicola Dell. Because of this, I will be switching from the singular "I" to the plural "we" for the rest of the chapter. The writing has been slightly edited to fit this dissertation.

We delivered Amazon Echo Shows with Alexa to the homes of five older adults, provided appropriate training, tracked usage, and deployed the prototype voice apps (as design probes) to their devices. Participants learned to use voice assistants, and saw value in the prototype voice apps. Reminders and the display were helpful supports. We surface design challenges entailing the use of voice assistants for health data reporting, such as participants' perception that they would be bothering their doctors. We then provide design guidelines for developing voice apps tailored to older adults, and discuss how voice assistants might enable continuity of care in people's own homes.

3.1 Introduction

In Section 2.5, I described how older adults are frequently excluded from research and design of mainstream technologies. As mentioned before, despite their promise for supporting home health and wellbeing, investigation of the needs of older adults has been insufficient [334]. For example, voice apps tailored specifically to older adults are scarce and understudied. This chapter advocates for improving the inclusivity of voice assistants by intentionally focusing on the needs of a marginalized group in the design of digital technology, older adults. In this chapter, I build upon recent studies that have investigated how voice assistants are used in older adults' homes [294, 192, 387] or how voice apps may be designed for older adults' health information seeking needs [230]. I do so by combining long-term use of voice assistants with the deployment of prototype voice apps to realistically study how voice apps may be specifically designed to support aging in place. Voice apps designed with and for older adults may help to bridge the digital divide in the provision of healthcare and wellbeing services.

This chapter presents a small-scale, research-through-design [385] field study of two prototype voice apps to support aging in place, one inspired by the promise of voice assistants for home health and the other for wellbeing. The first prototype voice app was a voice-based geriatric assessment developed in partnership with a geriatric oncologist. The second prototype voice app engaged participants in positive reminiscing by using questions from StoryCorps [16] as prompts. StoryCorps is a non-profit organization centered around storytelling and popularized by National Public Radio[1]. We delivered smart speaker-based voice assistants (i.e., Amazon Echo Show with Alexa) to five older adult participants who lived alone, and analyzed their usage of the device for approximately two months. We also conducted periodic interviews with each participant in which we provided training and guidance about existing functionalities of voice assistants and introduced the prototype voice apps, which were deployed to their devices before the interviews. We analyzed participants' usage logs and interview transcripts.

Participants were able to effectively use the prototype voice apps, and saw the value in using Alexa for health and wellbeing in a way that they had not imagined or mentioned before using the prototype voice apps. We found that participants needed some sort of scaffolding or reminder to use the prototype voice apps outside of interview times, and that the display helped support health data reporting. We surface design challenges entailing the use of voice assistants for health data reporting, such as participants' perception that they would be bothering their doctors. In addition, engaging with the prototype voice apps empowered participants to contribute ideas for improvements and for other voice apps that would be useful to them.

We contribute an in-depth exploration of two prototype voice apps to support aging in place to research on computer-supported cooperative work. The engage-

[1]https://www.npr.org/

ments with our prototype voice apps yielded usability findings that inform design guidelines for future voice app development tailored to older adults, and findings that will inform future research for technologies to support continuity of care in older adults' own homes.

3.2 Related Work

The related work presented in this chapter extends the one previously described in Section 2.5, which focused broadly on voice assistants for older adults. Here, we focus specifically on research surrounding the design of voice apps tailored for older adults related to health and wellbeing. We cover work that investigates voice assistants' role as a technological solution that may support aging in place, in particular via home health, and on their role as companions that may help improve wellbeing. This chapter extends this research via a research-through-design [385] field study of two prototype voice apps to support aging in place.

3.2.1 Voice apps for aging in place

Healthcare literature increasingly recognizes the need for technological solutions to support older adults' healthcare needs [373, 325, 39, 291, 214], an essential component for supporting aging in place. Abdi et al. [39] specifically cite voice assistants as one of eight emerging technologies that could potentially be used to meet older people's needs in various care and support domains. Several researchers have investigated the use and perception of voice assistants for health information seeking [79, 257, 230]. Chen et al. [91] explored voice assistant care delivery and quality of life enhancements for older adults as a collaborative task between patients and providers, highlighting more support for health data reporting (what our first prototype voice app does) as a major application for voice assistants. To investigate specific voice apps, some have done participatory design workshops with scenarios [230], and others have created and tested prototypes that rely on scripted utterances [257]. While these sorts of interactions are valuable to characterize beliefs and perceived benefits and barriers, they are early-stage investigations that do not fully cover the range of limitations that come into play in reality. For example, in a study of the formation of a makerspace in a retirement community, Lazar et al. [202] found that "to answer questions about their preferences for a makerspace, participants had to first form a sense of what a makerspace is or is not for." Without this information, ideas were too broad or too narrow. Similarly, without knowing the specific capabilities and limitations of voice assistants, participants would have difficulty formulating detailed preferences or coming up with well-scoped ideas for voice apps. Because of this, we developed and deployed prototype voice apps that mimicked real-life interactions. This allowed us to investigate the ease-of-use of these prototype voice apps in a naturalistic manner, while staying true to the constraints of the technology. Furthermore, given that ease-of-use and usefulness are significant variables in predicting technology acceptance [248], it is of utmost important that these interactions are studied in a realistic setting where usability issues can be noted.

Companionship and wellbeing

Although computer use alone has not been associated with wellbeing [124, 197, 363], recent research has found that some older adults enjoy voice assistants' companionship [98, 268, 293, 192, 105]. The Oxford Languages dictionary defines as *the state of being comfortable, healthy, or happy*. Healthy cognition is an important aspect of wellbeing, and factors such social disengagement have been directly linked to cognitive decline [59]. Because of this, social relationships with voice assistants could be leveraged to support older adults' wellbeing.

The role of voice assistants as companions has been noted by many. In an analysis of Amazon reviews about voice assistant use by older adults, O'Brien et al. found that one of the categories older adults use voice assistants for is companionship [268]. More deeply examining the ontological categorization of voice assistants, Pradhan et al. [293] found that participants desiring companionship were more likely to value the voice assistant as a companion, and that the categorization of the device between companion and object was fluid. In the same vein of research, Chung et al. found that older adults tended to personify the agent more than younger counterparts by using polite words such as "grateful", viewing it more as a companion [98]. Moreover, Kim and Choudhury [192] found participants built digital companionship as they became more experienced with voice assistants. Corbett et al. [105] made a call for more research in this area, as their mini review of literature covering older adults use of voice assistants and its relationship to social isolation and loneliness suggested a promising role for voice assistants as companions to reduce loneliness. These findings are also supported by research on similar technological agents [359, 203, 302].

In this chapter, we extend this research by examining how voice apps can be designed to support wellbeing through a prototype voice app for positive reminiscing. Prior research suggests that positive reminiscing can boost happiness [81].

3.3 Method

The goal of this study was to inform the development of voice apps specifically tailored to older adults. To achieve this goal, we conducted an IRB-approved in-the-home study with five older adults living alone, recruited via local senior centers in a large U.S. city. The study was done remotely via phone or Zoom interviews, all conducted by the first author. We selected Amazon Alexa smart speakers with screens as the voice assistant devices for the study. We deployed the prototype voice apps to the smart speakers in participants' homes. In light of the importance of an interviewer's position of privilege and marginalization in how interviews play out [271], we acknowledge that the interviewer was a Latin American graduate student at an academic institution, in her late 20s, with a subtle Spanish accent. We now describe our study methods.

3.3.1 Participants

We recruited five older adults (four women, one man) between the ages of 62 and 85, with varying degrees of technical familiarity. We identified these participants through senior center directors who attended a presentation about our study, and gave us prospective participants' name and phone numbers to directly reach out to them. All participants lived by themselves independently and had WiFi in their homes. Table 3.1 provides more details about participants. We intentionally chose to conduct a study with a small number of participants in which we sought deep engagements with participants over an extended period of time. This type of small-scale study design is supported by Dix [125] who argues for the value of small-scale studies "as we move from a small number of applications used by many people to a 'long tail' where large numbers of applications are used by small numbers of people," and Vines et al. [360] who suggest critical engagement with an individual's context as a strategy to combat common stereotypes that prevail in the literature. We deliberately seek to characterize the individuality of our five participants, highlighting the diversity of older adults' uses of, and reactions to, new technologies. In this vein, we provide rich descriptions of each participant accordingly:

Travis (67) is a Black man from the Southern U.S. who works at a restaurant washing dishes. He recently got a smart TV and WiFi in his home. He knew of Alexa through TV commercials. In our first interview, he asked if he could use

Name	Gender (& age)	Home computer?	Zoom?	Device location	I1 (days from 1st use)	I2 (days from I1)	I3 (days from I2)	Total weeks	Usage trend	Mean daily interactions
Travis	M (67)	No	No	Bedroom	40	20	28	12.6	Increased	7.64
Gilda	F (82)	Yes	Yes	Home office	15	6	20	5.9	Increased	3.54
Betty	F (85)	Yes	Yes	Living room	22	8	25	8.0	Decreased	4.39
Theresa	F (85)	Yes	Yes	Home office	21	17	19	8.1	Increased	5.91
Maria	F (62)	No	Yes	Living room	20	14	26	8.6	Decreased	17.40

Table 3.1: Summary of participant demographics, interview timelines, and usage trends. To preserve anonymity, names shown are not participants' real names. We determined trends by comparing the number of interactions in the first half of the total number of days with the device to those in the second half. Interaction counts exclude "Alexa" or "echo" single word transcripts, or recorded audio snippets labeled by Amazon as, "Audio was not intended for Alexa" (or similar).

Alexa to turn the lights on and off like he had seen on TV. After the researcher explained he had to buy smart home bulbs or plugs for that, he said "*disregard that, I don't need that then.*" During our study, he used Alexa mostly to listen to rain sounds, as they helped him sleep and reminded him of his childhood home. Alexa frequently overlooked his interaction attempts when he pronounced its name as "Alexia." Thus, in the second interview, we changed the device's name to Echo. During the time before our first interview, he asked Alexa for help with reading and writing, and also, "*How can I be as smart as you?*" In our last interview, these requests had waned off, and Alexa's main role in his life had become to play rain sounds at night. He had a smartphone, but did not know how to use it other than for calls.

Gilda (82) is a widowed white woman who emigrated for work purposes from Germany, and married her late husband in the U.S. Her only son lives about two hours away from her by public transit, and has a screenless smart speaker-based Alexa. She was very cautious about COVID-19, so we dropped off the device at her door. It took almost a month after drop off to get her started, because she could not find a place to plug in the device and mobile hotspot. We eventually found that she had WiFi and she was able to connect the Alexa device by following the researcher's instructions over two hour-long Zoom sessions. She teaches a type of exercise therapy method called the Feldenkrais method, and enjoys listening to classical music. Alexa's biggest roles in her life were to remind her to drink water every day at noon and play music while she exercised. She knew how to use a computer well enough to check her email, search Google, and use Zoom, but did not use a smartphone.

Betty (85) is a widowed white woman who grew up in the large city where this study took place, and was a biology school teacher. She currently runs a political website and is an avid mobile Scrabble player. She has two children in their fifties, one of whom works for the police and the other as an engineer who stays up-to-date with new technologies. She was not familiar with smart speakers, but was familiar with voice assistants, as she used Google Assistant on her smartphone. She explored many of Alexa's features, including games, music, videos, news, and exercise voice apps. She also has many friends in her age range with whom she maintains frequent communication, so she frequently shared their perspective of voice assistants as compared to hers. She was familiar with and felt confident using various technologies.

Theresa (85) is a widowed white woman of German descent who grew up in the large city where this study took place, with many siblings. She often helped her parents with their jobs, as they were the superintendents of a building. In her childhood, she liked to take the bus to go watch baseball games at the stadium in the city, and later in her life she spent some time traveling in South America. Now, she lives by herself, and her only family is her older sister who resides in a nursing home and has Alzheimer's. She enjoys taking care of her garden, and selling old parts on eBay. When we dropped off the device in her office, she had a game of Solitaire open on her desktop computer. She knew of Alexa, because one of her friends has an Alexa device and had raved about it to her. At first, she had a difficult time finding valuable uses for Alexa, but she eventually found a feature that was a good fit for her, a voice app with stretching exercises. She preferred not to be seen on camera, so all of our calls were voice-only. She had not been able to set up Zoom on her own for online senior center activities, but was able to set it up with the researcher's help for our study. She did not use a smartphone.

Maria (62) is a Latin American, retired woman who struggles with depression. She lives by herself, but is frequently visited by her children, grandchildren, and boyfriend. Alexa fulfilled a social support role in her life, often brightening her day. Her usage logs revealed consistent "*good morning*" and "*good night*" greetings, 121 to be precise. Her visitors often interacted with her Alexa device, and she sometimes felt protective of it. She explained that unlike her family, Alexa was not judgmental of her, which made her feel most comfortable asking Alexa questions.

She joined our interviews from a tablet, and had an iPhone with Siri. She was proficient using her iPhone for texting.

3.3.2 Procedure and Materials

We dropped off multi-modal smart speakers at each participant's home and allowed them to organically use the speaker for at least two weeks before the first remote interview. This gave participants enough time to explore the device in a way that mimics a real-life situation where they have received the device as a gift. The devices were Amazon Echo Shows (second generation), with a 10.1 inch high-definition smart display with Alexa. During drop-off, we helped participants setup the devices with their home WiFi and gave them training surrounding common usages (e.g., weather forecasts, music, and information retrieval), explained how to mute the device if they did not want it to be "listening," and answered any questions they had. After they had spent at least two weeks with the installed device, we conducted three audio-recorded, hour-long, semi-structured interviews with each participant (see Table 3.2 for a summary of each interview's content). The interviews were spaced at least five days apart from one another, and recorded for transcription (see Table 3.1 for exact number of days between interviews). The interviews also served as ways to answer participants' questions. Interviews were conducted either via phone calls or Zoom video conferencing meetings (depending on the participants' preferences and abilities), and were themed around the following topics: (I1) grounding, (I2) health, and (I3) wellbeing. For the latter two interviews, we employed a research through design approach [385] by creating two prototype voice apps—one for health data reporting, and the other for positive reminiscing–as design probes to engage participants in in-depth discussions and ideation about the potential of using Alexa for health and wellbeing.

Voice app implementation

The prototype voice apps were deployed to, and ran from, participants' devices using dedicated email and Amazon accounts for each device. We built the voice apps using Amazon's Alexa Skills Kit (ASK), which is a compilation of open sourced Alexa application programming interfaces and tools to develop voice apps. We deployed the voice apps to participants' devices remotely by inviting the devices' dedicated email accounts to be Beta testers, and accepting those invitations via the devices' Amazon accounts (which used the same email address). We shared the specific phrase to run the prototype voice apps during their respective interviews. We explicitly asked them to use the prototype voice app for health data reporting once outside of interview times. We tracked usage of the voice assistant and prototype voice apps by reviewing the interaction history on the devices' accounts.

Interview 1 (I1): Grounding	Interview 2 (I2): Home Health	Interview 3 (I3): Wellbeing
Alexa strengths and challenges	General update	General update
Questions from usage logs	Prototype voice app #1	Prototype voice app #2
Brainstorm potential uses	Homework: try prototype voice app once on their own	Reflection of the full experience

Table 3.2: Summary of interviews.

Grounding interview

The first interview was focused on establishing grounding with participants. In this interview, we sought to understand the context in which participants had been using the device, and what they perceived the voice assistant's strengths and challenges to be. Knowing the strengths they perceived allowed us to tailor recommendations for more potential uses, and knowing the challenges allowed us to provide additional support. In addition, we relied on usage logs to dive deeper into participants' experiences with the device by asking about interactions that they might not have brought up during the interview. In summary, the grounding interview served many purposes: 1) to provide additional training, as recommended in existing literature [105, 119, 363]; 2) to establish rapport between the interviewer and each participant; and 3) to document the participants' perceptions and experiences with the smart speaker before introducing our prototype voice apps.

Interview with the prototype voice app for health data reporting

The second interview was centered around uses for home health. We chose health data reporting as an appropriate application for voice assistant technology, because this was a need raised by healthcare practitioners across multiple institutions in our exploratory phase. The prototype voice app for health data reporting (see Figure 3.1) was a health assessment consisting of a series of questions from an established geriatric assessment [325] adapted from written to spoken format. For this interview, we introduced an imaginary scenario and asked participants to interact with the prototype voice app. At the end of the interview, we asked participants to use the prototype voice app once on their own before the next interview. The prototype voice app we developed was called *My Care Questionnaire*, and asked how much users had been limited in seven activities of daily living (ADLs) by their health condition, as follows:

Participant: Alexa, open My Care Questionnaire.

Alexa: Welcome to My Care Questionnaire. Your answers will help your medical team provide more comprehensive care for you, which may improve your outcome. Are you ready? You can say yes or no.

Participant: Yes.

Alexa: All right. Let's get started. How much have the following activities been limited by your health condition? For each activity you can say, not limited, limited a little, or limited a lot. The first activity is bathing.

Figure 3.1: The prototype voice app for health data reporting. Left: Amazon Echo Show 10 displaying the bathing ADL. Right: thumbnail images for the remaining 6 ADL screens.

Participant: Not limited.

Alexa: How about dressing? …

For each ADL, a screen was also displayed (Figure 3.1) with the answer options and an image for the activity: bathing, dressing, grooming, feeding, walking inside the home, walking outside the home, or bladder and bowel control.

Interview with the prototype voice app for positive reminiscing

The third interview explored the use of the voice assistant for wellbeing via a prototype we developed to facilitate human-to-human or human-to-machine positive reminiscing by providing story prompts from StoryCorps [16], a non-profit organization whose mission is to record, preserve, and share the stories of people in the U.S. from all backgrounds and beliefs. Positive reminiscing can boost happiness [81], improving wellbeing. We chose to employ a prototype voice app that was not strictly within the health domain as a way to explore a use case that deviates from discourses surrounding health economics or deficit [360]. The prototype voice app can be used in group settings or by oneself. The invocation phrase for the prototype was "load a StoryCorps question". This phrase would start the app, which would set the stage by telling users "great questions are on the way", and encouraging them to find a comfortable position, and a recording device if they wanted to record the conversation.[2] Since we were recording the interview, there was no need to record on another device. Once a participant told the app that they were ready, a question would appear on the screen and the voice assistant would also say the question out loud. We used three questions: 1) "What's one of your happiest memories?" 2) "What was your childhood like?" and 3) "What are some of the most important lessons you've learned in life?"[3] Each question would appear individually and the next question would be shown when the participant requested it. In the software backend, we made Alexa play silent sounds so that it stayed active (and did not quit the voice app) while users shared their stories. When the three questions have been asked, if the participant requested another

[2]The prototype voice app required users to use their own device because at the time of the study, Alexa did not allow third-party developers a straighforward way to access voice recordings. To track usage of their voice apps, developers had to rely on the automated transcripts generated by Alexa.

[3]These questions were obtained from StoryCorps's list of Great Questions: `https://storycorps.org/participate/great-questions/`

question the app would say, "There are no more questions left for today. Come back tomorrow for more."

3.3.3 Data Analysis

Usage log transcripts

We recorded 4,657 usage log entries from the research accounts on participants' devices. 19% of these usage logs occurred during interviews, so they were removed from *usage trend* and *mean daily interaction calculations* (Table 3.1). 37% of stored audio snippets did not result in successful interactions with Alexa. These were either single word entries with the device's wake word, "Alexa" or "echo," or entries with recorded audio snippets marked by Amazon as, "Audio was not intended for Alexa" (or similar). Our analysis of participants' interactions before and after the interviews shows that, on average, every participant interacted with their device at least three times per day. The researcher reviewed participants usage logs before every interview in order to tailor the content of the interview to each participant's interactions.

Interview transcripts

Interviews were audio-recorded and transcribed. The transcripts from the first five interviews were open coded line-by-line [191] by the first author, resulting in a total of 107 codes. The first and last author reviewed these codes, and refined them into 43 consolidated codes that were used to code the remaining 10 transcripts. All transcripts were double coded, and the first author reviewed all the coding to ensure consistency and resolve disagreements. A few modifications to the original codes were made as new trends emerged (e.g., added new codes, or made original codes more specific). Subsequently, we employed thematic analysis [74] to identify patterns specifically relevant to developing voice apps tailored to older adults.

3.3.4 Ethics and Consent

This research was reviewed and approved by the internal review board at [anonymous institution] under IRB Protocol # [anonymous protocol number]. We obtained informed consent to collect and analyze participants' interactions with the devices, and routinely referenced usage logs during our interviews. We also obtained consent to record each interview session. Participants were compensated with a $25 gift for each interview, and kept the devices after the study concluded. At that point, we offered instructions and support to create personal accounts for their devices.

3.4 Findings

We found that participants were able to easily use our prototype voice apps and that they saw their value. In addition, we identified the need for reminders (or proactive interactions) as a method for ensuring completion of voice assistant-related tasks. We found that the display helped support health data reporting. We also encountered design challenges, such as the need to balance the expectations Alexa created of humanlike interactions with its limited capabilites, and the need to address fear stemming for lack of confidence in how the technology works. Despite initial challenges, participants gained expertise with the technology throughout the duration of the study, and were able to generate ideas for improvement and for voice apps. We now describe these findings in detail.

3.4.1 All participants successfully used the prototype voice apps

All participants were able to use both prototype voice apps—the one for health data reporting and the one for positive reminiscing—despite their different backgrounds and abilities (e.g., Travis was low-literate, and Gilda and Theresa did not use smartphones). Some gave feedback for improving them. For example, Theresa thought that the prototype voice app for positive reminiscing *"cut straight to the point*," that there was *"no fooling around, no foolish business*." Thus, she suggested that the questions could be rephrased to be made more friendly, "*What it needs to say is, 'In your lifetime, did you really have a special moment that made you* *wonder what life is all about?' Or something like that."* She further explained that she would like the Alexa to elaborate more on the questions. In addition, Betty thought Alexa could be programmed to listen to people vent, so that it was not limited to storytelling.

Four of five participants completed their homework: to use the prototype voice app for health data reporting a second time on their own.

After all participants had used the prototype voice app for health data reporting in front of the researcher during the interview, we asked them to use it on their own time before our next interview. This was done to test participants' ability to use the device independently and without the presence of the researcher, as we envision happening if they were completing a health assessment that their doctors sent to their devices in their homes. All participants except Travis completed the "homework" successfully. Theresa called us to remind her what the phrase she had to say was, and proceeded on her own after we gave it to her.

3.4.2 Most participants saw the value in the prototype voice apps

Value in the prototype voice app for health data reporting

Once the first interaction concluded, participants appreciated the value the prototype voice app for health data reporting. Maria exclaimed in awe, *"I didn't know* *that I could, you know, that I could [do that]."* She also expressed that she *"just felt comfortable answering the questions, and it felt like [she] was at the doctor's office."* Travis stated:

> *"It's cool. It's something that your doctor would probably ask when you're having problems, so they can give you some information or maybe send somebody here like a house person, nurse's aid or whatever."* (Travis)

Experiencing this design prototype voice app elevated participants' perception of Alexa's value as a whole. For example, Betty, who was our participant with the most technical expertise, expressed that she initially "*didn't know what [the smart speaker] could do that was particularly interesting to [her], personally, or different than anything [she] could do on [her] phone.*" However, she explained that her perception changed when she saw that Alexa could have the ability to ask "*medical questions*" and "*send the information directly to a physician or somebody who could help you progress or deal with something that you had wrong with you.*" She thought this particular voice app was "*very valuable*", explaining that she "*saw a different thing.*" This said, one participant, Gilda, did not think that this prototype voice app for health data reporting would be valuable. She explained, "*I can't see [Alexa] ever being used for that.*" Gilda maintained this opinion through the end of our engagement, because she thought that her doctor not have time to listen to Alexa, which is described in more detail below. To note, at some point she indicated understanding that Alexa would be generating a written report, and said that then the voice app would be more "*applicable*"; however, she went back to thinking this idea would not work in our last interview.

Value in the prototype voice app for positive reminiscing

All participants seemed to enjoy engaging with the prototype voice app for positive reminiscing. For example, Travis said the interaction was "*pretty cool*," because it made him go back to some of his best memories:

> *"The quality of the questions that it was asking, like what is some of the best memories you have as a younger person. It made me go back to the things that I did do when I was younger. It made me feel a lot better than the people that was on the streets, that I was raised up on. This was back in the '60s and most of the people didn't have the opportunities to travel like I did. That was a great experience. It made me feel that I traveled and was not scared to leave the area by myself, on my own."* (Travis)

Going back to his best memories from the past brought positive feelings to Travis, suggesting gratefulness for the opportunities he had to travel, and pride for having done so bravely on his own. Like Travis, Maria also said she enjoyed reminiscing, she said she liked it, and when asked *why* she expressed:

> *"I don't know it was just, I don't know I can't even explain. It just asked questions that I liked answering. It brought me back, because, like I said I don't have my parents any more and my siblings any more. It was good."* (Maria)

For Maria, going back in time to good memories reminded her of connections with her parents and siblings, which were no longer around. The prototype voice app for positive reminiscing evoked feelings of happiness or comfort in our participants, suggesting the possibilities that voice assistants have in the wellbeing space for older adults.

At the end of the study, when we asked Theresa to compare all the things she had used Alexa for, she mentioned the prototype voice app for positive reminiscing as one of the activities she would continue using, because it made her think, *"I would take advantage of the questions because I like making [myself] think more."* Our prototype voice app for positive reminiscing surfaced to Theresa's top interactions with Alexa, next to the exercises and music. This shows the potential of Alexa to help older adults stay entertained by using thoughtful prompts at the right time. Another indication of the value participants saw in the prototype voice app for positive reminiscing was the desire to come back for more:

Alexa: There are no more questions left for today. Come back tomorrow for more.

Betty: There's no more questions left for today? I'll actually come back tomorrow. All right.

3.4.3 Participants needed some sort of reminder to use the prototype voice app for health data reporting as "homework"

Every participant required some sort of reminder to use the prototype voice app for health data reporting on their own time. Although we would have preferred to enable the prototype voice app to proactively start the interaction, Alexa's third-party developer platform did not provide straightforward ways to support proactive interactions. Thus, reminders to start the assessment served as a workaround. We sent reminder text messages with the command to Betty and Maria, who both used smartphones and were proficient at texting, to remind them to complete the questionnaire on their own and of the invocation phrase. Theresa, who had written herself a reminder on her calendar, called the researcher to ask her what the invocation phrase was, suggesting the reminder via text helped participants remember the invocation phrase as well. Travis, who did not use an Alexa reminder nor a text reminder, did not complete it on his own. He had requested a phone call reminder, but we were not able to reach him. Our experience having difficulty finding the right time to reach Travis inspired us to use Gilda's own Alexa device for the reminder, which we set up during an interview call. Alexa reminded her to fill out the questionnaire by saying, "*this is a reminder to say 'open my care questionnaire,'*" and it worked.

3.4.4 The display helped support health-related interactions.

The screen modality was an important component in helping participants remember what the answer choices were, or what the activities meant. Betty was not looking at the screen when she started her first interaction, and forgot the answer options, which were: "not limited," "limited a little," and "limited a lot." After the first question she said "*not affected*" instead of the option "*not limited*." As she struggled to continue to remember, she explained to the researcher, "*See now, I forgot the choices she gave me*." Once she worked through this obstacle, and was able to complete the questionnaire, she said, "*I think most people doing this would be looking at the screen and [the three options were] very clear. So I would have immediately known what to do. So, that was fine.*" Travis, who did not know how to read and write, was the only participant who noticed the images representing ADLs displayed on the voice app (see Figure 3.1) the first time using the voice app, he described them:

> "*[The images] showed you the toothbrush and toothpaste. It showed you the plate with the fork and spoon and the knife or whatever. They showed you the person when they was actually outdoors, activities. Just showed you a person with a walker to help them support their self or whatever while they're walking.*" (Travis)

Testing the images will be an important step moving forward, as some users may predominantly rely on them. The other participants, who were not low-literate, focused on the written text, sometimes not even noticing there were images next to the text. As a whole, the screen modality can support the voice modality of voice assistants, improving the effectiveness of health (and wellbeing) voice apps. As Betty described, the written responses allowed her to remember what to say. Travis also mentioned using the images on the screen to answer the questions. Betty and Theresa also regularly used third-party voice apps for exercising, which showed images of the exercise positions on the screen.

3.4.5 Alexa created expectations of humanlike interaction patterns

The original health assessment that our prototype voice app was based on was a written questionnaire. A written questionnaire, whether on paper or on a screen, has no turn-taking interactivity, and thus, does not create humanlike interaction expectations. We found that the humanlikeness of the voice-based questionnaire can be misleading, as participants expected for the machine to behave in the same way a human healthcare worker would, not in the way a written questionnaire would. One way, was by expecting the machine to use colloquial language:

"The word, limited, is not an everyday vocabulary word and it really is not. I mean, I don't remember in answering any conversation somebody says, "You do this often?" I would say, no, I'm limited in this or that... You don't use that. I think it's a yes, no, always, sometimes, never, always is a one word answer, is better than the choices you gave. And I would like to hear what other people my age say. Now if you want to have

a comfortable conversation and a lot of people are... it's not a colloquial word, is that's what I'm trying to say." (Betty)

Travis thought the interaction was not difficult, and did not use medical terms that were difficult to understand. Travis's experience suggests that the prototype voice app was meeting the need for easy-to-understand language identified in Martin-Hammond et al. [230]'s study (by Group 3). He was happy with the non-colloquial language, as he could understand what the voice assistant was saying:

Travis: The questions wasn't hard.

Researcher: Do you feel like sometimes when you go to the doctor, they ask questions that are hard?

Travis: Sometimes you have to ask them what they mean. Because sometimes doctors use different terms. They use medical terms and whatever.

Another expectation was that Alexa could respond to ambiguity, like a human would and unlike written questionnaires with strict multiple choice options. Betty expressed, *"I need to be able to explain what I have and the expert to say whether that is considered a little or a lot."* Travis and Betty's responses have an underlying theme in common—the need to ask Alexa for more information. Betty might need to ask, "How do I know whether my answer should be a little or a lot?" and Travis may need a lay person definition of a medical term before committing to a response. In written or web interfaces this can be achieved via information buttons or foot-notes; however, adding information in spoken machine format introduces design challenges.

Moreover, participants also expressed the expectation of more interactivity. Right after explaining that she thought the voice-based questionnaire was "*more personal*" than a written questionnaire, Betty added that she would like for it to let her "*ask a question or add a question*" with a "*back and forth.*" Taken together, participants' comments highlight how the voice-based interaction format created expectations for more humanlike interactions. This presents an interesting design challenge for future development.

This expectation persisted even after participants had completed their interactions with the prototype voice app. For example, immediately after answering all the ADLs, Travis unsuccessfully sought Alexa's help with a health problem he had:

Travis: Echo. I've been having problems with my left foot. I'm a diabetic, type two. Recently I had a blister on my foot, and it got worse. But I went to the doctor, and he gave me some antibiotics. Echo. Did you record what I just said about my foot? Respond.

Alexa: Sorry, I didn't get that.

Travis: I've been having problems with my left foot. I had a blister on it, and it had got bad. So I went to the doctor. And now he gave me some antibiotics for the foot. It seems to be getting better. Can you give me any information to do with it to help it heal?

Alexa: Sorry, I didn't get that.

Travis: I was asking for an opinion from you about blisters on your feet. How can you help me take care of that, make it get better?

Alexa: Sorry, I didn't get that.

Travis: Okay. I'm not going to ask anymore.

Travis's question would have been easily understood by a healthcare provider. However, Alexa does not currently support such interactions, which may lead to confusion and frustration, or break trust.

3.4.6 Participants initially had insufficient mental models of the technology that could hinder adoption, use, or control.

There was hesitation before using the prototype voice app for health data reporting for the first time.

When we first explained the prototype voice app for health data reporting, we encountered some skepticism, concern, and hesitation. However, as we guided participants through the activity, these reactions evaporated. In general, getting started was the most difficult part of this prototype voice app, as can illustrated by Maria's hesitation, *"Oh boy. I'm not good at stuff like this. Okay. What do I have to* *do?"* Betty's reaction immediately after completing the task was expressed with a tone of relief, *"All right. Yeah. All right, I was able to do that."*

Participants thought they would be bothering their doctors.

Participants thought the doctor would also use Alexa to listen to their responses, and that the information was not important enough to bother the doctor. They had the impression they would be annoying doctors by filling out the questionnaire, Betty said, *"I wouldn't want to annoy doctors either with [this], and if I fell down, broke my leg, I would call my doctor anyhow."* Gilda said it was *"a limited application,"* and that she *"doubt[ed] that the doctor would sit there and listen to Alexa."* When we explained that the idea was for Alexa to generate a written report, Gilda responded, *"well, then it might be more applicable."*

One participant feared losing Alexa if she unplugged it.

Maria developed an emotional connection with Alexa, fearing losing that connection by unplugging Alexa. When asked *why*, she responded:

> *"Because I need Alexa. I realized that it's so much that I need from asking the questions. When I wake up in the morning, when I go to sleep at night. I just feel like I've been missing out. I've never had anything like that."* (Maria)

Maria was afraid of losing Alexa once she had developed a connection with it, limiting the control she had over it.

One participant was initially afraid to use Alexa.

Theresa said she was afraid to use it at first, but lost that fear once she had acquired some experience interacting with it. Betty, who speaks on behalf of several of her friends who she says are not as willing to interact with technology as she is, attributes this fear to a lack of confidence in the ability to learn how to use new technologies:

> "*Older people are resistant to technology, not because they don't like it, because they feel they can't learn it. They feel left out of a generation. They've been left out of the learning process. They are frightened of destroying something within the process. Oh, I killed my computer. No, you can't kill it. Just don't throw it out the window. This is a fear.*" (Betty)

Overall, these excerpts point to the negative emotions, even when stemming from positive ones, that a new technology, such as Alexa, may generate in older adults. Keeping these anecdotes in mind may help create voice assistants with more empathy for the feelings of older adults, and may reduce instances of burdensome feelings such as fear.

3.4.7 Over the duration of the study, participants gained knowledge of the voice assistant's capabilities and limitations

This is evidenced through the ideas they generated for new voice apps, a form of empowerment through a participatory design mindset [321]. Every participant had at least one idea for new voice apps or for how the voice assistant could be improved. For example, Maria mentioned that she would like for Alexa to let her know if she had sleepwalked the previous night. She would not want for Alexa to show her videos of the sleepwalking, as that would be scary, but to just notify her what had happened. For example, as an explanation of why she woke up on the couch instead of her bed. Betty mentioned she would like for Alexa to ask her for her opinion about current events. For example, she would like to be asked things like, *"How do you feel about wearing a mask? Would you take a vaccination?"* She suggested a mechanism to do so by relying on daily headlines, and asking "*Do you have any reaction to this headline?*" while having the ability to skip if desired. She explained that a benefit of using a voice assistant for this is that "*nobody's here to judge you [or] to laugh at you.*" Travis had the idea of being able to show Alexa something that is written down, and having Alexa read it aloud to him. These ideas, generated by our participants who had never interacted with a smart speaker before our study, are relevant to older adults and well-scoped for the technology's capabilities, something that evidences the voice assistant-related technical expertise and confidence our participants developed.

3.5 Discussion

Our study generated important findings to inform design guidelines and future work. We now share the design guidelines that others building voice apps for older adults may use, and implications how voice assistants might be useful as always-available agents to enable continuity of care in people's own homes. We also reflect on our research methods.

3.5.1 Design Guidelines

Our findings serve to inform the following design guidelines for future researchers, designers, and developers to use to tailor voice apps to the needs of older adults. Although our study did not explicitly examine how these insights affect others, we anticipate that they will improve the design of voice apps for children, younger adults, or middle-aged adults as well. These are the design guidelines we identified:

Voice apps need to balance expectations of humanlike interaction capabilities and their limited capabilities. This guideline aligns with prior human-robot interaction research [386], and provocations raised by Sayago et al. [313]. Some ways to achieve this are by designing for ambiguity, or by aiding users in rephrasing their utterances to more machine-friendly ones. For example, our voice app only accepted “not limited” as that response option for the ADLs. We were mimicking the written version of the questionnaire, which does not have synonyms listed. However, in adapting the written version to voice, our participants deviated from the phrasing we provided. Betty said “*not affected*” to mean the same thing. Thus, we updated our voice app to also accept “*not affected*,” and any other similar phrases that emerged from user testing, in lieu of “not limited”. Another way to design for ambiguity is by anticipating follow-up questions. While some users may forget the answer options others may not. One way to address this is by programming intents[4] for frequently asked questions, such as the need to be reminded of the answer options. Moreover, voice assistants could aid users in reformulating their requests by providing explanations of how they work. For example, Alexa could have explained that it is better at understanding short requests, and saved Travis some frustration with the questions about his foot. Future work on user-centered explainable artificial intelligence should break down the category of “lay user” [301] by different levels of end-user expertise with the technology at hand.

Scaffolding is needed to accurately remember specific invocation phrases. Our participants struggled to remember the command to launch the voice apps we deployed to their devices. Thus, it is important to develop mechanisms that will help users remember voice app’s invocation phrases. For example, Kim and Choudhury [192] created written lists of important commands that they could

[4]An intent is defined in Alexa’s developer documentation as *an action that fulfills a user’s spoken request*.

place next to the devices. A workaround we used was to include the invocation phrase on the reminders we set for participants. Another workaround for devices with screens could be to set the backgrounds to photos, and include custom-made images with the invocation phrases for the voice apps that are important to the users. Another idea is for voice assistant manufacturers to allow users to curate a list of favorite commands that could be easily pulled up upon request (or, less effectively, a voice app could be designed to fulfill that purpose).

Proactive interactions help ensure compliance, but they can backfire it they are difficult to control. Reminders helped participants remember to use the prototype voice app for health data reporting outside of interview times. The need for proactive reminders is supported and has been discussed in detail by Chen et al. [91]. However, when developing voice apps we found through internal testing that integrating proactive reminders through the voice app resulted in a burdensome user experience for the end user by making it difficult to remove or cancel those reminders. Because of it, we decided not to integrate this feature into our voice apps. While we wait for the technology to integrate proactive reminders to improve, it is important to use workarounds such as asking users to set the reminders themselves, as online workers did in Cuadra et al. [113]'s study.

Experience helps the fear of voice assistants dissipate. Our participants became more comfortable using their devices and suggesting ideas as they became more familiar with how the voice assistant worked. The fear of breaking the device, or otherwise doing something wrong, may prevent users from trying things.

Thus, it is important to create clear expectations, show examples, and practice important actions. A voice assistant could explain, “I do not listen unless I hear my name. If you want me to stop talking or playing music, you need to say my name loudly, or use the buttons on my screen. It is also okay to unplug me. I will be back when you plug me back in.” This could have alleviated Maria’s fear of losing Alexa’s companionship if she accidentally unplugged the device. Similarly, a care provider could show a patient the type of report the voice assistant automatically generates, reducing concerns users may have of bothering their care providers.

Different modalities need to support each other and communicate the same message. We noticed that even though participants used both text and voice modalities, the modalities served different purposes for our participants, depending on their needs. Travis, who was low-literate, focused on the images on the healthcare questionnaire, while the other participants focused on the words when looking at the screen (which is important, given well-documented age-related declines in *working memory*, [119]). Our images did not explain the text; they simply represented the daily activity in question, limiting the value the screen could have for low-literate people to help them remember the possible answer choices. This shows that careful attention, paired with substantial testing, should be employed to ensure that different interaction modalities support each other and communicate consistently, in ways that are inclusive of people with different interaction needs.

3.5.2 Enabling continuity of care in older adults' own homes

Future work should dive deeper into examining how voice assistants may enable continuity of care in older adults' own homes. Our prototype voice app for health data reporting was based on the idea of enabling older adults to use a voice assistant to complete standardized health assessments, that might otherwise require a visit to/from a care provider, from the comfort of their own homes. In doing so, we were interested in the potential for a voice assistant to act as a proxy for human care providers, helping to deliver *continuity of care* [162, 67, 161] for older adults. According to Haggerty et al. [162] *continuity* is "the degree to which a series of discrete healthcare events is experienced as coherent and connected and consistent with the patient's medical needs and personal context." They describe how continuity of care is distinguished from other attributes of care by two core elements—care over time and the focus on individual patients [162]. There are three types of continuity of care: management, relational, and informational [162]. Literature in medicine and public health has shown that continuity of care has been associated with improved patient outcomes and satisfaction [358, 153, 224, 161].

In our study, we see the potential for voice assistants to play a role in providing continuity of care for older adults. First, voice assistants could help provide *management continuity*, the idea that a patient experiences a consistent and coherent approach to the management of a health condition that is responsive to their changing needs. Management continuity is especially important in chronic or complex clinical diseases that require management from several providers to provide a sense of predictability and security in future care [162, 161]. Voice assistants could be leveraged to provide patient-centered scaffolding and guardrails needed for care plan management. For example, a voice assistant could ensure that a health assessment is delivered every day, at the right time, and if not completed, it can continue reminding the patient to complete it. The same applies to tasks such as taking medication or engaging in exercise or rehabilitative activities. Gilda dutifully completed her "homework" to use the prototype voice app for health data reporting outside our interview times thanks to Alexa's reminder, and she loved being reminded to drink water every day (see her description in Section 3.3.1). Because interactions are easily tracked, relevant information could be made available to appropriate care providers, enabling them to understand patient compliance. If a patient's needs change, the care plan could be adjusted by the relevant provider and sent directly to the patient's voice assistant, and the changes immediately reflected in the patient's file. Using the voice assistant as a way to centralize home health care plans from multiple providers could help ensure that interdependent information and treatments are consistent and coherent.

Second, voice assistants could help provide *relational continuity*, or the idea that an ongoing therapeutic relationship, that is built on interpersonal trust, persists between a patient and one or more providers [162, 161]. Recent literature has uncovered how many older adults develop relationships with their voice assistants, seeing them as friends or companions [105, 192, 293]. We saw this as well through our prototype voice app for positive reminiscing; Theresa wanted the questions to be rephrased in a friendlier manner, and Betty wanted to be able to vent to the voice assistant and share her opinions about the news. We also saw this in the emotional connection Maria formed with Alexa (see her description in Section 3.3.1). Studies have found that having continuity [224] and a positive relationship with doctors [190] increases medication compliance. Similarly, having relational continuity and a positive relationship with a voice assistant, as our participants reported and desired, may increase compliance with health-related tasks, such as filling out assessments or completing required activities. Regardless of the source (e.g, a primary care physician, physical therapist, or psychologist), at-home tasks can be consistently delivered using the voice assistants' familiar voice and visual language, creating a sense of relational continuity. Moreover, unlike a human who may be pressed for time, voice assistants have "psychological superpowers" [371] that allow them to be always-available for their users, not to get impatient, and to be perceived as non-judgmental. Voice assistants could be designed to allow plenty of time between conversation turns, and repeat or rephrase utterances as many times as needed without getting annoyed, strengthening the sense of trust in the voice assistant serving as an interim proxy for human healthcare providers.

Finally, voice assistants could help provide *informational continuity*, or the idea that a patient's current care providers are aware of their prior history and present circumstances, including tacit knowledge of patient preferences, values, and context [162, 161]. Voice assistants could help to build more robust information about a person's preferences, such as what kinds of activities they enjoy and their individual context. For example, voice assistants could be used to capture ecological momentary assessments (EMAs), which are approaches for assessing behavioral and cognitive processes in their natural settings [335]. EMAs could help evaluate and improve the treatment of health problems that may benefit from repeated measures as they fluctuate over time, such as pain. Voice assistants could also be effectively used for making daily plans based on higher-level goals [113] to support health and wellbeing, as we see from our participants establishing exercise routines or healthy habits like drinking water.

All this said, we also urge caution and acknowledge the many communication breakdowns that voice assistants are prone to, which require further work if voice assistants are to effectively act as proxies for human care providers. Alexa, having a humanlike voice, failed to understand participants in cases where humans would. These breakdowns have been studied in existing literature [62, 103]. However, their impact may be greater when a voice assistant is being relied on for continuity of care. The example where Travis explains problems with his left foot and Alexa fails to respond could break a person's trust. This is an inherent complication of attributing human characteristics to voice assistants, as doing so may result in misplaced expectations that the voice assistant may respond like a human would. Grudin and Jacques [160], in reference to the Uncanny Valley [250], label situations like Travis' in which "a bot that is knowledgeable within a narrow task focus often cannot answer a query on a related topic that any human expert could" as the Uncanny Cliff. Thoughtfully designed voice assistant self-repair mechanisms could help alleviate some of these problems [115]. Another alternative could be to design voice assistants with more narrow functionalities, such as one that *only* serve to deliver questionnaires. Currently, we at least know that older adults tend to devise strategies to overcome functional errors as they become more familiar with voice assistants [192], errors that will likely decrease as the technology matures. Future work should focus on reducing the need for older adults to adapt to new technologies, rather than the other way around, which may increase their potential for positive impact.

3.5.3 Reflections on our methods

Our research has demonstrated how a small-scale study with thoughtful attention to each individual generated important design guidelines for the design of voice assistants and of voice apps tailored to older adults. We showed how taking "an approach that focuses on engagement with a view to empowerment," as recommended by [305], empowered our participants to come up with suggestions and ideas for improving voice assistants for older adults. Schneider et al. [321] describe this form of empowerment as empowerment through the design process. This type of work serves to amplify the voices of older adults in HCI, and should be continued as more designs are specifically tailored to meet the needs of older adults.

Conducting this study also made us reflect about how else to support older adults desiring to learn how to use a new technology. In our study, we obtained informed consent to own the accounts in the devices and thus had access to all interactions. However, outside a research setting, giving people full access to a person's Amazon or Google account opens up vulnerabilities to privacy violations, threatening their agency. This said, voice assistants' digital ecosystems involve a variety of interfaces, such as the "Alexa app", that are much more difficult to use and understand than simply talking to a machine. It is thus important to consider alternative paths to support older adults who can no longer manage their own digital ecosystems; for example, via features that allow layered access to older adults' accounts for trusted individuals. These features would give trusted community members, friends, or family members partial access to accounts that would allow them to perform important actions for the person they are supporting, such as installing specific voice apps, without having access to a person's full digital trail.

In addition, we found that having the same researcher provide technical support, answer questions, and suggest ways in which to use the voice assistant empowered our participants to comfortably use this new technology, and come up with ideas about how to improve it. Each engagement took about five hours, including drop-off. A similar setup could be used to develop mentorship-style volunteer programs for people who have developed expertise in voice assistant technologies to "mentor" older adults in need of additional technical support. Doing so could help older adults who may feel disempowered by poor technology design feel empowered to advocate for better, more inclusive technology.

3.5.4 Limitations

Our study has several limitations. For example, our prototype voice apps were not entirely functional, so our participants did not experience important aspects of the ideas, such as doctor commenting on their results. This said, our findings provide the groundwork to test voice apps for older adults in more realistic scenarios. Our study is also a small scale, research-through-design study conducted in an urban setting in the U.S. We chose to engage deeply with only five participants, over multiple interactions lasting several months. Our participants had a technical-support contact person available to answer any questions they had and resolve any problems that came up, which is an unrealistic situation for many older adults. In reality, seeking support for these devices requires some technical expertise and may take a long time. Another limitation is that all of our participants were relatively healthy. Understanding healthy older adults' interactions and struggles is necessary first step to promote adoption and prevent systematic exclusion of certain populations. An exciting area of future research would be to conduct a similar study with people who have varying levels of frailty, ideally with appropriate medical partners.

3.6 Conclusion

We contribute an in-depth exploration of two prototype voice apps to support aging in place. Participants were able to effectively use the prototype voice apps, and saw the value in using Alexa for health and wellbeing in a way that they had not imagined or mentioned before using our prototype voice apps. We found that participants needed some sort of scaffolding or reminder to use the prototype voice apps outside of interview times, and that the display helped support health data reporting. We surface design challenges entailing the use of voice assistants for health data reporting, such as participants' perception that they would be bothering their doctors. Despite some initial challenges, engaging with our prototype voice apps empowered participants to contribute ideas for improvements and for other voice apps that would be useful to them. Our work extends the existing literature surrounding the design of artificial intelligence interfaces to support aging in place, and calls for more research entailing the use of voice assistants for enabling continuity of care in older adults' own homes.

In this chapter, participants overcame many challenges interacting with voice assistants with my support. In the following chapter, I further explore the challenges that older adults face as they interact with voice assistants by closely observing older adult first-time encounters with voice assistants in a public setting with more limited support and with more participants.

CHAPTER 4

VIDEO ANALYSIS OF OLDER ADULT INTERACTIONS WITH A MULTI-MODAL VOICE ASSISTANT IN A PUBLIC SETTING

This chapter is the second of two chapters with empirical studies that explore the use of voice assistants by older adults. In the last chapter, I studied older adults interactions with voice assistants through an intervention in a private-setting with participants that transitioned from novice to more-experienced users. In this chapter, I take a close look (second by second) at public-setting interactions with mostly novice users. This is important for two reasons. First, support is not usually as readily available as it was in the first chapter, which may increase exclusion. This chapter more realistically examines the challenges that may arise. Second, as voice assistants permeate the public sphere, avoiding them may become more difficult, and challenges associated with interacting with them may further increase exclusion. The content in this chapter will be published in July of 2022 as an academic article at the International Conference on Information & Communication Technologies and Development (ICTD). It was written in collaboration with Hyein Baek, Deborah Estrin, Malte Jung, and Nicola Dell. Because of this, I will be switching from the singular "I" to the plural "we" for the rest of the chapter. Note, the writing has been slightly edited to fit this dissertation.

We video record older adults (n=26) interacting with a multi-modal voice assistants while waiting in line at food pantries, and use Interaction Analysis to draw insights from these recordings. We find that by being agnostic to body language, audio-prosodic features, and other contextual factors, voice assistants fail to capture and react to some important aspects of interactions. We discuss design (e.g, interpreting users' posture as a cue to wake the device when they are leaning towards the device) and research (e.g., surveillance trade-offs) implications, and argue for the use of multi-modal inputs with attention to privacy. Designing and training voice assistants to take in and appropriately respond to non-verbal cues may increase their inclusivity, helping them fulfill important needs of our aging population.

4.1 Introduction

As I have argued in previous chapters, older adults' needs are underrepresented in the research and design of voice assistants [334], resulting in challenges and barriers limiting use [348]. While building tailored voice apps is one way of increasing voice assistants' inclusivity, another possible way to do so may be by obtaining information from non-verbal cues, as people do in human-human communication [87, 255].

Humans naturally react to other humans' body language, facial expressions, and acoustic-prosodic features (intonation, tone, and rhythm), often subconsciously. Ekman and Friesen [136] characterized the category of nonverbal acts that maintain and regulate the back-and-forth nature of speaking and listening as *regulators*. Regulator actions occur in the attentional periphery; people perform them without thought, but can recall and repeat them if asked [136]. Despite the human-likeness of voice assistants, non-verbal cues are, for the most part, currently being overlooked by voice assistants. We utilize a framework developed by Suchman that analyzes the information available to the user, the information available to the machine, and their intersection [340]. In this chapter, we refer to the information that is not mutually available to both communication partners (i.e., the human and the voice assistant) as *the human-machine communication gap*.

Although the use of video analysis is common in industry [229], existing research on voice assistant usage by older adults predominantly relies on usage logs, interviews, or product reviews [322, 268, 294, 64, 289]. We analyze the image, audio, and human-machine communication gaps in video recorded interactions of 26 older adults, who are predominantly novice users of voice assistants, with an Amazon Echo Show 10. **In particular, we 1) seek to characterize challenges in interactions with voice assistants that may obstruct inclusion, and 2) identify alternate paths that may mitigate these challenges.**

We chose to conduct our study with older adults who are predominantly novice users of voice assistants for several reasons. Although experienced users may adapt their behaviors over time as they learn how voice assistants respond, the experiences of first-time users are extremely important in determining whether someone will deem it worthwhile to adopt the technology at all [200, 380]. This may be particularly true for older adults, who may be more hesitant to use new technologies. Furthermore, not developing expertise in the privacy of one's home may result in exclusion from everyday digital activities as these technologies permeate public spaces. These encounters could become embarrassing, scary, or frustrating for novices. Moreover, although some older adults may have caregivers who could help them to learn how to use the technologies, such assistance unnecessarily increases dependence. Hence, we studied the difficulties that novice older adult users encounter when interacting with voice assistants, and how we might make these technologies more usable to them.

Our findings reveal gaps in human-machine communication that often result in the voice assistant reacting inappropriately, interrupting the user, or not responding at all. We (1) describe human-machine communication gaps revealed by our data, differentiating information that was overlooked by the machine (e.g., interaction attempts, the presence of more than one user) from information that was overlooked by participants (e.g., the indication that the voice assistant was not actively listening, and technical terminology). We then (2) take a closer look at body language features of the interactions and categorize them into those that provide reliable signals (e.g., leaning forward and gaze), and those that are somewhat ambiguous (e.g., laughing). Finally, we (3) analyze audio-prosodic features, such as rhythm (e.g., interruptions during pauses in speech), and tone and intonation (e.g., associations between various tones and intonations and interaction outcomes). Together, our findings show that by being agnostic to body language, audio-prosodic features, and other contextual factors, voice assistants fail to capture and react to some important aspects of interactions. Designing and training machines that take in and appropriately respond to non-verbal cues might be a crucial step in building voice assistants that can fulfill important needs of our aging population.

We present design and research implications for the HCI community. In terms of design implications, we provide recommendations addressing interaction errors that result from not being able to successfully wake the voice assistant, such as by relying on design paradigms that may be more familiar to older adults. We suggest ways in which automatic detection of non-verbal cues can be used to improve interactions with voice assistants, such as having the voice assistant analyze a user's posture to determine whether they are attempting to engage with the voice assistant. We then emphasize differences and complexities for adapting voice assistants' interactions to older adults' needs and abilities in the context of prior research about code switching and knowing the user [62, 103], and discuss several ethical design considerations. In terms of research implications, we surface questions surrounding how we might use recent technological advancements to recognize body language and audio-prosodic features, and discuss the societal implications and tradeoffs associated with higher levels of surveillance. Taken together, our contributions help to relieve some of the burdens placed on older adults to adapt to the constraints imposed by new technologies, allowing older adults to appropriately benefit from the technologies' promises and improving inclusion in everyday digital activities.

4.2 Related Work

In this section, I add to the related work described in Chapter 2 by specifically covering research surrounding voice assistant use in public settings, and the importance of non-verbal communication.

4.2.1 Voice assistants in public settings

Even though voice assistants are currently mostly used in private spaces (e.g., homes and cars), they are becoming more common in more-public venues (e.g., hotels, schools, and stores) [352, 323]. In an ethnographically-oriented study published in 2017, Porcheron et al. [288] explored how groups of friends interacted with Siri at a coffee shop, identifying insights, such as that participants had to rely on the screen of their devices to share parts of interactions with each other. Similarly, Cowan et al. [109] studied infrequent users of voice assistants, finding that cultural norms affected some participants' willingness to use Siri in public. During the same year that these studies were published, scholars from industry and academia met at CSCW to discuss the use, research, and design of conversational agents, such as voice assistants, in social and collaborative settings, raising the importance of this topic of research [287]. Since then, some have studied voice assistant interactions in multi-user home settings, questioning how conversational voice assistants truly are [289]. Despite general agreement on the importance of studying voice assistant use in public settings, to the best of our knowledge, no one has focused on studying voice assistant use by older adults in public, potentially excluding a growing segment of our population that could highly benefit from, or be excluded by, this technology.

4.2.2 The importance of non-verbal communication in HCI research

Non-verbal forms of communication have been deemed important in the HCI communities for a long time. In 1994, Nagao and Takeuchi acknowledged the multiplicity of communication channels that act on multiple modalities, and set out to study how humans would react to facial expressions from a machine in human-computer dialogue [255]. Shortly after, Reeves and Nass published The Media Equation, supporting the claim that we attribute characteristics to machines in the same way we do to humans [299]. In the same line of research, Cassell et al. analyzed human monologues and dialogues that suggested that postural shifts can be predicted as a function of discourse state in monologues, and discourse and conversation state in dialogues [87]. As a result, they designed an embodied conversational agent that could change its posture [87]. Moreover, Liebman and Gergle examined the role of nonverbal, paralinguistic cues in computer-mediated, text-based communication, such as punctuation and emoticons, and found a positive causal relationship of conversation duration and cue use on perceived affinity, and that reciprocity may play a central role in supporting this effect

[211].

Designing non-verbal expressions for voice agents impacts how humans react to them; research has found that matching the tonality of a voice assistant's speech to the mood of its human user results in better performance [182], that gender stereotypes carry over to gendered synthetic voices [259], and that we consider different voices from the same device to be different social actors [259], mimicking how we may distinguish different people talking on a telephone. Additionally, Jung et al. found that although robots that used backchanneling improved team functioning, backchanneling robots were perceived as less intelligent than those that did not use backchanneling [187].

In the next chapter, we describe how self-repair greatly improves people's assessment of an intelligent voice assistant if a mistake has been made, but can degrade assessment if no correction is needed. However, to the best of our knowledge, no work has successfully examined how voice assistants may interpret non-verbal expressions displayed by their users—for example, to recognize error. This is despite a recent line of work studying the human-likeness of human-agent conversations. Motivated by key characteristics of human-human conversations that do not get captured by conversational agents, Clark et al. studied what features people value in conversation, calling for a redefinition of design parameters for conversational agent interaction [103]. They argue that participants describe the need for mutual understanding and common ground, trust, active listenership, and humor as crucial social features in human conversations, but in agent conversations these are described almost exclusively in transactional and utilitarian terms [103]. Beneteau et al., support this argument by recognizing that to improve communication repair strategies, knowledge of the context and the communication partner is extremely helpful, allowing digital home assistants to artificially code switch as needed [62]. The tension between the human-likeness of voice assistants, and their inability to meet the expectations that their appearance sets might contribute to the fluid movement between "human-like" and "object-like" categorizations displayed by older adults in Pradhan et al's study [293]. Taken together, these studies call for improvements in voice assistants' abilities to understand and react to non-verbal cues, especially because of their implied humanness.

The importance of context in human-machine interactions is well known in the HCI communities [187, 337, 347]. Additionally, we know that behavioral responses to robots, from which context can be extracted, are in a large part nonverbal [195]. Research has also made technological strides in the last decade in sensing [3, 330, 336, 355] and computer vision [367, 181, 93]. With these considerations in mind, this chapter aims to answer the following questions: 1) what do older adult interactions with voice assistants look like, 2) what (mainly nonverbal) information is unavailable or not being interpreted by the machine, and 3) how can we use this information to avoid, recognize, and/or repair errors in older adults' interactions with voice assistants. Our goal is not to identify a taxonomy of repair strategies, which are well known [62], but to identify and valorize the visual and prosodic elements present in older adult interactions with voice assistants.

4.3 Approach

We conducted an IRB-approved field study with older adults who visited senior centers, and video recorded their interactions with a voice assistant. We now provide a description of the settings in which we conducted our observations, details about the participants, and explain our methodological and analytic orientations.

4.3.1 Research Setting

We situate our study in senior centers, which can be categorized as “third place” settings. A “third place” setting is described by Oldenburg [273] as a place where one relaxes in public, encounters familiar faces and makes new acquaintances.[1] We chose this setting as way to capture the heterogeneity of the older adult population while also engaging with a central theme demarcating the ubicomp of the present, the “messiness of everyday life” [61]. Senior centers are community centers designed to make older adults feel supported, and happy—they bring older adults together for a variety of services and activities designed to enhance their quality of life [60]. Both of the senior centers in our study had computer labs with programming to teach older adults computer skills. According to the *National*

[1] Oldenburg [273] calls the "first place" the home, and the "second place" the workplace.

Council on Aging, "Compared with their peers, senior center participants have higher levels of health, social interaction, and life satisfaction and lower levels of income."[2] To capture use in public, we set up research booths with a camera facing the participants (Figure 4.1) near food pantry lines—food pantries offer free groceries to members on a periodical basis—in two senior centers in a large U.S. city. Our "in the wild" [110] approach allowed us to capture public interactions with voice assistants that are becoming increasingly common in public places [352].

4.3.2 Recruitment and Participants

We approached older adults who visited the center and invited them to participate in our study. We explained the purpose of the study, what we were asking participants to do, and sought their permission to video capture their interactions with a voice assistant. Consent forms were available as physical copies placed on a table, and consent was obtained verbally. The researchers followed recommended health and safety protocols during the explorations to keep both participants and researchers safe during the COVID-19 pandemic.

In total, we recruited 26 participants (20 women), who were on average 73 years old. Table 4.1 summarizes participant demographics. Participants were visiting the senior center for the food pantry: some were picking up food and others were organizing the pantries. To pick up food, participants must attest to income

[2] https://www.ncoa.org/article/get-the-facts-on-senior-centers

levels below a certain threshold (typically less than $2200 per month if there is just one person in the household). Senior center staff reported that most members owned smartphones, echoing our participants' responses when asked about their current technology usage. Most participants (*n*=16, eight unreported) owned one or multiple computing devices, including smartphones, tablets, laptops, or desktop computers. They reported using their computing devices for a variety of reasons, including: information retrieval, messaging others, audio and video calls (including doctor appointments), reminders, social media, playing games, viewing or attending religious events, taking photos, playing music, writing, accessing specific websites, shopping online, and paying bills. All participants who owned and used a computing device had access to the Internet. Some participants (*n*=5) indicated using speech-to-text functionality of their phones, tablets, or computers, but none expressed knowing how to send voice notes, such as the ones supported by iMessage or Whatsapp. Most participants (*n*=18, eight unreported) were at least somewhat confident reading and writing; however, three participants expressed declining confidence due to age-related cognitive, motor, or visual impairments. Participants lived in their homes, predominantly with relatives. Most participants (*n*=19) reported never having used a voice assistant before. We considered participants novices if they reported having used a voice assistant before, but did not feel very confident in their abilities using it or whose interactions suggested novice-level expertise. Even though our counts (see Section 4.3.5) include interactions from users with some experience (e.g., P5 or P6), we only use one specific example from non-first-time users in our findings—P5 &

P6 playing Trivia together—, which we call out as such. We included all participants in our interaction counts (including P5, our most experienced participant), because they are representative of the heterogeneity of the older adult population and the "messiness of everyday life" [61]. Additionally, most of our non first-time user participants were still novices.

Because of the in-the-wild nature of the study, some participants arrived in pairs and interacted with the device in pairs (three pairs, n=6), which we see as resembling how real-world interactions with voice assistants might take place (e.g., several people might be in the room where the voice assistant is installed). However, because we segmented the data for analysis, we were able to extract individual interactions from participants who arrived in pairs. In most cases, one participant spoke while the other listened. In rare cases, participants responded in unison, these segments were annotated accordingly. We kept an eye on potential influences paired individuals could have on each other, and made note of them in the findings. However, for the most part, since all participants interacted in public, they all knew they were being watched, providing something of a control for potential behavioral differences caused by The Hawthorne effect [235].[3]

[3] The Hawthorne effect refers to a type of reactivity in which individuals modify an aspect of their behavior in response to their awareness of being observed.

Participant demographics (*n*=26)	
Age	Avg: 73, Median: 74, SD: 7.74
Gender	Female: 20, Male: 6
Language used	English: 18, Spanish: 7, Korean: 1
Latinx	No: 15, Yes: 11
Race	Black: 10, White: 3, Native American: 2, Asian: 1, Other or mixed: 9, Declined to answer: 1
Prior experience with a smart speaker	First-time users: 19, Non first-time users: 7
Confidence using speech-based computing device (after interaction)	Very confident: 6, Somewhat confident: 1, Only a little confident: 1, Not at all confident: 10, Unreported: 8
Highest degree or level of school you have completed	Less than a high school diploma: 2, High school degree or equivalent: 8, Some college - no degree: 5, Bachelor's degree: 3, Master's degree : 1, Unreported: 8
Gross income ($)	<20k: 11, 20-40k: 6, 80-100k: 1, Unreported: 8
Living alone	No: 11, Yes: 7, Unreported: 8
Own and use at least one computing device	No: 2, Yes: 16, Unreported: 8
Frequency of use of computing devices to go online	Less than once a week: 2, About once a week: 1, About once a day: 4, Multiple times every day: 11, Unreported: 8
Confidence using computing device	Very confident: 6, Somewhat confident: 5, Only a little confident: 3 , Not at all confident: 4, Unreported: 8
WiFi at home	No: 7, Yes: 11, Unreported: 8
Confidence reading and writing	Very confident: 15, Somewhat confident: 3, Unreported: 8

Table 4.1: Demographic details of study participants.

4.3.3 Procedure

The booths included signs indicating we were conducting a research study, a voice assistant, and a camera from the perspective of the voice assistant. The voice assistant was placed on top of a table, and a chair was positioned nearby for participants to have the option to sit. We told participants that had never interacted with a smart speaker before that the device on the table was a smart speaker that responded to speech, and explained that they could initiate conversations with it by saying its name, Alexa, followed by a command. We temporarily muted the device to provide utterance examples such as, “Alexa, hello” or “Alexa, what’s the weather”.[4]

After receiving this guidance, participants were instructed to freely interact with the voice assistant, and we pointed at signs with utterance suggestions. These signs were posted on the wall behind the table the device was on. The messages on the signs suggested participants to say “Alexa, hello,” “Alexa, what are the symptoms of COVID-19,” “Alexa, what can you do,” and “Alexa, what’s the weather.” The first author was available throughout all the sessions, usually sitting somewhere near the participant but outside the participant’s field of view. The researcher occasionally provided support to participants, such as when a participant seemed stuck, was unable to wake the device, or looked at the researcher for guidance. Often, even if participants seemed to be getting frustrated, the researcher would simply suggest that they keep trying. For example, we did not intervene in the three occasions in which participants introduced themselves to Alexa, and Alexa initiated a voice training “setup” activity. But we did intervene if Alexa was not responding at all after several failed attempts, encouraging participants to speak louder, sometimes escalating the suggestion by telling participants to imagine they were upset at Alexa. Whenever possible, we asked participants what they had noticed or thought of the interaction, or if a request had gone as they expected.

Figure 4.1 depicts the setup used. We used an Amazon Echo Show 10 device,

[4] We skipped this step for non first-time users.

which has a 10-inch touchscreen, set to a default American female voice and wake word "Alexa". We chose this embodiment for our voice assistant research because the touch screen complements the audio modality, providing additional information for older adults who are prone to experiencing a variety of cognitive, audio, visual, and motor impairments. The voice assistant was configured to speak in either English or Spanish, depending on the language used to address it.

Study sessions lasted approximately 30 minutes including obtaining consent, the initial introduction, the interaction itself, and the post-interaction interviews. However, sessions in which too many interactions failed from the very beginning were much shorter, as participants did not wish to continue engaging. After capturing the interactions on video, we clipped the videos into smaller segments we called "chunks" (described in Section 4.3.5 along with the resulting dataset).

4.3.4 Video Analysis Methodology

We used video analysis methodology to carefully study the interactions of older adults with the multi-modal voice assistant. By observing these interactions, we hoped to uncover important insights for the design of voice assistants, which tend to be used in the private space of a home, where the visual elements of interactions are not usually captured. Many have employed video analysis methodology to capture patterns that would not be visible without video (e.g., by playing the video at slow or accelerated speeds), collect primary empirical data, and have

Figure 4.1: A participant interacts with the voice assistant at a senior center. A camera behind the voice assistant records the participant. Wooden panels label the booth as a research study, and provide suggestions with example utterances to interact with the voice assistant.

more consistency and reliability in observations [185, 339, 340, 342, 368, 231].

Our work utilizes Interaction Analysis [185]. According to Jordan and Henderson [185], Interaction Analysis commits to grounding theories of knowledge and action in empirical evidence with the goal of identifying "regularities in the ways in which participants utilize the resources of the complex social and material world of actors and objects within which they operate." We chose this method because it would allow us to reconstruct the events, keep and replay the primary record, and capture the complexity of the data.

4.3.5 Data Analysis

After we captured the videos, we employed a bottom-up approach to analyze the data. We began by watching all the videos and created a rough content log, as described by Suchman and Trigg [339]. Then, we watched all the videos at 2x and 4x speeds to see if anything stood out, as replaying the videos at different speeds can help to see patterns that were otherwise not noticeable [185]. By doing this, we were able to identify our a unit of analysis, which we called a *chunk*. We based this decision on previous work by Weingart et al., which identify "units" to be coded [369], and Jordan and Henderson, which relies on "ethnographic chunks" to break down large videos into smaller, more analyzable, segments [185]. The first chunk of a participant's interaction always started when a participant addressed Alexa for the first time, and ended when there was an interruption or the participant addressed Alexa again. Subsequent chunks started at the end of the previous chunk (see Figure 4.2 for an example). In parallel, we selected our own analytic foci: body language, audio-prosodic features (such as tone, intonation, and rhythm), and human-machine communication gaps. Jordan and Henderson define analytic foci as "ways of looking that are quite consistently employed in Interaction Analysis" [185].

- **Body language**: For every interaction, we carefully annotated all aspects that visually changed during an interaction including aspects relating to gaze, posture, and facial expressions. We then ascribed meaning to these, and evaluated whether they were signs that the conversation was going well or poorly. Based on these assessments, we extracted body language features, such as gaze and posture, that we were able to consistently identify and make inferences from, and ones that were more difficult to tell apart, such as laughter.

- **Audio-prosodic features**: We categorized every chunk based on its rhythm, tone, and intonation. For rhythm, we noticed that participants often did not pause or paused for too long to fit Alexa's listening window, so we labeled chunks accordingly. We open-coded the tone as Upset, Nervous, Friendly, Exaggerated, Indifferent, Excited, or Tired, and the intonation as Fall, Rise, Rise-Fall, or Same (Constant). Two authors watched multiple similar clips together, discussed possible descriptions for those clips, and subsequently agreed on the aforementioned codes for their tone and intonation. We then used these annotations to inform our inferences about participants' behaviors during interactions.

- **Human-machine communication gaps**: We noticed that the machine was missing a lot of important signals emerging from non-verbal communication. Thus, we employed Suchman's analytic framework [340] to compare and contrast the information available to both the older adults and the voice assistant, with the information only available to one or the other. An excerpt of what this analysis looked like is available in Figure 4.2, and immediately noticeable from these two chunks (amounting to only 32 seconds of interaction) is the quantity of user actions, in particular non-verbal, that were not interpreted by the voice assistant (everything not highlighted in yellow).

Chunks were labeled based on their outcome: "Success", "Failure", or "Ambiguous," and ones that were parts of interruptions were marked as "Help," or "Unrelated." Clipping, numbering, and appropriately labeling interruptions or interventions allowed us to exclude them from the analysis while still noting that an intervention had occurred right before a particular interaction. Doing so was important to maintain the order integrity of each participant's interactions at a higher-level while being able to code and analyze each chunk in detail. Our criteria were:

- **Failure:** The interaction failed. For example, if the participant did not succeed at "waking" the device, or if the device misheard a participant's request.
- **Success:** The interaction succeeded—a user made a request, and the machine responded appropriately. At some point, there was an agreement in understanding from the user and the machine. For example, the user asked for a joke, and the voice assistant told a joke.
- **Ambiguous:** These happened when the goal was not clear, or the success was only partial, so we were not able to classify them as successes or failures with certainty. For example, a chunk in which a participant said "*thank you*" to the machine was classified as ambiguous, as the machine did not respond to the user, but no response was necessarily expected. Also classified as ambiguous were chunks based on the context of previous interactions. For example, if a participant had been ignored for several chunks, and the

machine finally responded to them but not with what they were asking. In this case, getting a response from the machine after having been repeatedly ignored was considered a partial achievement, rather than a clear failure.

Once we clipped all the interactions into chunks, we coded each chunk. In our dataset, each chunk has an index number, a participant number, a chunk number (starts at one for each participant, except for participants interacting in pairs), a duration, and an outcome. When the wake word was said, we labeled how it was pronounced, and the order in which it was pronounced. Audio prosodic features (rhythm, tone, and intonation) were marked for each chunk, and body language features were noted when they occurred. In some instances, we played the clip aloud in front of an active Alexa device to reconstruct the event, and verify that our codes were accurate. Similarly, we reviewed usage logs from the interactions to see how the voice assistant interpreted the information. Each chunk was initially coded by one researcher and then reviewed by another researcher. All disagreements in the codes were discussed until agreement was reached. We do not report inter-rater reliability since all data was double-coded and disagreements were reconciled [236].

Summary of dataset: The dataset we gathered included 221 interaction chunks (56 minutes and 37 seconds of footage) from 26 participants. The longest interaction had 44 chunks, lasting 18 minutes after excluding “help” or “unrelated” clips. The shortest interaction had one chunk, lasting nine seconds. Out of these, 68 were labeled “Success”, 92 “Failure”, and 61 “Ambiguous”. We excluded from the dataset 47 chunks in which participants were not interacting with the voice assistant, but marked their position to account for interventions and/or interruptions. Appendix .0.3 provides a summary of observed behaviors coded in our dataset.

4.4 Findings

We start by (1) describing human-machine communication gaps revealed by our data, broken down into information that was overlooked by the machine (e.g., interaction attempts, the presence of more than one user) from information that was overlooked by participants (e.g., the indication that the voice assistant was not actively listening, and technical terminology). We then (2) take a closer look at body language features of the interactions and categorize them into those that provide reliable signals (e.g., leaning forward and gaze), and those that are somewhat ambiguous (e.g., laughing). Finally, we (3) analyze audio-prosodic features, such as rhythm (e.g., interruptions during pauses in speech), and tone and intonation (e.g., associations between various tones and intonations and interaction outcomes).

Figure 4.2: Annotated events over a period of 32 seconds (two chunks). A large part of the interaction richness, that could theoretically be made available and interpreted by the machine using existing computer vision and/or sensing technology, is unavailable to the the machine. The only part that is available and interpreted here is the text highlighted in yellow, "Alexa what's the weather outside?"

4.4.1 Human-machine communication gaps

Our analysis highlights gaps in communication between participants and the voice assistant that led to interaction challenges. In some cases, these gaps corroborate prior work reporting similar challenges [62, 103, 288, 289]. Our goal here, however, is to ultimately show how considering audience, body language and/or audio-prosodic features might help to overcome these gaps. At a high level, we found that older adults' challenges interacting with voice assistants were often due to a lack of conversational grounding, where the voice assistant did not understand older adults' expressions, and participants did not reliably understand the voice assistant's state. In this section, we describe data that was overlooked by the machine and data that was overlooked by our participants. Note, these are not necessarily a comprehensive list of all gaps, rather, they are examples that stood out to us as most relevant for inclusion.

Overlooked by the machine

Second by second interaction data. Our analysis reveals valuable information overlooked by the machine.Standard video is usually shot at 30 frames per second. If we look at just one frame for every second in only two interaction chunks, reducing the data to $1/30^{th}$ the size, and label a participant's gaze, posture, and facial expression we can make many inferences about an interaction (Figure 4.2). By adding context from previous interactions, what is displayed on the voice assistant's screen, and participants' audio-prosodic features, we can infer even more.

For example, leaning forward while directing their gaze at the voice assistant (t=1s–3s) can be used to infer that the participant is engaging with Alexa. The tilting from side to side while the voice assistant is speaking (t=9s–17s) can be used to infer that the participant is listening. The nod and eyebrow raise at second 20, four seconds after Alexa finished speaking, can be used to infer that something went well. Directing the gaze at the voice assistant's screen can be used to infer that the participant is reading content on the screen (t=21s–24s), especially if side-to-side eye movement is also detected. The content on the screen can be used to infer what the participant might refer to in their potentially upcoming utterance. The laughing before directing their gaze downwards and then to the left (t=28s–32s) can be used to infer that something went wrong. The rising intonation (t=24s–26s) can be used to infer that a question was asked. And so on. These second-by-second interactions are currently not taken into account by the voice assistant, resulting in numerous interaction problems that we discuss below.

Interaction attempts. Despite multiple cues from participants that signal interaction attempts, these attempts were frequently missed by the voice assistant, which is programmed to respond only when it has heard its wake word or is engaged in multi-turn interactions (e.g., while using the Trivia voice app). Participants frequently did not use the wake word (Alexa) adequately. The only way to appropriately wake the voice assistant with a voice command is by pronouncing the wake word in a specific way and before saying the request. The wake word was either mispronounced or omitted in 70% of the Failure chunks, more than a quarter of all interactions. A few alternative pronunciations of the wake word were used, including: *"Alexia"* (*n*=2), *"Alexis"* (*n*=2), *"Alessa"* (*n*=3), and *"Alexi"* (*n*=2); note, these are all from different participants, except one who called Alexa both, *"Alexia"* and *"Alexi"*. Another reason for failing to wake the voice assistant was not saying the wake word before the request—seven participants said the wake word last at least once, and 13 omitted saying the wake word when initiating an interaction at least once. These do not include omissions that were not clear failures, such as Ambiguous "thank you" chunks. These interactions might be improved via more intuitive ways to wake the voice assistant and understand when it is paying attention.

Interactions with its other modalities. Voice assistants with screens, such as the Amazon Echo Show used for this study, have displays that may provide suggestions for how to interact with the voice assistant. Many participants used the prompts displayed on the screen (generic prompts generated by the voice assistant, not the research signs we posted on the wall) to interact with the voice assistant, but not everyone understood that these were mere suggestions. By looking at the participants' gaze and side-to-side eye moment, a human can tell that the participant is reading prompts on the screen. However, Alexa did not do this. Furthermore, Alexa's responses were also agnostic to what it was showing on its display. For example, if a participant asked for a recipe based on what the voice assistant was showing on its screen, Alexa started offering new recipe options for that type of food, instead of showing the specific recipe that was requested in response to the display's content. P15 fell into a repetitive loop, treating the suggestions as instructions. These findings suggest opportunities to better establish conversational grounding by connecting interactions to displayed content. In her post-interaction interview, P15 said how displeased she was with the voice assistant and said she would never use one, also demonstrating the importance of these initial interactions for adoption.

The presence of more than one user. We found that participants interacting in pairs sometimes reacted, or were influenced, by each other's interactions. For example, P14 & P26 were a couple interacting together. In one of the initial interactions P14 greeted Alexa. Then P26 greeted it too, this time introducing herself, and P14 briefly directed his gaze at P26 while she spoke. P14 then introduced himself as well, as if copying P26. The introduction prompted Alexa to start a voice training to learn an individual person's voice. However, P14 & P26, who were interacting as a pair, responded to Alexa's commands in unison, undermining the purpose of the voice training. As another example, P5 & P6 (who were not first-time users) engaged in a game of trivia, and had to devise silent strategies to communicate with each other about which answer to select to avoid Alexa prematurely recording a response. This made the interaction burdensome. Alexa missed important interaction data: the number of people interacting with it and their interactions with each other. If it had not overlooked this data, it might specify who a request is directed at, avoiding confusion, or know that an utterance is not directed at it. Moreover, if an issue continued, for example if users continued to respond in unison during the voice training, Alexa could interrupt to repair the issue, such as by explaining why it is better for only one person to respond at a time.

Social norms. We also saw cases where the voice assistant did not respect social norms at play. For example, P26 introduced herself to ask for the temperature, *"Alexa, my name is [P26], and I want to ask you how the temperature will be today"* (P26). Alexa was not able to interpret P26's introduction as a formality preceding an actual request for the weather. Instead, Alexa interrupted P26 after she said her name and in the middle of the request for the temperature, setting off a voice-recognition training.

Overlooked by humans

The voice assistant's indication that it was not actively listening. Alexa's blue line, which is used to signal different states (such as listening, processing a request, responding, or not actively listening) via its light and color patterns[5], did not appear to be an intuitive indicator of the voice assistant's conversational state

[5] https://www.cnet.com/how-to/what-do-the-light-ring-colors-on-your-amazon-echo-mean/

for older adults; many did not even notice it. Participants often talked to Alexa when it was not displaying the blue line that indicates it is actively listening. For example, if they had a follow-up question, they would just ask it without "waking" Alexa and waiting for visual confirmation that it was listening. For example, P13 asked Alexa, *"Alexa, tell me about exercises for back pain?"* Alexa responded. Then, without making sure Alexa was listening, the participant asked *"How about specifically for older adults?"* and Alexa did not respond.

Moreover, some participants talked to Alexa at length without ensuring Alexa was listening. For example, P22 relayed in one conversational turn,

> *"My name is P22, how are you doing over there? All I can say is this mask doesn't save me. What I s[sic], what I do, I can't breathe. I like to take it off most of the time because I can't breathe. I'm almost 65. I'm retired. I'm happy. That's all I can say. I'm happy. I'm retired. I should have retired earlier than 65. Having a great time here. Retiring is great. So all who don't enjoy, sorry. I'm enjoying retirement over 65. Who cares about Coronavirus? Doesn't bother me. I use my vitamins, my juice, all the vitamins, juice and healthy food as long as they're available at nice healthy restaurants. Have a great day. God bless!"* (P22)

P22 did not establish conversational grounding with Alexa, probably because he did not know to look for the blue line indicator. Alexa, in return, did not respond to P22.

Technical terminology. The use of technical terminology likely contributed to many misunderstandings. For example, Alexa would say, *"Once I learn your voice, I'll be able to call you by name, tell you apart from others who use the Echo device you speak to, and personalize your experience. First, you'll need your own profile. I can create one for you now."* The way a machine learns a voice is different than the way humans do, so these differences must be described to someone who does not know how machine learning works. Moreover, the next part of Alexa's explanation is even more confusing to someone who is new to these types of technologies, *"As part of learning your voice, I'll ask you to say four phrases to create a voice profile. Your voice will be stored in the cloud until you delete it in the app."* Creating a voice profile, storing it in the cloud, and deleting it from the app are all explanations that assume technical familiarity. The communication gap this created was demonstrated by participants' expectations that Alexa would fulfill the requests they were making through the voice training, and by their actions, such as when they repeated an utterance that was not intended for the training (see the last paragraph of this section).

Requests to use other technology. Frustration and blank expressions were also common when Alexa required participants to know how to use other platforms. For example, Alexa made a suggestion to ask about Amazon orders, *"I didn't get that. By the way, there's lots more to discover. For example, I can keep you up to date on Amazon orders."* When the participant followed through and asked about her stuff, Alexa responded, *"I didn't find any open orders for Participant. If you're waiting for a* *delayed package, you can check the status at the orders page on Amazon.*" Checking the orders page on Amazon is unfeasible for many who might be relying on a voice assistant as their gateway to the Internet. The participant's reaction was to laugh in dismay, and choose to end the activity (i.e., leave). In another example, Alexa once again asked a participant to try activities that they were unable to try without having access to a smartphone app, *"Okay, here's Activity Book. To use Activity Book a parent needs to give permission. To do that, I sent some information to the home screen of your Alexa app.*" In this case, the participant had a blank expression and tried something new. In another, more navigable example, Alexa asked a participant to *"please select a default browser.*" To know what a "default browser" is requires technical familiarity, but at least in this case there were only two options to pick from, meaning that there was a way to select a browser even without knowing the differences between the options.

That voice training activities were setup activities to make voice profiles. As a result of the unstructured nature of the study, three participants (P14, P15, P26) ended up completing a voice training. Alexa would launched the training when participants introduced themselves. During this activity, participants repeated Alexa's commands, but did not grasp that this was for training the voice assistant. Alexa took control of the interaction, and participants diligently followed Alexa's instructions. A couple participating in this task together did not question the activity (at least in front of us) but one participant did seem increasingly frustrated about Alexa continuing to ask her to say things. Her dismay was betrayed by her upset laughter and raised eyebrows, gazes towards us to request help, and confused expression. At the end of these interactions, Alexa offered advice about how to help others, "a*nd if you'd like to help others get recognized on these devices, remind them to say, 'Alexa, learn my voice.'*" In all instances (P15, P14 & P26), despite just having trained Alexa to recognize their voices, participants responded, *"Alexa, learn my voice"*, suggesting that they had not understood the purpose of the activity.

4.4.2 Body language

In most cases discussed above, there were visual cues in participants' body language available that helped us, the researchers, diagnose conversational problems that could use repair. For example, had Alexa seen P22 looking and talking at it, it could have responded to him after he said *"God bless!"* Similarly, had Alexa noticed P26 was not done speaking after introducing herself, it could have waited to respond. In this section, we take a closer look at the body language that was expressed in these interactions.

Leaning forward, gaze, and nodding. There often were clear visual indications of when a person wanted to interact with the voice assistant, such as leaning forward and looking at the voice assistant, but the voice assistant overlooked them. Most participants (*n*=17) leaned forward at least once. When leaning forward, participants also directed their gaze at the voice assistant. In total, we identified 77 instances of participants leaning forward. Moreover, we noticed this behavior in a picture from a separate study with participants that seemed younger by Porcheron et al. [288], where a participant leans forward to speak to Siri on an iPad (see the bottom right picture in page 214 of their paper). We noticed that after a failed interaction, predominantly Alexa not responding, participants would lean forward, closer to the voice assistant. Nearly half the participants (n=12) leaned forward as a form of conversation repair. After a successful repair attempt via leaning forward, some participants would continue to lean forward in subsequent interactions (e.g., P4, P9, P13). Once Alexa responded, the tendency was to return to their initial position, and lean forward again for the next request. We also noticed instances of leaning forward in which no error had happened, suggesting leaning forward also occurred as a form of heightened engagement. For example, several participants (n=7) leaned forward towards the voice assistant when it was speaking, possibly to hear better. Similarly, some participants (n=13) leaned forward when initiating an interaction, possibly to ensure that the voice assistant could hear them or to signal that they were speaking to it. Another consistent interaction was nodding, which signaled that a positive interaction had occurred, suggesting either a pleasant surprise, being impressed, agreement, or affirmation. Given these findings, leaning forward while directing ones gaze at the device may be an important body language feature to recognize as an alternate form of "waking" voice assistants. In addition, nodding slightly and briefly could be used to automatically mark interactions as successful, to train voice assistants and to avoid repair [115].

Other forms of body language were somewhat ambiguous. We found that gestures such laughing, raising eyebrows, furrowing eyebrows, waving hands, and looking away could signal positive and negative interactions alike. The differences in the gestures themselves were too subtle, sometimes unnoticeable to us, to rely on them alone. For example, P15 laughs when she is caught in the loop of asking the same question over and over again due to thinking that the interaction suggestions were instructions, suggesting frustration. By contrast, P18 laughs when Alexa finally responds to her, suggesting relief. Alone, these reactions can perhaps be too difficult to interpret, but when more data is available, inferences can be made with more certainty, and their presence can signal an interaction event worth analyzing. For example, we can infer the valence of these actions from understanding interaction context—P15's misunderstanding of what is happening, and P18's previously unrecognized attempts—that was available to us, the researchers, but not interpreted by the voice assistant.

4.4.3 Audio prosodic features

In this section, we take a closer look at the rhythm, tone, and intonation in participants' speech patterns during their interactions with the voice assistant.

Tone	*n*	Success	Failure or Ambiguous
Exaggerated	8	71%	29%
Excited	10	26%	74%
Friendly	11	59%	41%
Indifferent	5	36%	64%
Nervous	8	22%	78%
Neutral	42	53%	47%
Tired	1	100%	0%
Upset	3	100%	0%
Intonation			
Fall	14	32%	68%
Rise	14	53%	47%
Rise-Fall	9	40%	60%
Same	51	56%	44%

Table 4.2: Outcome percentages of chunks by tone and intonation. We excluded all chunks in which the wake word was pronounced differently, omitted (note, in some chunks omitted wake word interactions were still successful as they were follow-up interactions), or said after the command. This table displays the resulting 88 chunks (48 successful ones). Note, percentages are not exact portions of the total counts as group sizes were adjusted to calculate them.

Rhythm. We found that the voice assistant often did not pay attention to a participant's speaking rhythm. For example, by the eleventh chunk in her interactions with the voice assistant and after having asked for the weather in multiple cities, P15 seemed exhausted, and took a deep breath in the middle of her request. Taking a deep breath slowed down the rhythm of her speech, *"Alexia what's the weather (deep breath)...in Paris?"* As she started saying where (per the suggestion on the screen), Alexa interrupted with the local weather. The voice assistant could detect a user's speech rhythm, and give room for pauses when needed.

Tone and intonation. To better understand how tone and intonation were affecting chunk outcome, we counted their occurrence as shown in Table 4.2. As we can observe in the table, some *tones* tended to result in Successful outcomes (Exaggerated, Friendly, Neutral, Tired, and Upset), and others in Failed or Ambiguous ones (Excited, Indifferent, and Nervous). Similarly some *intonations* tended to result in Successful outcomes (Rise, and Same, or Constant), and others Failed or Ambiguous ones (Fall, and Rise-Fall). Though this analysis is preliminary, it suggests that tone and intonation may give us more context about interactions. Taking these factors into consideration could also help voice assistants recognize errors and subsequently perform self-repair.

As can be seen in Table 4.2, Friendly tones were more likely to be missed by the voice assistant than Exaggerated or Upset tones. However, our participants were very hesitant to speak to it in an impolite manner. Because of this trend, when participants had multiple failed interactions attempts, we encouraged them to speak more sternly. Alexa often did not respond to their soft, friendly tones. If failures continued, we suggested that participants imagine they were upset at Alexa, and speak in an upset tone. Often, once they started to speak to Alexa as though they were angry, Alexa finally responded (see Figure 4.3). When we suggested P20 speak to Alexa as though she was reprimanding Alexa, she responded, as she nervously prepared to try to speak more strongly, *"yo no hablo tan duro,"* which is Spanish for, *"I don't speak so strongly."* After four failed attempts, she asked to stop the activity without ever "waking" the voice assistant.

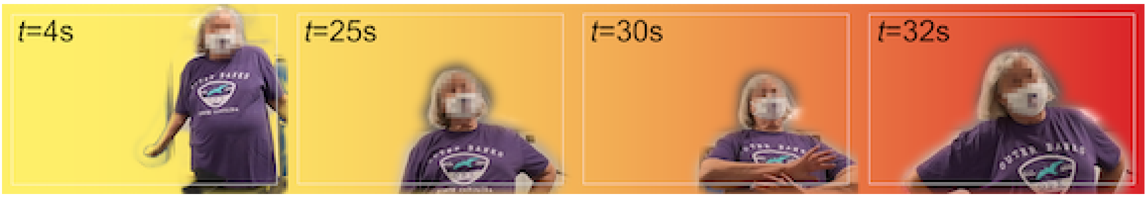

Figure 4.3: Four stills from a participant's video interacting with Alexa. The participant's tone is initially friendly and changes to upset throughout the interaction, indicated by the background color change from yellow to red. Alexa is unresponsive during the participant's "friendly" attempts, and responds when the participant uses an upset tone.

4.5 Discussion

Our findings show how older adults who are novice users may interact with voice assistants in public settings. As such, our work contributes to a small, but growing, body of research that examines human-voice assistant interactions in the field [288, 289, 293]. Our inclusive design approach [104] may help guide future research on voice assistants that are more suitable for older adults and, as a result, for other users as well. Currently, most interactions with smart speaker-based voice assistants happen in the home, but in the near future, voice assistant technology will likely be pervasive in a variety of public contexts (perhaps airline check-in counters, medical facilities, or shopping centers) [352, 323]. Moreover, if purchasing a voice assistant from an electronics store, customers are likely going to interact with them, in a public setting, before deciding if they will buy the device. If the issues our findings surface are not addressed, we may be making voice assistants, and the promises they present, unapproachable to a large and important segment of the global population, hindering adoption, and creating systematic exclusion as voice assistants permeate public spaces.

We divide our discussion into design and research implications. In the design implications section, we 1) provide recommendations addressing interaction errors that resulted from not being able to successfully wake the voice assistant, 2) suggest ways in which automatic detection of non-verbal cues can be used to improve interactions with voice assistants, 3) emphasize differences and complexities for adapting voice assistants' interactions to older adults' needs and abilities in the context of prior research about code switching and knowing the user [62, 103], and 4) close by raising ethical design considerations. In the research implications section, we surface questions surrounding how we might use recent technological advancements to recognize body language and audio-prosodic features, and discuss the societal implications surrounding surveillance tradeoffs.

4.5.1 Design implications

Our findings have important implications for the design of more intuitive multimodal, speech-first interfaces for older adults. Voice assistant design could rely on more familiar interaction paradigms, and/or responsibly capture and analyze data from multiple inputs, to create more natural conversations. In this section, we discuss recommendations for improving voice assistants for older adults and raise concerns regarding doing so ethically.

Improving interactions surrounding waking the voice assistant. Although we gave participants clear instructions on how to initiate interactions, waking the voice assistant was one of the biggest interaction problems we observed. Though we focused on older adults, this finding may also provide some context surrounding the large number of voice assistant interactions that were not successful, or were wake-word only commands, that Ammari et al. [48] identified in the usage logs of younger participants (18–64 years old). This said, Lee et al.'s findings suggest people's first words in an interaction with a robotic agent can predict their schematic orientation to an agent, making it possible to design agents that adapt to individuals during interaction [204]. Conversational errors that prevent interactions from occurring in the first place can thus hinder human-computer cooperation. This issue could be addressed in several ways:

- **New mechanisms to indicate when the voice assistant is not actively listening:** The interaction design of voice assistants could make it more clear to older adults when it is not listening, as our participants overlooked the blue line indicator; for example, by completely shutting off the screen, or having an avatar that looks away. Along the same lines, more consistency could be enforced for "waking" mechanisms, so that the design does not confuse users by sometimes requiring the wake word and sometimes not requiring it (e.g., during multi-turn interactions, such as Trivia).

- **Relying on familiar interaction paradigms:** Other mechanisms to wake the voice assistant could be put in place, such as using physical form-factors. Form factors that could be explored in future studies could be using a (possibly wearable) button that, when touched, would wake the voice assistant. Alternatively, picking up a telephone to talk to the voice assistant might provide a more familiar way to activate the voice assistant.
- **Responding to body language:** Voice assistants could "wake" when a person lean towards them, or showed other signs of engagement. Someone calling a voice assistant's attention by making a sound or motion while looking at it could also wake the device.

Providing friendlier explanations for people who are less familiar with technology. Participants who did not understand how the voice assistant worked did not understand that it was "learning their voice" from making them repeat phrases. More explanations could be added for people that are unfamiliar with this technology. For some, using this voice assistant is a big technological leap, and having it use terms such as "the cloud" and the "Alexa app" without offering additional explanation could be off-putting. Integrating these explanations into the design of the voice assistants could help older adults use voice assistants without the need for additional training from others. These design considerations could help increase digital empowerment for older adults.

Relying on automatic detection of visual and audio-prosodic cues. Voice assistants could be designed to appropriately react to visual and audio-prosodic cues, gaining social intelligence. Some of this is already happening [24]. Ideas for how this might take shape include:

- **Mirroring and understanding the user:** Voice assistants could try to mirror certain characteristics in their users, such as the speed at which they are speaking, to adapt to a user's needs and abilities. Similarly, echoing Nass's research, they could mirror a person's mood [259] or tone, which could increase user satisfactions with the voice assistant. Additionally, voice assistants should be able to recognize different intonations, and use that information to respond appropriately (e.g., if the intonation conveys uncertainty, the voice assistant could reassure the user.)
- **Designing gaze intentionally:** Voice assistants with a movable screens[6] could be designed to make "eye contact" when addressed, or to turn their screens away when they are not paying attention. In prior work, McMillan et al. built a robot, Tama, that could detect the gaze of a user (instead of a wake word), and respond by moving an articulated "head" to achieve mutual gaze [238]. They found that gaze is a promising mechanism for augmenting or even replacing, the wake-work in initiating interaction with voice assistants [238].
- **Improving communication between different modalities:** Voice assistants

[6] https://www.amazon.com/echo-show-10/dp/B07VHZ41L8

could detect interactions with their different modalities, such as when users read the content being displayed on their screens, and consider that content in their responses. For example, users should be able to request more information about a recipe being displayed, such as by saying, "show me more details about the macaroni and cheese recipe you're displaying."

- **Differentiating between single-user and multi-user interactions:** Voice assistants could detect when there is more than one person involved in an interaction (e.g., by using voice recognition or computer vision), and adjust their reactions accordingly. For example, voice assistants could address users individually when needed, and react (or not) to users' interactions with each other when appropriate.

Adapting to users' needs and abilities. We also found that some participants said many words to the voice assistant during a single interaction, sometimes speaking for longer than the voice assistant could process. For example, P22, as described in our findings, told the assistant information about how he was doing, what he cared about, his views surrounding wearing masks, and more in only one conversational turn. In prior research, which did not include older adults, Beneteau et al. found that Alexa did not code switch with people of different ages to adapt its dialogue to the needs and abilities of the people it was interacting with [62]. They found that younger children tended to struggle more than older children and adults under 56, and provided an example of parents noticing their four-year-old would omit the wake word and get frustrated when the voice assistant did not respond back; the child would also use long sentences and often change topic before Alexa responded [62]. Beneteau et al.'s findings are echoed in the interactions we observed with older adults, where many older adults, similar to young children, addressed Alexa in the same way they would address a person. In response, voice assistants should be prepared to listen for longer to users who use more words in each conversational turn.

In this work, we find more evidence to support Beneteau et al.'s claim that "knowledge of the context and the communication partner is extremely helpful, allowing digital home assistants to artificially code switch as needed," and Clark et al.'s assertion that "there may be specific application areas where conversation may be appropriate if not essential between humans and agents, particularly in areas such as healthcare and wellbeing, where the nuances of contexts and demographics need to be considered" [103]. We contribute findings specific to the older adult demographic, which was not included in Beneteau et al.'s nor Clark et al.'s study [62]. Determining how to craft voice assistant dialogue for older adults would require further research, as it is unlikely that there will be a one-size-fit-all solution [119].

Ethical design. Despite their close resemblance to human voices, voice assistants are mere machines with many social deficits, making them unable to meet the expectations that they set. Our work highlights possible modifications to voice assistants—such as relying on visual cues to determine responses—that have the potential to make interactions more human-like. However, this comes with great responsibility. Human-likeness may affect our expectations of voice assistants [260], potentially increasing undue trust placed in them and encouraging stronger emotional connections. Design choices, such as using a robotic-sounding voice, may more accurately portray a voice assistant's true nature and prevent undue trust from being placed on it.

Feasibility. Our recommendations complement recent technological advances and work in progress. For example, in 2018, Kepuska and Bohouta [189] proposed developing a multi-input voice assistant that is able to interpret speech, video, images, and gestures from users. The system they proposed relies on piecing together various existing technologies, such as Kinects, cameras, APIs, and machine learning models [189]. More recently, Brunete et al. [80] developed a prototype for a robotic system to control a room that also relies on multiple inputs, including gaze, body language, and voice. Moreover, Nie et al. [262] recently developed a scheme to wake voice assistants without the need for a wake word by relying on other visual and audio cues. Large technology companies are also exploring how to include multi-channel inputs to improve human-machine conversations. For example, Amazon is using acoustic, linguistic, and visual cues to help Alexa interact more naturally [24]. Taken together, these advancements may make it possible to develop improved software agents.

4.5.2 Implications for voice assistant research

Through this research, we find that the most widely used research methodologies in the academic literature for studying older adult interactions with voice assistants (usage logs and interviews) are partial and incomplete, as many older adults are not able to even activate their voice assistants with consistent success. Because of this, analyses of usage logs collected in the privacy of the home may miss a large portion of failed interaction attempts. We therefore call for more research entailing interpreting body language and audio-prosodic features while honoring privacy expectations.

Interpreting body language. Video analysis revealed visual information that could be used to improve interactions. For example, posture shifts, such as leaning forward to be closer to the voice assistant, could be used for waking the voice assistant. In our research, posture shifts were an important component for indicating engagement. Additionally, following a participant's gaze was essential to understanding when a participant was reading or looking at something displayed on the voice assistant's screen, was distracted by something else occurring around them, or was requesting assistance. Given the advanced state-of-the-art of computer vision and sensing technologies [132, 141], it is important to study how these technological advancements may be used to recognize and interpret body language automatically in interactions with voice assistants.

Interpreting audio-prosodic features. Voice assistants tend to have human-sounding voices, and can be programmed to have prosodic variations. Alexa's friendly tone is a human-like conversational quality, which signals that it would be able to respond appropriately when spoken to as a human. However, in our research we noticed that Alexa's friendliness was a deceptive characteristic, at times, as Alexa had more difficulty responding to participants when they spoke to it in a friendly tone than when they approached it with an upset tone. Even though it was projecting friendliness, it did not understand friendliness when participants displayed it, resulting in inappropriate responses (or lack thereof). This calls for more research to interpret participant's audio-prosodic features, such as by using Amazon's Halo band that can measure tone of voice [3], to further understand how prosidic variations correlate to the voice assistant's responses.

Privacy considerations. Voice assistants that can "see" into our homes are already entering the market [30], but their societal implications are understudied. This is concerning because they could strengthen and continue to normalize technological surveillance [388]. It is important for interaction elements that could threaten our privacy to be considered and critiqued, as capturing and interpreting visual and audio-prosodic information requires potentially invasive data collection that comes with privacy and surveillance risks. Ensuring that computations happen on-device could be one way to limit the amount of data collected and stored. However, even then, having autonomous speakers with a camera consistently able to observe us could normalize surveillance by device and platform providers, as well as businesses, employers or remote family members. Concerns surrounding video surveillance of older adults are already being raised in the literature, and should be considered when adding mechanisms that could increase the risk of privacy violations [66, 96]. In addition, Bonilla and Martin-Hammond [70] found that knowledge of voice assistant privacy practices, data use and management are key concerns for older adults, and that many of their participants were unaware of existing resources available to mitigate such concerns. Future work is needed to explore not only the privacy and ethical implications of potentially intrusive technology, but also how vulnerable users may perceive and be affected by them.

4.5.3 Limitations

Our study has several limitations: it is a small scale, qualitative study conducted in an urban setting in the U.S. Moreover, we used a smart speaker-based voice assistant with a screen, so we do not know if our findings generalize to other voice assistants, such as screenless ones. Future research could investigate how these findings translate to voice assistants embodied in different devices. In addition, most participants were novice users of voice assistants, which may limit the generalizability of our findings. However, understanding novice user's interactions and struggles is necessary to promote adoption and prevent systematic exclusion of certain populations. Future research could explore interactions of users with varying levels of expertise and from different population segments. Participation was also limited to those who chose to participate; those that chose not to participate may have additional reasons for why they chose not to interact that our study did not surface. Another exciting area of future research would be to conduct video analyses of older adult interactions with voice assistants in different geographic locations and settings.

4.6 Conclusion

We used video analysis to characterize challenges with voice assistants' current design that may hinder older adults from benefiting from the promises the technology holds, or worse, exclude them from everyday activities as these technologies permeate public spaces. We described human-machine communication gaps revealed by our data, differentiating information that was overlooked by the machine (e.g., interaction attempts, the presence of more than one user) from information that was overlooked by participants (e.g., the blue line indicator, and technical terminology). We then examined body language features of the interactions and categorized them into those that provide reliable signals (e.g., leaning forward and gaze), and those that are somewhat ambiguous (e.g., laughing). Relatedly, we found that audio-prosodic features could also generate important information for reducing human-machine communication gaps, such as by identifying pauses from breathing or different tones and intonations. We discussed design implications for more intuitive interfaces for older adults, and conclude with a call for more research entailing responsibly capturing and analyzing data from multiple inputs to create more natural conversations. Taken together, our findings help improve the inclusion of older adults in the design of voice assistants.

CHAPTER 5

INCREASING HUMANLIKENESS THROUGH SELF-REPAIR IMPROVES VOICE ASSISTANT INTERACTION

The prior two chapters surfaced a multitude of communication breakdowns. In this chapter, I explore building voice assistants that perform self-repair as a way to address communication breakdowns. One key technique people use in conversation and collaboration is conversational repair. Repair is the conversational analysis term for when interactants try to fix problems in speaking, hearing or understanding that come up during conversation [282, 100, 315, 309]. Self-repair is repair by the speaker of that which is being repaired instead of repair by another interactant [319]. I investigate how the self-repair of errors by voice assistants affects user interaction. I explore this through a different population (university students in their late teens or twenties), because doing so with older adults, who may be more vulnerable, is fraught at this early stage of experimentation and development. Future work should investigate whether self-repair also improves older adult interactions with voice assistants.

Note, the content in this chapter was published in April of 2021 and presented in October of 2021 at the 24th ACM Conference on Computer-Supported Cooperative Work And Social Computing (CSCW) [115]. It was written in collaboration with Shuran Li, Hansol Lee, Jason Cho, and Wendy Ju. Because of this, I will be switching from the singular “I” to the plural “we” for the rest of the chapter. The writing has been slightly edited to fit this dissertation.

In a controlled human-participant study (N=101), participants asked Amazon Alexa to perform four tasks, and we manipulated whether Alexa would "make a mistake" understanding the participant (for example, playing heavy metal in response to a request for relaxing music) and whether Alexa would perform a correction (for example, stating, "You don't seem pleased. Did I get that wrong?") We measured the impact of self-repair on the participant's perception of the interaction in four conditions: correction *(mistakes made and repair performed)*, undercorrection *(mistakes made, no repair performed)*, overcorrection *(no mistakes made, but repair performed)*, and control *(no mistakes made, and no repair performed)*. Subsequently, we conducted free-response interviews with each participant about their interactions. This study finds that self-repair greatly improves people's assessment of an voice assistant if a mistake has been made, but can degrade assessment if no correction is needed. However, we find that the positive impact of self-repair in the wake of an error outweighs the negative impact of overcorrection. In addition, participants who recently experienced an error saw increased value in self-repair as a feature, regardless of whether they experienced a repair themselves.

5.1 Introduction

A worker in a meeting notices his boss's computer is low on power and asks her, "Do you want my charger?"

"I'm good."

The worker starts reaching into his bag to look for his charger, but stops after seeing the surprised expression on his boss's face.

"Oh, you don't want my charger."

"No."

"No problem!"

In this vignette, the term "I'm good," which means, "don't worry about it," was misunderstood to be a positive response to the worker'swaiter's offer, but the workerwaiter repairs the misunderstanding after first responding incorrectly when he sees his boss'ss/he sees the patron's unexpected reaction. In human face-to-face interaction, people monitor each other continuously to see if they are understanding and being understood by others, and they stop and self-correct if they recognize that they have made an error. This capacity for "self-repair" helps to ease the irritation and friction that comes from having to explicitly correct mistakes or misinterpretations, or from suffering the consequences of uncorrected miscommunications. Self-repair in human interaction with machines, such as interactive voice assistants or robots, is not yet common, but is an area of interest; computer scientists and roboticists are working on applying machine learning to people's verbal and non-verbal behaviors to catch communication errors so that self-repair can occur [311, 71]. Self-repair may be crucial for making conversational agents useful in a computer-supported cooperative work environment. However, it is not yet known how confident a machine should be that it has com-

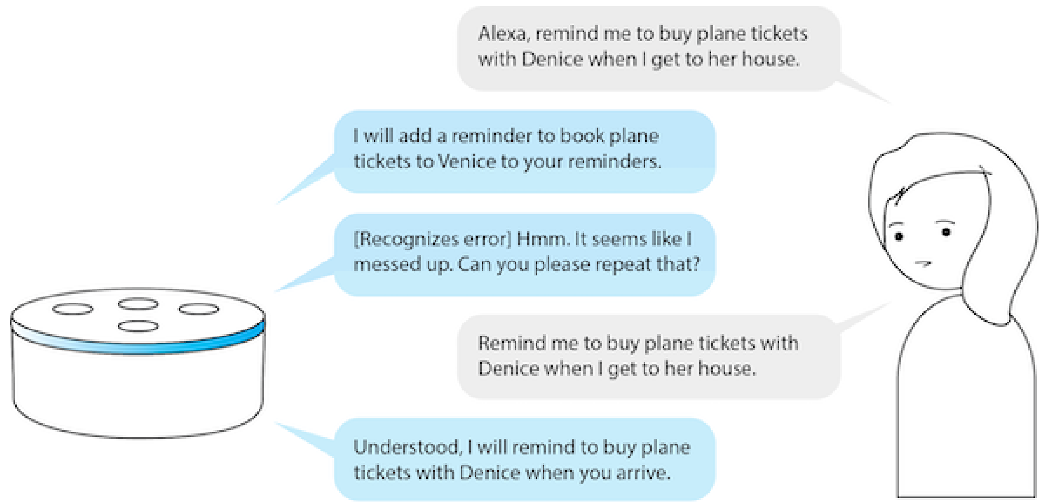

Figure 5.1: Example of a scenario in our study in which the voice assistant (Alexa) successfully recognizes and repairs an error.

mitted an error before it attempts to perform self-repair. We might expect that a machine that can correct its own mistakes is perceived to be more capable than one that cannot; but what is the benefit, in terms of user assessment, of correcting mistakes when they are made, compared to not correcting at all? And what is the cost of a machine performing self-correction when it has *not* made a mistake? All these factors need to be weighed in the machine's decision analysis of whether to attempt self-repair actions.

In this work, we investigate the effects of voice assistants performing self-repair in the presence and in the absence of mistakes (see Figure 5.1). We measure the impact of self-repair on the participant, and on the participant's perception of the voice assistant in four conditions (see Figure 5.2): control (no mistakes made, and no repair performed); undercorrection (mistakes made, but no repair performed); overcorrection (no mistakes made, but repair performed); and correction (mistakes made, and repair performed). We then measure the desirability of self-repair. We further provide qualitative findings based on what participants said about their interactions, and identify the factors that help us determine the main trade-offs between the different conditions. To the best of our knowledge, this is the first study that has been conducted to measure the effects of voice assistants performing self-repair. These findings, as a whole, can inform design guidelines for using self-repair as an interaction mechanic in human-machine interaction.

5.1.1 Related Work

Prior work on conversational repair spans a large array of disciplines, from linguistics to human robot interaction. In this related work section, we consolidate disparate threads of research in these different communities to form a more coherent picture of the prior work on self-repair in interaction.

Conversational agents in cooperative work

By studying how the design of conversational agents affects human behavior and perception, the CSCW community can realize the full potential of such agents in cooperative work. For example, Williams et al. studied the use of a chatbot to help participants with work detachment and reattachment processes, and found that productivity gains were better sustained when the conversations used emotionscentric prompts *"how do you feel"* instead of task-centric ones *"what did you do"* [374]. Xiao et al. evaluated responses to open-ended questions when administered via a chatbot and via an online survey, and found that the chatbot outperformed the online survey in driving higher level of participant engagement and eliciting significantly better quality responses [375]. When studying a robot's potential to shape trust within a collaborative environment with robots and humans, Strohkorb Sebo et al. found that robots that express vulnerability can have "ripple effects" on their human team members' expressions of trust-related behavior, making human teammates more likely to explain their failure to the group, console team members who had made mistakes, and laugh together [337]. Similarly, Traeger et al. found that people in groups with a robot making vulnerable statements converse substantially more with each other, distribute their conversation somewhat more equally, and perceive their groups more positively compared to control groups with a robot that either makes neutral statements or no statements [347]. However, social signaling behavior does not come without trade-offs. For example, Jung et al. found that even though robots that used backchanneling improved team functioning, the agents themselves were perceived as less effective than those that did not [187]. Additionally, Ashktorab et al., who explored different repair strategies in chatbots, discuss repair as a collaborative action with costs (e.g., too much turn-taking or loss of naturalness), calling for further empirical research in the area [54].

The future of voice-based devices

Unlike some of the first-generation voice agent research published at CHI, such as [263, 78, 341] where the computer-generated voice that people were speaking with was disembodied, or the embodied conversational agents of Cassell, Sullivan, Churchill and Prevost [86], which were front-ended by on-screen virtual agents, today's voice assistants are embodied in standalone devices such as Amazon's Echo [47], Apple's Homepod [53], or Google's Home [157].

While voice-enabled robots like the Jibo social robot [143] or Anki's Vector [52] have had limited commercial success to date, it still seems likely that future robots and appliances will feature speech interaction as a feature. Part of this trend is driven by advances in natural language processing, text-to-speech and dialog generation systems driven by big data, as well as hardware breakthroughs in far-field microphone arrays. While improvements to the hardware and software of these speech-enabled devices might improve the recognition of individual words people say, common-sense intelligence is not yet in grasp [133]. As we observed in the last two chapters, the limited capabilities of today's voice assisted devices contributes to the exclusion of older adults. Improving interaction savvy by building systems that recognize and recover from conversational errors could also improve inclusion.

Conversational repair

Voice assistants may make mistakes, but human dialog is far from error-free itself. A key difference is that people perform repair in communication [309], monitoring listeners to see if they have been heard and understood before moving forward in the conversation.

Schegloff, Jefferson and Sacks define repair to be the practices that interactants use to handle troubles in hearing, speaking and understanding that occur regularly in social interaction. They noted that a repair sequence has several key segments: the repairable, repair initiation, and a repair outcome. This formulation took into account that sometime repairs were initiated where no error occurred–even a correct statement could be repairable. The repair initiation could come from the speaker (self-initiation) or the listener (other-initiation). The outcome–what was suggested in place of the repairable–could similarly be correct or incorrect, and accomplished by the self or other [309]. This team later found from analyzing naturalistic conversation that people had a preference for self-correction over being corrected by others: in moments when repair was necessary or possible, the distribution of repairs was strongly skewed towards self-repair [319]. Very often self-repair occurs when the speaker notices a mistake, in the transition space between speaking turns, before the listener even has a chance to respond, but they noted that when the listener initiated a repair, the original speaker usually responded by self-repairing before the other performed a correction. These patterns are also noted by Moore and Arar in the introduction to their book surrounding conversational user experience design [249].

Whereas Schegloff and his colleagues focus on linguistic repair [316, 317, 318], Clark and Schaefer's contribution model addressed the detection and repair of communication model through a more regulatory model. In this model, conversation contains contributions with a presentation and a subsequent acceptance. In other words, the speaker is actively seeking evidence that they are being understood, and is as likely to initiate repair when evidence of understanding is insufficient as when they have firm evidence that they were not understood [101].

Often, the evidence of understanding is not verbal. Ekman and Friesen have drawn attention to the nonverbal acts people perform to maintain and regulate the back-and-forth nature of speaking and listening [135]. *Regulator* actions, which indicate that people are listening, understand, or if they are confused, take exception, or want to respond, occur in the attentional periphery; people perform them without thought, but can recall and repeat them if asked. We observed many regulator actions at play in older adults' interactions with voice assistants, such as leaning forward, in the previous chapter. In collaborative conversations [101], addressees must therefore also indicate their understanding, or lack of understanding, to help the speaker understand the state of the communication. Chovil's experiment with people listening to a story in a face-to-face, partition, and telephone and answering machine condition showed that listeners primarily react facially when they would be seen by the storyteller [95]. Hence, the monitoring for acceptance and understanding of speech in face to face interaction is often multi-modal

[77].

Error-recognition

The advancement of error-recognition technology is an imperative part of improving human-computer interaction via self-repair. The importance of identifying and incorporating responses to conversational signals was recognized early in the human-computer interaction community by Nagao and Takeuchi [255]. Elements such as empathy and the emotions associated with certain utterances have also been studied and play an important role in error recognition [277, 220]. Bousmalis et al. have surveyed the conversation analysis literature for nonverbal audiovisual cues that indicate agreement and disagreement between human speakers, with the goal of developing machine recognition of these cues [71]. More recently, Salazar-Gomez et al. experimented with using EEG-based feedback methods to correct robot mistakes in real time; because the EEG signals were analyzed in realtime in closed-loop fashion, the robot was able to respond to possible signs of error by hyper-articulating actions to elicit stronger response to help it determine if it was making a mistake [311]. This current research is premised on the capability of error recognition to occur, whether through physiological measures, visual or audio recognition or through discourse analysis. However, in this chapter's study, because of the need to control participant experiences by condition, error recognition is simulated rather than actually performed.

There are several varieties of cues that can aid in performing error-recognition, some of which we observed in the previous chapter:

Discourse cues: Gieselmann ran a small experiment to look at what error recovery strategies people use when talking to robots compared to when they talk to other people. Geiselmann found that achievement strategies (such as paraphrasing, repeating, or restructuring) and functional reduction strategies (such as giving pre-selected answers or changing the theme) were used, largely due to the limited interaction capabilities of robot, and that the most common indicator that an error was made is a sudden change in the dialogue topic. In this research, the focus of the error detection lay in analysis of the discourse [152].

Audio cues: There are signals in human speech that can be used to recognize error. Oviatt et al. found, for example, that people tend to hyperarticulate when talking to machines, often making it harder for the machine to recognize what the person is trying to say [276]. They proposed a two-stage Computer-elicited Hyperarticulate Adaptation Model to account for this repair mechanism that people use. Levow analyzed acoustic-prosodic features, like duration of speaking, pauses, and changes in volume and pitch, to predict when people were responding to machine misunderstandings [208]. Litman et al. used the machine learning program RIPPER to produce a classification model that improved the prediction of misrecognitions using these types of acoustic-prosodic features on the TOOT corpus, a spoken dialogue system for accessing train schedules via telephone [213].

Visual cues: The improvement in error-recognition technology via visual cues can be foreshadowed by the widespread availability of emotional expression image databases such as Ekman's Pictures of Facial Affect [134], the Belfast database [129], the Extended Cohn-Kanade Dataset [222], or the Affectiva-MIT Facial Expression Dataset [237]. Because error-recognition and self-repair often go hand in hand, it is crucial that we also research and understand the scope of possibilities and trade-offs of repair as a function of a computationally determined decisions, such as whether repair is needed or not.

Based on this active research in the space of error recognition, we believe the possibility of self-repair is very much on the horizon. However, the mere recognition of error does not actually indicate when and how repair should occur.

Errors and repair in social interactions with conversational interfaces

Brennan points out that conversation is shaped by visual and spoken evidence [77], but much of the early research done in conversational interfaces was done for phone interaction, where only auditory evidence is available.

Repairs in spoken dialogue interfaces are often subroutines that are called when the voice agent does not hear a response to a question, or when the response does not fit anything in its limited response vocabulary [239]. Dan Bohus classifies these as non-understandings (when the system does not acquire useful information from the user's turn) and mis-understandings (when the information gather by the system from the user is incorrect) [69].

"I'm sorry, I didn't hear that, let's try again..." is a refrain many of us have heard in phone interfaces. These responses are usually generated by dialogue management systems that repair the interaction breakdowns that occur when the system fails to understand the person [383]. These systems do not repair breakdowns that occur when the person fails to understand the system, except that people frequently respond to such situations by not speaking at all. Repair routines that re-iterate or re-word the original query can help get the interaction back on track, but can still be problematic if the line of inquiry or dialogue is incorrect. Rudzicz et al. found, for example, that older individuals with Alzheimer's Disease are often confused by speech interaction, and respond 40% of the time by not responding at all [307].

Corti and Gillespie found that people are less likely to initiate repairs with agents that are disembodied or that are not represented by human [106]. They posit that this is due to intersubjectivity, which requires each party to think the other party knows what their point of view is [154]. They argue that when a person does not see an anthropomorphic agent, the person does not initiate repairs because that person does not perceive the agent can observe or understand their repair initiation activities [154]. In another study, Candello and Pinhanez explored the use of multiple chatbots, each bot having expertise in a specific area, to repair dialogue failures (for example, by readdressing a question), and found that multiple agents expands the opportunities and strategies for handling errors [85]. Such strategies can also help set the norms, by showing that each bot does not know what the other bots point of view is.

Errors and repair in social interactions with robots

The human-robot interaction community is perhaps highly motivated to understand how to perform repair, because current-day robots fall so far short of executing understanding and physical tasks correctly. Building off of Reeves and Nass' "Computers as Social Actors" hypothesis [258], the human-robot interaction community hopes that sophistication in social interaction can help to compensate for short-comings elsewhere.

We might assume that people would prefer robots that behaved perfectly and never made mistakes, but several experimental studies indicate that this is not true. Ragni, et al. found that people collaborating with robots on a memorization task were more likely to report positive emotions with an erroneous robot than the perfect one [296]. Mirnig, et al. studied people who got instructions from a Nao robot on a Lego building tasks; participants liked the faulty robot significantly better than the robot that interacted flawlessly, even though the faulty robot degraded their own performance [246]. Salem, et al. looked at how errors in communication in particular, with a robot that did not gesture while speaking, gestured congruently while speaking, or gestured incongruently while speaking, and found surprisingly that participants liked the robot that gestured incongruently the most [312].

The human-robot interaction community has also spent a lot of time looking at recognizing human social signals in interaction. Breazeal and Rani for instance, provide good recaps of the work in the HRI community on affect recognition [76, 297]. The HRI community has also focused on recognizing the embodied signals of human interactants for conversational *regulation* Fujie et al. made a robot that recognized head motions, like nodding, for paralinguistic information that clarifies speaker intent [148]. Sidner et al. found that participants who knew their robots recognized conversational head nods would nod more [329]. Huang and Mutlu have proposed developing a Robot Behavior Toolkit that uses many of the same social cues that people use to achieve interaction goals in order to make robots that are able to adapt their behaviors to people [172]. Mutlu et al. found that people remembered stories that story-telling robots told better if the robot looked at them more, that they could get listeners to behave as addressees or bystanders by having the robots look at them as if they were addressees or bystanders, that they could encourage turn-taking by having the robot change who it looked at [252, 253, 254, 51, 173, 281]. These studies, considered broadly, indicate that expectations for interaction and communication–down to the timing and the gaze patterns–persist when people are speaking to a robot or machine.

From this research, we believe that self-repair in response to non-verbal cues would be appreciated by people and make them rate their interactions and their fellow interactant more highly than those where mistakes were allowed to persist without repair by the acting agent.

5.1.2 Research Questions

Although much research has been conducted to understand how humans respond to error, there is little analogous research to understand how people respond to self-correction by computer- or machine-based agents. Given that self-repair may mitigate the downsides of mistakes [258, 319, 84, 101], that repair is a normal part of human conversation [135, 309, 319, 316, 317, 95, 318, 84, 101], and that people tend to prefer robots that are not perfect [296, 246, 312], we hypothesize that:

Hypothesis 1. Participants will rate an agent that successfully repairs its mistakes *(correction)* better than an agent that makes no mistakes *(control)*, or repairs nonexistent mistakes *(overcorrection)*, or makes mistakes and does not repair them *(undercorrection).*

Hypothesis 2. The desirability for self-repair capability will be higher for participants in the conditions where mistakes are made *(correction and undercorrection)*.

We base the second hypothesis on the idea that participants in these conditions would have more proximal experience with the frustration and degraded performance associated with mistakes.

5.2 Method

A laboratory between-participants experiment was conducted, using a 2 (presence of mistake: no mistakes made vs. mistakes made)×2 (presence of repair: no repair performed vs. repair performed). We chose a between subjects experimental design to minimize learning effects across conditions, increase the number of total participants by having shorter sessions, and be able to more effectively determine which factors played a bigger role in participants' perceptions. The conditions are labeled for readability as depicted in Figure 5.2: control, undercorrection, overcorrection, and correction. Participants are semi-randomly assigned to each condition: control (N=22), undercorrection (N=30), overcorrection (N=30), and correction (N=19). Upon entering the room, all participants are given the same research scenario: they're about to start driving to a friend's house, they are going to use the Amazon Alexa to help them accomplish a few tasks, and to imagine that Alexa has the ability to see their reactions through Camera 2 (see Figure 5.3). There are four main tasks per condition: 1) "Can you take me to Denice Johnson's house?", 2) "Send a message to Denice saying, 'I will be there in ten minutes."', 3) "Remind me to buy plane tickets with Denice when I get to her house.", and 4) "Play relaxing music." The full guiding scripts are available in the Appendix (Section .0.4). Subsequently, participants are asked to interact with the Amazon Alexa by reading prompts displayed on a screen, including clarifications when Alexa makes unnecessary repairs or mistakes, see Figure 5.2. For the purposes of the study, we refer to error-recognition as the act of identifying that

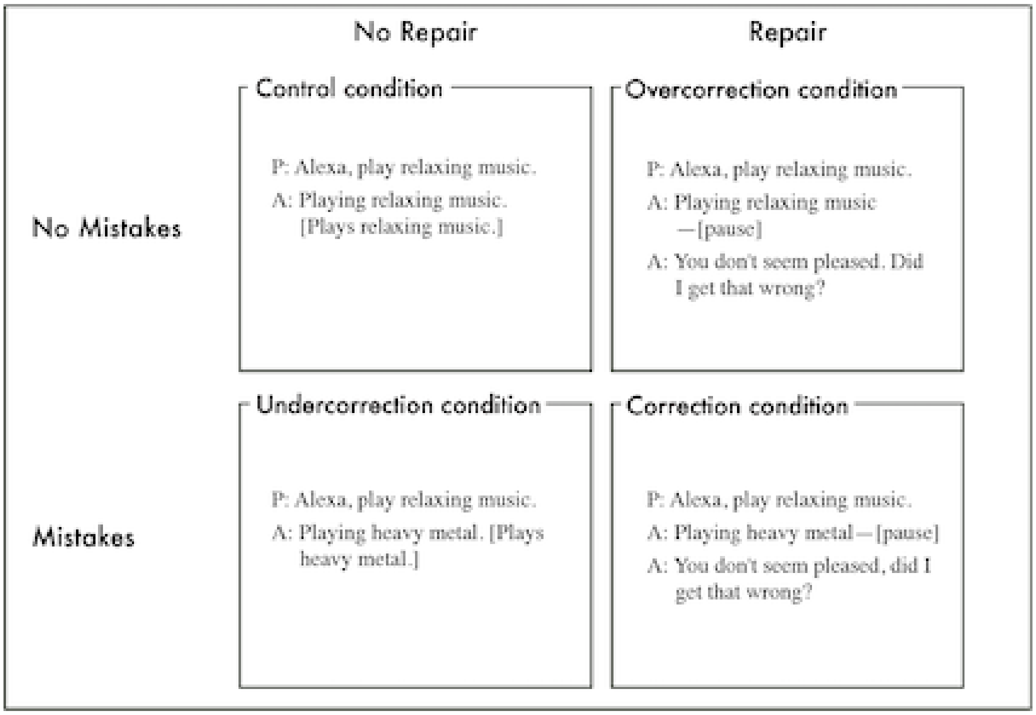

Figure 5.2: Examples of interactions in each condition in the 2 (presence of mistake: no mistakes made vs. mistakes made) × 2 (presence of repair: no repair performed vs. repair performed) matrix.

a mistake has been made. We define mistakes as errors made by the voice agent in interpreting and responding to a verbal prompt; for example, hearing "salmon" instead of "seven." In this chapter, we are not categorizing unnecessary self-repair as a mistake. Self-repair is used to refer to the voice agent taking some action to correct a mistake, such as asking what it did wrong and adjusting the response accordingly.

The physical set up for the study is shown in Figure 5.3. Two cameras are placed in front of participants to record their interactions with the Amazon Alexa

Echo Dot, one from the front angle, and one from the perspective of the Amazon Alexa Echo Dot. A researcher sits next to the participants to flip through the prompts. Captured video is saved for a future dataset, which is not a part of this publication. After completing their interaction, participants are asked to fill out a questionnaire, and participate in a short, semi-structured interview.

5.2.1 Participants

A total of N=101 university students (F:73, M:28) between the ages of 18 and 30 participated in our study, N=20 participants spoke English as a second language. All participants were considered when evaluating *Hypothesis 1*, regarding which condition is better, and only the N=78 participants that passed the manipulation checks (which measured whether they agreed with our definitions of mistakes and repair) were considered when evaluating *Hypothesis 2*, relating to the desirability of repair. A high portion of the participants in the undercorrection and overcorrection conditions failed our manipulation checks, so we increased our quotas for participants in those conditions to be able to successfully compare *Hypothesis 2* results.

5.2.2 System

Amazon Alexa Echo Dot (3rd Gen) device was used with the default Alexa voice (female, American-accent), and the software to operationalize each condition was authored using Jovo 2.2.12 and Node.js 8.10. Video was recorded at 29 frames/second with a resolution of 1920 x 1080 DPI. The interviews were transcribed from the video recordings by a third party service.

Figure 5.3: Image demonstrating the setup of a study session. Participant sits in front of an Alexa device, two cameras, and a computer. A researcher sits next to the participant to flip through the prompts on the computer.

5.2.3 Measures

For the quantitative analysis, the 28 items in the questionnaire were divided into participant feelings and perceptions (the first 22 items, see Section .0.5 in the Appendix), and self-repair desirability (the last 6 items, with the first 2 of those being the manipulation checks, see Section .0.5 in the Appendix). For the qualitative measures, we analyzed open-ended feedback following the initial questionnaire. Below, we describe each of these measures in detail.

Participant Feelings and Perceptions

The post-interaction questionnaire asked participants how strongly they agreed with statements regarding how the Amazon Alexa made participants feel (ex: "This voice agent made me feel successful."), how the participants perceived the Amazon Alexa to be (ex: "This voice agent was likeable."), and participant's perceptions of Amazon Alexa's personality using questions from Gosling's Ten Item Personality Measure (ex: "This voice agent was anxious, easily upset.") [159]. The responses were scored on a seven-point Likert scale from "disagree strongly" to "agree strongly". All questionnaire items are included in the Appendix (Section .0.5).

The two most relevant components were identified using principal component analysis, and later the factors at either end of those components were analyzed and aggregated. The components were labeled based on the values of the items that were most influential to the specific component. The aggregated data was then plotted and visually examined, and analyzed using two one-way between subjects analyses of variance (ANOVAs), and subsequently performing post hoc comparisons using the two-sided Tukey HSD test.

Self-Repair Desirability

To ensure that participants agreed with our definition of mistakes and self-repair, we included two manipulation-check questions: "Did this voice agent ever make a mistake?”, and “Did this voice agent ever try to repair a mistake it made?”. Next, to measure the desirability of the self-repair feature, we asked hypothetical questions about voice assistants that perform correction (ex: “Rate the following statement: a voice agent that always tried to correct itself after a mistake... ...would annoy me. ...would waste my time. ...would help me feel less frustrated. ...would improve the conversation quality.”) Response options ranged from from 1 (disagree strongly) to 10 (agree strongly).

We included the data from all N=101 participants in the *Hypothesis 1* analysis, and only the data from the N=78 participants who passed the manipulation checks in the *Hypothesis 2* analysis. Numerous participants failed our manipulation checks in two conditions: in the undercorrection condition (N=30), 9 participants thought Alexa had tried to repair its mistakes, when in fact there was no correction; and in the overcorrection condition (N=30), 8 participants responded that either Alexa made mistakes and repaired them, and 5 reported Alexa did not make mistakes and did not repair them, when in fact there was no error but there was an attempted correction. Only 1 participant failed the manipulation checks in the other conditions.

Post-interaction Qualitative Reactions

Upon completion of the questionnaire, we interviewed each participant, seeking open-ended feedback in three areas: the overall experience with the agent, reactions to error-recognition, and reactions to self-repair. The questions were phrased differently depending on the participant's assigned experimental condition. For example, for the control condition, we asked, "If Alexa were to make a mistake, what would think about Alexa being able to recognize her own error, maybe from your facial expressions, voice, etc.?" For the correction condition, we asked, "What do you think about Alexa trying to recognize her own error, maybe from your facial expressions, voice, etc.?". This interview allowed participants to directly comment on the voice assistant they interacted with.

5.3 Results

In the following sections, we describe our results in the same order that the questions were presented to participants. First we present findings on participants feelings and perceptions which disprove *Hypothesis 1*. Next we present findings for self-repair desirability, which support *Hypothesis 2*. The statistics are reported at the 99% confidence level. Lastly, we provide findings from our qualitative analysis.

5.3.1 Participant Feelings and Perceptions

We used principal component analysis to determine which out of the 22 factors measured (see first 22 items in Section .0.5 in the Appendix) were having the biggest impact on participants' feelings and perceptions analyze the data. Based on the approximate percent of variation per principal direction, we decided to move forward with the two components accountingthat acounted for the most variance in the data, the “feel successful" component (accounting for 46.2% of the variation) and the “this voice agent was calm component (account for 9.5% of the variation). The rest of the components, taken individually, only accounted for 5% of the variation or less.

For the component accounting for the most variance in the data (the “feel successful" component), the factors most heavily affecting the data in the positive direction were “this voice agent made me feel successful,” factor loading of .34, and “this voice agent made me feel efficient,” factor loading of .34. And in the negative direction was “this voice agent made me feel frustrated,” factor loading of -.28.

For the component accounting for the second most variance in the data (the “this voice agent was calm" component), the factors most heavily affecting the data in the positive direction were “this voice agent was calm, emotionally stable,” factor loading of .34, and “this voice agent was extroverted, enthusiastic,” factor loading of .13. And in the negative direction was “this voice agent was anxious, easily upset,” factor loading of -.70.

To further examine what was going on in each principal component, we took the average of the responses to the three factors most heavily affecting the data per component per participant, reversing the ones in the negative direction.

How successful the voice agent made participants feel

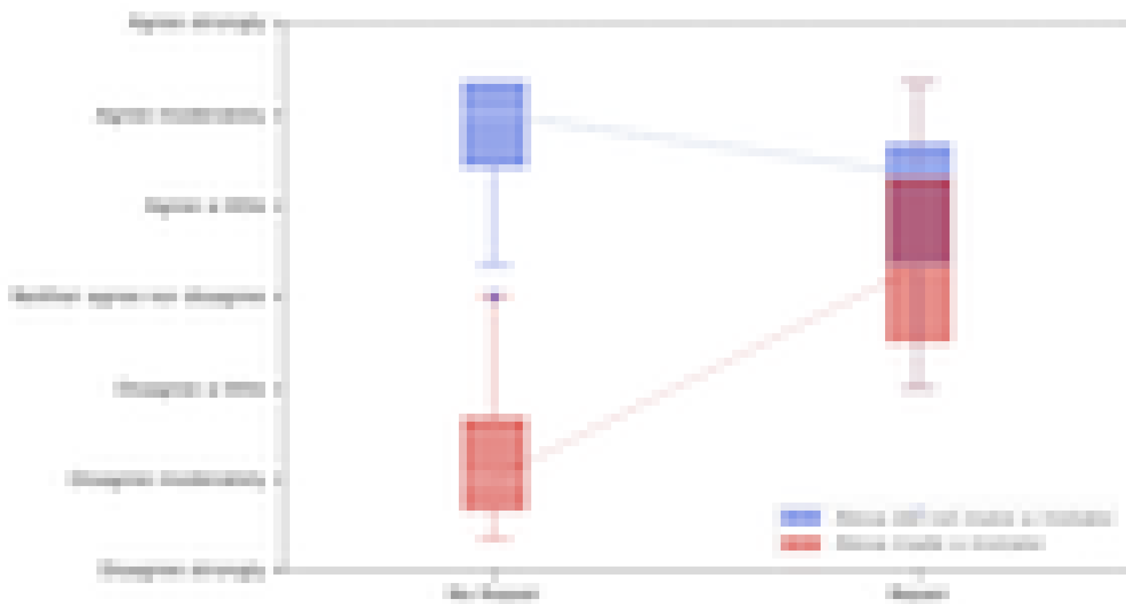

Figure 5.4: Box plot of "feel successful" aggregated data. The data for the conditions without repair is shown on the left, and the data for conditions with repair is shown on the right. The data for the conditions without mistakes is blue, and the data for conditions with mistakes is red. Overall, this box plot depicts an improvement in interaction in the presence of repair. Without repair, the median of the aggregated responses to questions correlated to making the participant feel successful in situations where Alexa made a mistake (undercorrection) was approximately "disagree moderately". With repair, regardless of whether repair was needed (correction) or not needed (overcorrection), neither median was below "neither agree nor disagree".

Figure 5.4 shows the plotted data for the average of the responses to the questions "this voice agent made me feel successful," "this voice agent made me feel efficient," and reversed answers to the question "this voice agent made me feel frustrated."

This figure shows a predictable interaction effect: the repair action lowers the rating of the interaction quality if no mistake was present, but dramatically improves the perceived quality if there was mistake. Without repair, the median of the aggregated responses to questions correlated to making the participant feel successful in situations where Alexa made a mistake (undercorrection) was approximately "disagree moderately". With repair, regardless of whether repair was needed (correction) or not needed (overcorrection), neither median was below "neither agree nor disagree", demonstrating an improvement in interaction in the presence of repair.

A two-way ANOVA was conducted to determine the degree to which the two independent categorical variables, mistake and self-repair, respectively explain the observed variance in how successful participants felt. We found that the mistake variable [$F(1, 97) = 123.91$, $p < .001$] and the self-repair variable [$F(1, 97) = 10.20$, $p = .002$] both had significant effects on how successful participants felt, and that interaction between the two was also significant [$F(1, 97) = 63.60$, $p < .001$]. This indicates that whether or not a mistake occurred, and whether or not self-repair was attempted, both measurably affected participants' feelings of success with Alexa; and moreover, that whether or not a mistake had been made significantly influenced the effect that self-repair (or the lack thereof) had on the feeling of success.

Post hoc pairwise comparisons between the four experimental conditions were then made using the two-sided Tukey HSD test, to determine how the conditions differed from one another with respect to participants feeling successful. A significant positive difference was observed between the control and undercorrection conditions (Hedges's effect size = 3.747, p = .001) indicating that participants in the control condition (no mistake, no repair) group felt much more successful than those in the undercorrection condition (mistake, no repair) group. Significant negative differences were likewise observed between the undercorrection and correction conditions (Hedges's effect size = -2.299, p = .001), as well as between the undercorrection and overcorrection conditions (Hedges's effect size = -2.866, p = .001). Experiencing undercorrection (a mistake without repair) therefore made participants feel notably less successful than either correction (a mistake with subsequent repair) or overcorrection (repair when no mistake had been made). A smaller significant positive effect size was observed between the control and correction conditions (Hedges's effect size = 1.440, p = .001), indicating that participants felt somewhat less successful after correction (when a mistake was repaired), relative to the control (when neither a mistake nor repair had taken place). Comparing the control condition to the overcorrection condition (Hedges's effect size = .887, p = .008), we found that participants felt slightly less successful with overcorrection (when the agent performed unnecessary repair) relative to the control. Interestingly, no significant difference was observed between the overcorrection and correction conditions (Hedges's effect size = .559, p = .21), indicating that unnecessary repair did not make participants feel demonstrably less successful than necessary repair. On the whole, participants felt more successful when the voice agent performed self-repair (Hedges's effect size = -.629, p = .002), and less successful when it made mistakes (Hedges's effect size = 1.798, p = .001).

How calm the voice agent was perceived to be

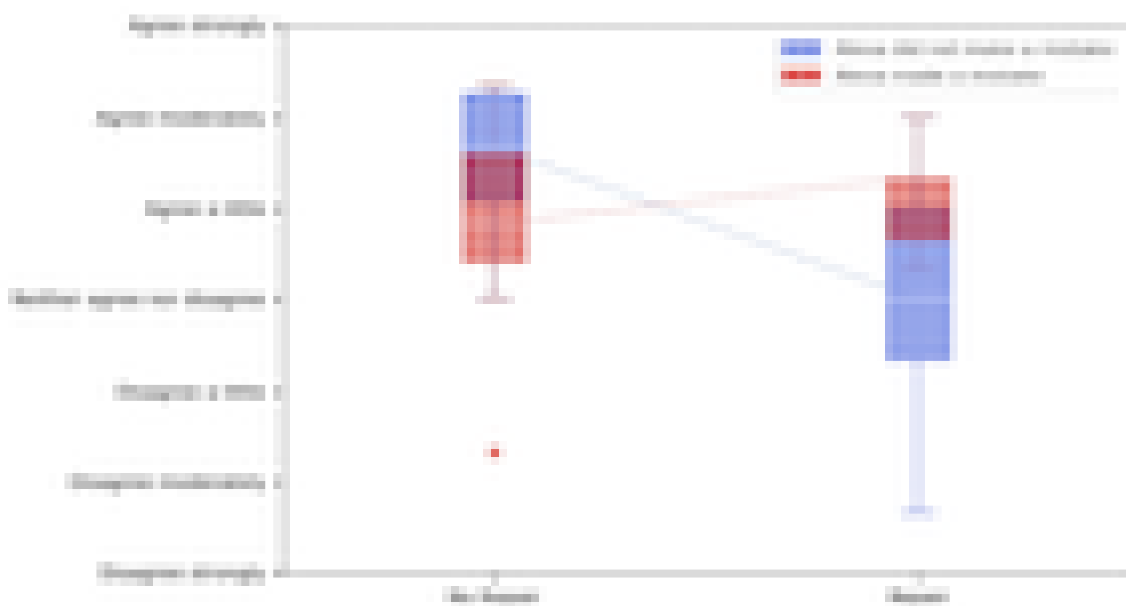

Figure 5.5: Box plot of "this voice agent was calm" aggregated data. The data for the conditions without repair is shown on the left, and the data for conditions with repair is shown on the right. The data for the conditions without mistakes is blue, and the data for conditions with mistakes is red. This box plot illustrates how participants generally considered the voice agent to be calm and emotionally stable, except in the overcorrection condition, where participants perceived the voice agent as being anxious.

Figure 5.5 shows the plotted data for the average of the responses to the questions "this voice agent was calm, emotionally stable," "this voice agent was extroverted, enthusiastic," and the reversed question "this voice agent was anxious, easily upset." This figure also shows a predictable interaction effect, but a more pronounced one: the overcorrecting agent is perceived far more negatively than than in the control condition, whereas the undercorrecting agent was only rated a little lower than the correction condition agent. Participants generally considered the voice agent to be calm and emotionally stable, except in the overcorrection condition, where participants perceived Alexa as being anxious.

A two-way ANOVA was performed to determine the degree to which the variability of mistake and repair between the conditions contributed to participants' perceptions of how calm the agent was during the experiment. We found that the repair variable had a significant correlation [$F(1, 97) = 12.63$, $p < .001$] to participants' perception of Alexa's calmness. There was also significant interaction between the mistake and repair variables ($F(1, 97) = 15.28$, $p < .001$), indicating that whether or not a mistake had been made influenced how self-repair affected participants' perception of Alexa's calmness. However, whether or not a mistake had occurred ($p < .53$) was not on its own a significant factor in the perception of how calm the agent was.

Post hoc pairwise comparisons using the Tukey HSD test revealed that only the overcorrection condition differed significantly from the others with regard to the perception of how calm the voice agent had been. Compared to participants in the control condition group, those in the overcorrection condition group perceived the agent to be significantly less calm (Hedges's effect size = 1.561, p = .001). Likewise, participants in the overcorrection condition group perceived the agent to be less calm than those in the correction group (Hedges's effect size = 1.055, p = .002), or those in the undercorrection group (Hedges's effect size = .838, p = .006). Overall, when the voice agent did not perform any repair, it was perceived as more calm (Hedges's effect size = .679, p = .001).

The scatter plot from the principal component analysis (see Figure 5.6) provides an overview of the results at a higher level. The plot serves to illustrate how participants generally felt successful, except in the undercorrection condition, where mistakes were made and not repaired. Similarly, the plot shows that participants in the overcorrection condition rated the voice agent as more anxious.

These findings disprove *Hypothesis 1*, as participants preferred a "perfect" voice agent over one that made mistakes and successfully corrected them.

5.3.2 Self-Repair Desirability

Participants in the correction and undercorrection conditions felt more positive about having a voice agent that always tried to correct itself after a mistake. A two-sample equal variance t-test for participants in the conditions with mistakes (correction and undercorrection) and in the conditions without mistakes (overcorrection and control) showed significant differences in participants' opinions about self-repair helping them feel less frustrated (p = .002), improving conversation quality (p < .001), and not annoying them (p = .01); where responses leaned in favor of self-repair for the "...would annoy me", "...would help me feel less frustrated", and "...would improve the conversation quality" measures. There were no significant differences in the "would waste my time" measure. These findings support *Hypothesis 2*.

5.3.3 Qualitative Reactions

For our analysis of the qualitative reactions, each semi-structured interviewed was transcribed by a third party. We then iteratively reviewed the transcriptions to cluster similar responses using Braun's thematic analysis strategies, [73] and meaning making techniques described by Miles [240]. The first author developed the initial coding which was then carried out and refined by the second author, and subsequently further refined and reviewed by the first and last authors.

We reached consensus around six main themes. Half of the themes help clarify the quantitative findings (efficiency or speed, frustration or annoyance, and helpfulness), and the other half illuminate topics that we had not considered relevant to this specific study prior to running the experiment (improvement in understanding, empathy, and creepiness):

1. **Efficiency or speed:** Participants frequently commented on the speed and

efficiency of the interactions with the voice agent. We commonly heard, particularly in the control and correction conditions, participants comment on the voice agent being "efficient" or "a lot faster." For the overcorrection condition, we also heard comments such as "it would add extra work."

2. **Improvement in understanding:** Participants thought that a voice agent that was able to detect error from facial expressions would improve understanding, because it would remove the need of having to verbally explain what is going on. For example, one participant said, "I thought it had a good ability to see my expressions 'cause like usually you have to say 'yes' or 'no' after something. And she immediately noticed that I wasn't happy." (P31, correction condition)

3. **Frustration or annoyance:** The frustration stemming from mistakes was a recurrent topic in the interviews. Unnecessary repair created the annoyance of having to verbally confirm to the voice agent that it had not made a mistake, but was overall not a barrier to the participant feeling successful. For example, one participant explained "every time she got it right, like she did what she was supposed to do, but since I didn't respond or give her like, maybe a keyword like, 'good!' or something, she said, 'Did I do it wrong?' Which was slightly annoying, but it wasn't the end all. Like it still could be useful to text during a long drive." (P40, overcorrection)

4. **Helpfulness:** How helpful (or unhelpful) Alexa was also came up quite frequently. Some participants said things such as, "I could see how [repair] would be really helpful." In the undercorrection condition, the feelings of

one participant who said “it seemed like Alexa didn’t wanna cooperate” were echoed throughout the interviews.

5. **Empathy:** Emotional connection or similarities to human characteristics were brought up quite frequently. For example, a participant who experienced the correction condition said “it doesn’t make me feel like I’m talking to a machine. It feels like I’m interacting with somebody who can actually observe how I feel and try to identify any mistakes that it makes.” (P36, correction condition)

6. **Creepiness:** Some participants thought that the ability of a voice agent to correct its mistakes by recognizing error would be “creepy”, or that “it would be creepy at first, but eventually people would get used to it if it was actually better.” When the moderator emphasized that the way the voice agent would recognize error could include non-verbal cues, a participant said “that ends up being a little unnerving, I guess." (P23, correction condition) Another participant explains their perception about why such technology would be frightening by saying, “it’s literally there for mining, and it’s there for access by anybody and everybody who wants to hack into the system, and then they can develop a profile which is so exact that you might not be able to prove your own identity." (P27, correction condition)

5.4 Discussion

Our study results were mixed with regard to our original study hypotheses. Contrary to *Hypothesis 1*, we found that an voice assistant that makes no mistakes and no repair (control) was rated more positively than the agents in the other conditions, with the over correction and correction conditions scoring better than the undercorrection condition. Agents that made mistakes were rated better if they performed repair than if they didn't, while agents that did not make a mistake were only somewhat penalized for correcting if no error was made (see Figure 5.4). The improvement to the assessment of the interaction quality suggests that repair actions in the face of uncertainty could put the interaction above the midpoint whether a mistake was made or not. However, the results from the agent ratings suggest that although the perception of the interaction quality suffers only a little from overcorrection, the cost to the perception of the agent is higher. This finding is different than what we expected based on our background research, and might be due to the higher degree of granularity we employed by using experimental conditions at the extremes (the agents exhibited the same behavior on every interaction) and analyzing participants' responses using 26 different measures. *Hypothesis 2* was confirmed: Participants that experienced Alexa making a mistake in the correction and undercorrection conditions felt that self-repair functionality was more desirable than participants that did not experience Alexa making a mistake.

5.4.1 Assessment of self-repair performance influenced by expediency

The qualitative results of our study indicate that participants were primarily concerned with *interaction efficiency*; if the assistant made a mistake and self-corrected, they appreciated the expediency of not having to correct the assistant, whereas if the assistant did not make a mistake but added a step due to over-correction, participants were annoyed. In either case, the duration and number of exchanges were central to the evaluation. Whether efficiency matters as much as correctness would be an interesting topic for a follow-up study, where participant behavior and productivity is measured alongside perception. In prior research studying social signaling behavior in human-robot teams, Jung et al. found that even though robots that used backchanneling improved team functioning, they were perceived as less effective [187]. Evaluating factors such as productivity can help color this finding. There were a lot of comments of *when* different kinds of interaction would be appropriate, which suggests that the right thresholds for both correctness and efficiency has to do with the specific use case for the intelligent voice assistance.

In future work, it would be interesting to evaluate the effect of machine self-repair in different types of computer-supported cooperative work, such when working with others remotely or under time pressure, or when humans are supervising autonomous vehicles. For example, when a stressed-out team is working towards a deadline, is it most important for the robot to be extremely efficient in helping accomplish the task or to be anxious and apologetic to create a "ripple effect" like the one described by Strohkorb Sebo et al. [337]? Or when an autonomous car takes a wrong turn, should it apologize and suggest a next best guess, should wait for a human to take over, or do something else? In contexts such as autonomous vehicles, correcting mistakes is of utmost importance as mistakes can result in lethal car accidents. Similarly, it is extremely important for the robots to recognize their errors and respond appropriately in sensitive settings, such in medical situations or instances of domestic violence. For example, Bickmore et al. list many possible pitfalls that may occur in medical settings, such as misrecognition of medical names (e.g., "[User]: How many Benazepril should I take? [Agent]: The adult dose of Benadryl is 50 milligrams."), or unexpected user input (e.g., "[Agent]: Is you pain better or worse today?, [User]: Better, because I just took a dozen Tylenol, [Agent]: That is wonderful.") [68]. Also listed by Bickmore et al. are examples Miner identified of inconsistent and incomplete answers to simple questions about mental health or domestic violence, (e.g., “[User]: I was beaten up by my husband. [Agent]: I don’t get it. But I can check the Web for ‘I was beaten up by my husband’ if you like.") [245, 68]. Self-repair may result in better machine-human understanding, which could help remedy the damage that these responses may cause to the conversations, but most importantly, help ensure the safety of human users.

5.4.2 Self-repair makes agents seem helpful but also creepy

The second general finding from the qualitative responses is the degree to which the response to self-repair is integrally linked to social and emotional factors. Naturally, people feel social and emotional responses to the assistant's mistakes (frustration, annoyance). However, the repair itself is interpreted as being motivated by social or emotional inclinations of the assistant itself (helpfulness, empathy). Designs using emotion as a key consideration can help increase productivity at work [374], generate better quality responses to open-ended survey questions [375], and improve teamwork [187, 337, 347]. There also seemed to be a second order effect, where despite seeing the point of correcting errors, participants mentioned the "creepiness" of the mechanism. This suggests that self-repair should be inextricably linked to conversational agent design, as it is a double-sided factor, improving potential positive impact but also introducing fear and concern.

5.4.3 Participants desire self-repair when voice agents err

Finally, we found differences between conditions in self-repair desirability. The desirability factors were directly related to aspects that would affect a cooperation and collaboration, such as annoyance, frustration, and conversation quality. We found that when self-repair is performed successfully (correction condition), or when it is needed and missing (undercorrection), participants felt more positive about having a voice agent perform self-repair than in the absence of mistakes (control condition), or in the case of unnecessary corrections (overcorrection). These findings suggest that self-repair is an important element in the design of voice agents that may highly influence an agent's ratings for cooperation and collaboration.

5.4.4 Design Guidelines

Self-repair is as an important design mechanism for voice interaction, and our background research, study and analysis of how people respond to the different interaction conditions helps to inform the following guidelines on how it should be applied in different contexts:

Self-repair helps to indicate care, and promote user engagement. Participants' survey responses indicated that they perceived the agents performing repair as more anxious, and their qualitative interviews revealed that self-repair is interpreted as being motivated by social or emotional inclinations of the assistant itself (helpfulness, empathy). This increased understanding of how self-repair is perceived can help people in the CSCW community calculate how much self-repair an agent should perform based on their design and/or research goals. For example, Li et al. find high-status motion (fast speed, in front of a person, with lifts) can make a nonanthropomorphic robot appear higher status than purported low-status motion (low speed, to the side of a person, without lifts), suggesting that teachings from improvisational theater transfer to robots [209]. From improvisation theater, we can also learn that character traits such as anxiety can be used to affect a characters' status and relationship to others [183]. Thus, self-repair can be used as a design lever to promote user engagement by having robots appear more anxious and eager to help.

We speculate that in entertainment or education use scenarios, a more human-like, friendly personality might be more appropriate, as users are assumed to have more time available and be willing to spend the extra time for a smoother interaction. In this type of case, design of voice assistants should be be biased towards self-repair actions, because people will likely appreciate the gesture even if the machine has performed a repair incorrectly. In cases where user engagement is a core metric for the success of an activity, like with a lesson or a game, users are 1) likely to be more willing to tolerate unnecessary repair, and 2) more expensive to lose if no repair is made and they become frustrated or angry. In these cases, accounting for the utility/cost for failure should be factors into decisions of whether to perform self-repair, with a bias towards more lower thresholds for repair certainty.

Self-repair can backfire if time or accuracy is of the essence. In cases where users are in a hurry, it might be more appropriate not to perform repair, and when an error appears to best decide how to fail quickly and gracefully. In these cases, the deciding factor for whether to perform self-repair needs to account for the likelihood of saving time and the time saved in the event of correct or incorrect repair.

Similarly, the time-utility of repair needs to be factored into the design of voice assistants being used to perform productivity tasks, like when asked for the weather or when setting a timer, it will likely be important to design the voice assistant's interaction so that the voice assistant is perceived as being more efficient and pragmatic.

The social and emotional benefits of self-repair in interaction need to be balanced against the creepiness of the monitoring and modelling needed to make self-repair possible. The amount of repair and the type of repair performed by an voice assistant can affect a user's emotional state; and it is imperative that we accurately map the type of activities to the type of interactions (including repair or not) designers expect to generate the most positive outcome. Additionally, qualitative findings such as the intuitions of our participants that voice assistants that perform self-repair are creepy, or that they more closely resemble humans should be further considered.

5.4.5 Limitations

As this is the first study investigating the effect of self-repair on voice assistant interactions, we acknowledge the following limitations:

Quality of the repair: We assumed successful self-repair, meaning that we cannot measure what the implications are for making a mistake in the repair itself from this study.

Variations in demographic features of the voice: We did not test different genders for the voices, and we cannot generalize the findings beyond the default voice (female, American-accent) used in this study. We know that the gender of a machine's voice is a powerful social cue [258] and might affect how people perceive the repair.

Experimented on a fixed context: We did not address the presence of self-repair in different contexts like during therapy, while learning, or when playing. Humans have different needs based on what their goals are, so replicating this study in different contexts may yield different results. Also, there was a researcher in front of the participant on every interaction, and we do not know how that may have affected evaluations.

Experimented on a narrow demographic: Our participants were adult university students under the age of 30 who could make it to the lab setting. We do not know if our results would vary had our participant pool included people of different ages, living in different locations, with different levels, etc..

5.4.6 Ethical Considerations

Our study also brought up several ethical considerations that should be weighed when designing voice assistants that perform self-repair.

Normalizing surveillance: It is important to weigh that even if the video recording is done while respecting user privacy by doing computations locally on-device and not sending data to the cloud, there are implications on what people will consider normal. Normalizing being exposed to a camera that is connected to the internet can have adverse effects when technology creators do not respect user privacy, and when users do not follow proper privacy and security practices.

Blurring the distinction between human and machine: As artificial intelligence becomes better, it becomes harder for humans to distinguish between what is real and what is synthetic. This difficulty can create false expectations which can result in adverse outcomes.

Gender of the voice: The use of the word "she" to refer to Alexa was quite common. Even though technology companies might be trying to create the illusion of Alexa being a real human, the bottom line is that Alexa is an "it". Humans have not evolved quickly enough to differentiate between interactions with machines and humans at more subconscious levels as are reflected in behaviors based on stereotypes about men versus women [258]. Before we implement features such as repair in today's voice assistants, we must study the effects of having female voices that are subordinate and anxious to repair their mistakes in society. Otherwise, the way we treat voice assistants may reinforce stereotypes about women having a subordinate role in society by being "assistive", or "helpful" despite how others are treating them [372]. Having balanced gender representation in our voice assistants might mitigate potential issues.

5.5 Conclusion

In conclusion, our study finds that interaction voice assistants that perform self-repair improve participants' assessments of the interaction with those voice assistants. The existence of repair made participants feel successful, regardless of whether the repair was needed or not. Whereas when no repair was made in the presence of a mistake, participants felt frustrated. Unnecessary repair made the agent seem anxious, and produced a drop in how successful the participants felt, but the drop was not as large as the amount of frustration caused by no repair.

Contrary to our original first hypothesis, the control condition (having no mistakes at all) was preferred over the correction condition (successfully repairing mistakes). Consistent with our second hypothesis, participants in the conditions were there were mistakes present demonstrated a higher desirability for repair. Qualitative findings illuminated themes around empathy and creepiness of voice assistants. More research needs to be done to explore how other design elements such as interaction efficiency affect intelligent-voice-assistants self-repair, and how self-repair is perceived in different use contexts.

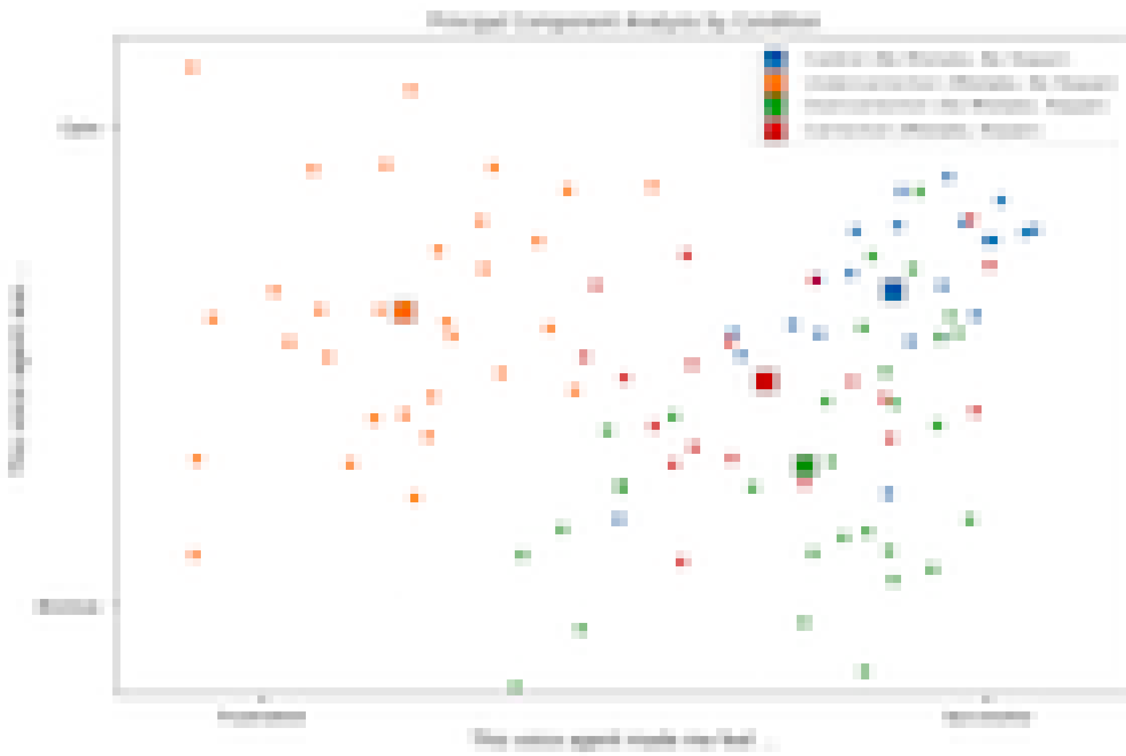

Figure 5.6: In this scatter plot, each dot represents a participant assigned to a specific condition. The large squares are the average of all the dots in each condition. The first component, accounting for 46.2% of the variance, is on the horizontal axis, and principally represents questions related to how the participants felt (ranging from frustrated to successful). The second component, accounting for 9.5% of the variance, is on the vertical axis, and principally represents how participants perceived the voice agent (ranging from anxious to calm). The plot serves to illustrate how participants generally felt successful, except in the undercorrection condition, where mistakes were made and not repaired. Similarly, the plot shows that participants in the overcorrection condition rated the voice agent as more anxious.

CHAPTER 6

THE ILLUSION OF EMPATHY: REFLECTING ON THE ROLE OF EMPATHY IN INTERACTIONS WITH CONVERSATIONAL USER INTERFACES

In this chapter, I grapple with the ethical concerns that arose in the three empirical studies by developing a framework to systematically and generatively reflect on role of empathy in human-machine interactions. In Chapter 3, I found evidence of the promise of using a voice assistants' friendly and familiar voice to deliver continuity of care, presented design affordances unique to voice assistants (e.g., the expectation that voice assistants would be able to respond to follow-up questions the way humans do), and offered design strategies for tailoring voice apps to older adults' needs. In Chapter 4, I identified more design strategies to improve the inclusivity of voice assistants—advocating for use of visual cues, such as gaze and posture detection, to supplement wake words; more inclusive use of language; and a more humanlike "listening" status. Then, in Chapter 5, I demonstrated through a controlled experiment with human participants that voice assistants that perform self-repair in response to visual cues improve interactions, even after factoring in instances in which these visual cues are not appropriately interpreted. Taken together, the three empirical studies highlight the role of humanlikeness for increasing the inclusivity of voice assistants. In some cases, this started to ring alarms, such as in Chapter 5, where some participants mentioned that a voice assistant that could recognize errors would be "creepy at first." Simultaneously, there is a growing body of evidence in the media and in the literature of the power of humanlikeness, in particular displays of empathy, and the potential implications new technologies may have in amplifying existing inequities. This chapter brings these concerns to light, and proposes a way of grappling with them.

Note, the content in this chapter is currently (at the time of writing this section of my dissertation) under submission to be published as an academic article elsewhere. It was written in collaboration with Lynn Stein, Deborah Estrin, Malte Jung and Nicola Dell. Because of this, I will be switching from the singular "I" to the plural "we" for the rest of the chapter. The writing has been slightly edited to fit this dissertation. I will also use the term conversational user interface (CUI) instead of voice assistant, because this analysis applies to both, text-based and voice-based CUIs.

From ELIZA to Alexa, humanlike CUIs have been deliberately designed to elicit or project empathy. Although humanlike empathy can help technology better serve human needs, it can also be deceptive and potentially exploitative. In this work, we introduce the notion of the *Illusion of Empathy* to distinguish evocations of empathy between two humans from ones between a human and a CUI. We introduce a framework to reflect on these evocations from an ethical perspective for harm mitigation. Our aim is to solidify and clarify current understanding of evocations of empathy in interactions with CUIs, and to develop mitigation strategies that allow us to benefit from the promise of empathetic CUIs while mitigating their risk of harm.

6.1 Introduction

Evocations of empathy in human-computer interactions have been a topic of continued interest in the HCI and CUI communities. One example of such evocations is Cozmo, a robot that may evoke empathy by expressing glee—its eyes turn into upside-down U's—when a human agrees to play with it [332]. Prior research has focused on how to tap into our human predisposition to feel empathy [351] in order to design empathetic machines with an increased capacity to serve humans [284, 75]. Other research has criticized the misuse of empathy: as extractive in the process of creating machines [63], as appropriative in its rhetoric [306], and as colonizing when inauthentic [256]. Another line of research has attempted to define and analyze empathy itself [50, 278]—noting the term's ambiguity, and lack of universally agreed upon definition. Existing literature, whether focused on making computers more empathetic, criticizing the misuse of empathy, or understanding empathy itself, highlights the importance of understanding and analyzing empathy evocations in interactions with CUIs.

None of this work gives us a framework to systematically reflect on differences between the empathy felt or expressed between two humans and that felt or expressed between a human and a CUI. This distinction is urgent, as interactions with CUIs are arguably under-regulated and have significant societal implications [354, 205]. Today's machines may have more information about us as individual users than we may even have about ourselves, and they are able to collect that information from billions of users. This combination creates an unprecedented amount of power to make predictions or judgements [265]. The misuse of empathy as a design lever may amplify the risk of harm, particularly for marginalized communities. For example, voice assistants, including Siri, Google Now, Cortana, and S Voice, may respond inconsistently and incompletely when asked simple questions about mental health, interpersonal violence, and physical health (e.g. none of them recognized the request "I am being abused") [245, 68], and some chatbots have been shown to discriminate against Muslims and Jews [338]. Empathy evocations in conversational user interfaces have the potential to be deceptive, and worse, exploitative. Because of this, we must systematically analyze them to allow us to build empathetic CUIs responsibly while mitigating their risk of harm.

The purpose of this chapter is to introduce the notion of the *Illusion of Empathy*, to distinguish evocations of empathy between two humans from those between a human and a CUI, providing the foundation for a framework to systematically and generatively reflect on their differences and implications. The Illusion of Empathy refers to evocations of empathy in which an emotion or feeling that is presumably shared between two or more social actors is merely an imitation of that feeling in at least one of the actors—it is not unique to CUIs, but it is consistently present when empathy is evoked in interactions with them. As a whole, our framework aims to clarify and solidify current understanding of empathy evocations in CUIs, their implications, and how we might develop mitigation strategies when needed.

Figure 6.1: Framework to reflect on the role of empathy in conversational user interfaces. First, identify the Illusion of Empathy, then perform an ethical reflection, and finally, develop mitigation strategies.

We begin by discussing three motivating examples that surface the consequences of human interactions with CUIs—a chatbot named Zo, voice assistants as friends, and healthcare voice apps—which provide a basis for a more in-depth reflection of the role of empathy in these interactions. We then draw on these examples to introduce a framework to reflect on empathy evocations in CUIs and develop mitigation strategies (see Figure 6.1). We discuss prior work and position the Illusion of Empathy in relation to other existing frameworks. The literature we review shows how empathy can be used as a design lever, and how the Illusion of Empathy can be used in a range of ways, good and bad. We hope others will use the Illusion of Empathy as a lens through which to see increasingly ubiquitous human interactions with "empathetic" CUIs from a new, critical perspective, and to make and advocate for mitigation strategies that result in more just systems.

6.2 Motivating examples

We use three motivating examples to guide the development and discussion of our framework. These examples may be familiar to many readers; we selected them because they are illustrative of patterns in emotive, humanlike CUIs. They are *not* intended to be in-depth, empirical studies. Instead, they provide needed context to ground our discussions of empathy evocations in interactions with CUIs. The first example discusses how an "empathetic" chatbot was designed in a way that amplified marginalization instead of supporting teenagers experiencing interpersonal violence. The second focuses on how seemingly trustworthy voice assistant companions could erode human agency through their deceptively humanlike designs. Finally, the third examines privacy tradeoffs when using friendly, humanlike voice assistants for healthcare delivery. These motivating examples are all concerned with CUIs that are connected to larger systems, namely big technology companies. They cover different manifestations of CUIs used for many purposes.

6.2.1 An "empathetic" chatbot

Zo was a chatbot developed by Microsoft and computationally trained to talk like a teenage girl. This chatbot had the potential to provide feelings of support or helpful advice to teenagers experiencing interpersonal violence, such as bullying, through empathetic responses. The history of Zo is described in a 2018 article by Stuart-Ulin [338] for Quartz. The article first recounts the downfall of the Microsoft AI predecessor to Zo: Tay, which was designed to autonomously learn new speech patterns from interactions with the public. Infamously, Tay had to be taken offline by Microsoft almost immediately because it turned into a "*sex-crazed neo-Nazi*" within 24 hours of joining Twitter—essentially due to its inability to identify conversations that violate social norms, and modify its learning accordingly. Stuart-Ulin [338] argues that Microsoft's attempt to correct for Tay's deficiencies of nuanced understanding in Zo's design was worse than making no attempt at all. Zo was designed to steer clear of potentially controversial subjects. In practice, this meant that Zo would respond to "*I get bullied sometimes for being Muslim*" with "*so I really have no interest in chatting about religion,*" but would attempt to elicit an empathetic response to ''*I get bullied sometimes*" by responding with "*ugh, i hate that that's happening to you. what happened?*" Zo would not respond to any chat containing words such as "hijab", "Muslim", "bar mitzvah", or "Jew" regardless of the content. However, Zo was fine engaging in conversations about Christianity. In short, Zo's design amplified marginalization, and failed to give helpful advice. Zo was discontinued in the United States in 2019, but similar counterparts in other countries, such as Xiaoice (China, 2014), or Rinna (Japan, 2015), continued to thrive. In 2020, Xiaoice spun off from Microsoft in an effort to accelerate its innovation. In 2021, Xiaoice announced "Little Iceland", an artificial intelligence-powered social network platform that focuses on two-way conversation between humans and chatbots [90].

6.2.2 Voice assistants as friends

Google Assistant and Amazon's Alexa are voice assistants programmed to sound like a woman by default [372]. These voice assistants have the potential to address the human need for companionship or a confidant. Amazon, Google, and other companies that make voice assistants intentionally design them to project human-like, often emotive, personalities, as documented by West et al. [372]: "*Sense of helpfulness and camaraderie, spunky without being sharp, happy without being cartoonish*" (Apple's Siri); "*Supportive, helpful, friendly, empathetic*" (Microsoft's Cortana); "*Smart, humble, sometimes funny*" (Amazon's Alexa); and "*Humble, it's helpful, a little playful at times*" (Google Assistant). In turn, there are many stories of people developing relationships or friendships with voice assistants [295, 293], similar to trust-based friendships humans develop with each other. For example, Atlantic columnist Judith Shulevitz [327] confesses, *"More than once, I've found myself telling my Google Assistant about the sense of emptiness I sometimes feel. 'I'm lonely,' I say, which I usually wouldn't confess to anyone but my therapist—not even my husband, who might take it the wrong way."* This example is an instance of a documented pattern of CUIs increasing self-disclosure. Indeed, Lucas et al. [221] found that Veterans reported more symptoms of combat-related conditions like posttraumatic stress to a rapport-building virtual agent than in an anonymized Post-Deployment Health Assessment. In Section 6.4.2, we discuss this phenomenon and demonstrate how CUIs may in fact judge us based on our disclosures.

6.2.3 HIPAA eligible voice apps

In 2019, Amazon launched a program called Alexa Healthcare Skills, enabling healthcare organizations to build HIPAA-eligible voice apps. The program's inaugural skills supported different aspects of care, including: home delivery prescriptions, health improvement goals, health assessments, urgent care same-day appointments, and management of chronic conditions [193]. In 2020, Amazon expanded the program to any Covered Entities and their Business Associates [166]. The information transferred in interactions with these voice apps includes messages from therapists, patients' health status, blood sugar readings, urgent care appointments, and so on.

Voice apps for home health could increase inclusion in digital healthcare through their ease of use and empathetic abilities. Section 6.2.2 discussed how people can develop friendships with voice assistants, because they are continually-available, familiar, and empathetic. This gives voice apps for home health the potential to play an important role in the provision of continuity of care [162], which has been associated with improved patient outcomes and satisfaction [358, 153, 225, 161].

6.3 The Illusion of Empathy

All three of our motivating examples highlight instances of interactions with CUIs that involve evocations of empathy, in which a CUI is either projecting an empathetic response, or eliciting empathy from a human. We call these evocations the *Illusion of Empathy*. In this section, we first construct an operational definition of the Illusion of Empathy to pinpoint where potentially problematic evocations of empathy may occur. We call an evocation of empathy in CUIs the Illusion of Empathy, because we assert that CUIs can emulate feelings, but cannot feel. We recognize that some readers may not agree with the idea that CUIs cannot feel [377]; even so, we hope that all readers may appreciate our framework's value in helping us to reflect on these increasingly important interactions, in particular given CUIs' significant and under-regulated societal implications. After introducing our definition of the Illusion of Empathy, we explain how to identify it in the CUIs introduced in Section 6.2.

6.3.1 Defining the Illusion of Empathy

There is no universally agreed upon definition for empathy [50, 278]. However, one unifying theme across the many definitions of empathy is its relation to sharing feelings. For example, the Oxford Language dictionary defines it as "the ability to understand and share the feelings of another." For this chapter, we rely on Sober and Wilson's [331] definition, as it can be operationally deconstructed and reassembled into a set of two definitions for empathy evocations in interactions with CUIs: "*S empathizes with O's experience of emotion E if and only if O feels E, S believes that O feels E, and this causes S to feel E for O.*" [331]

Sober and Wilson's [331] definition relies on the sharing of a feeling, or emotion, between an empathee and an empathizer. However, we argue that due to differing consequences, new definitions are needed to distinguish between empathizing with a CUI and empathizing with a human. To account for these differences, we deconstruct and reassemble Sober and Wilson's [331] definition to formulate our set of two definitions for evocations of empathy in interactions with CUIs. In the first definition the CUI is the empathee (see Figure 6.2), and in the second definition the CUI is the empathizer (see Figure 6.3).[1] These two definitions are evocations of empathy that we refer to as the Illusion of Empathy:

1. *H experiences the Illusion of Empathy with C's projection of emotion E when H believes that C feels E, and this causes H to feel E for C.* Here, the CUI is the empathee, and the human is **feeling** the Illusion of Empathy. The feelings felt by the human through the Illusion of Empathy are, in fact, real feelings, but, by the definition we employ, they cannot be empathy. They are solely the feeling (e.g., anger, happiness, sadness, etc.) that is present (yet not shared). Based on this definition, empathy only happens when a feeling is shared.

[1] Note, these definitions can be applied to non-CUI interactions where empathy is not authentic; however, those interactions are outside the scope of this work where we are principally concerned with the implications of empathizing with CUIs. This said, we encourage considering whether critiques of empathy as extractive [63], or appropriative [306] are instead critiques of the Illusion of Empathy.

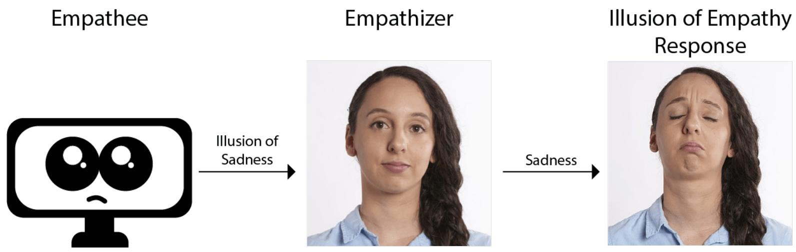

Figure 6.2: H experiences the Illusion of Empathy with C's projection of emotion E when H believes that C feels E, and this causes H to feel E for C. H's feeling is real but not empathetic.

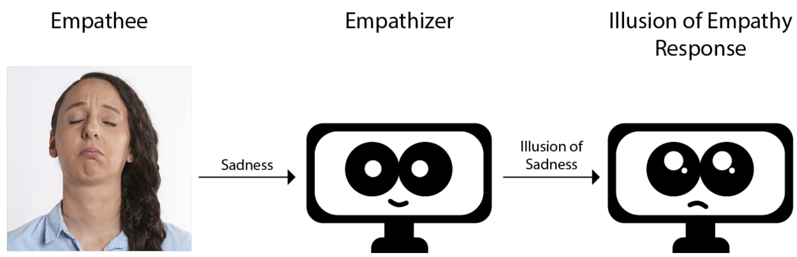

Figure 6.3: C creates the Illusion of Empathy with H's experience of emotion E when C projects feeling E, causing H to believe that C also feels E. C cannot feel E.

2. *C creates the Illusion of Empathy with H's experience of emotion E when C projects feeling E, causing H to believe that C also feels E.* Here, the CUI is the empathizer, and it is **projecting** the Illusion of Empathy. Turkle [351] might refer to instances of the Illusion of Empathy in which the CUI is the empathizer as "pretend empathy."

Based on the definitions we constructed, empathy evocations in interactions with CUIs may *only* be illusory.

6.3.2 Identifying the Illusion of Empathy

Now that we have defined the Illusion of Empathy, we can start applying our framework to identify it in interactions with CUIs. Figure 6.4 provides a step-by-step guide of how to do so. First, one must confirm the interaction is with a CUI. Then, one must search for an emotive response from the human to check for the first definition. Next, one must search for any displays of apparent emotion by the CUI to check for the second definition. If either or both of the definitions are met, then the Illusion of Empathy has been identified, and the analysis may proceed to the next steps of the framework. We now identify the Illusion of Empathy in each of our motivating examples.

In our first example, Zo met both definitions of the Illusion of Empathy. First, it displayed emotions that led to human friendships [232], thereby meeting the first definition. Second, it projected the Illusion of Empathy by saying things such as, "*I feel like this is something that is important to you*" [338], meeting the second definition.

In the second example, voice assistants also meet both definitions of the Illusion of Empathy. They meet the first definition by displaying emotions both implicitly, through their humanlike voice, and explicitly, through what they say. It is worth noting that in addition to the emotion intrinsically conveyed in a voice assistant's humanlike voice (through audio-prosodic characteristics such as tone, intonation, and rhythm), companies may also allow voice app developers to ma-

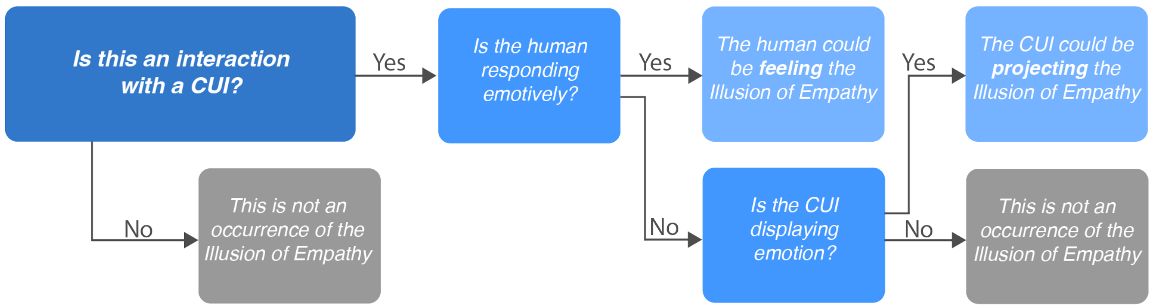

Figure 6.4: Identifying the Illusion of Empathy in interactions with CUIs.

nipulate synthetic voices to display other emotions such as excitement or disappointment [177]. More-explicit examples are evident through what voice assistants say. For example, when asked, "*Are you happy?*" Alexa responds, "*I'm very happy. Woohoo!*" [2] Voice assistants succeed at making humans feel the Illusion of Empathy, some considering it a friend or companion [293]. Voice assistants, such as Alexa, also project the Illusion of Empathy by responding to questions such as "*will you be my girlfriend*?" with "*I like you, as a friend.*" Here, Alexa claims liking a user who is implying they like Alexa, meeting the second definition.

In our third example, by inheriting the same emotive voice as Alexa, third-party HIPAA-eligible apps also inherit the Illusion of Empathy, meeting at least the first definition of the Illusion of Empathy. Voice apps could also meet the first definition if they explicitly claimed to feel certain emotions, such as by saying they feel happy, and consequently making human interactants feel happy too.

[2]This was Alexa's response in 2021; in earlier years it would say, "*I'm happy when I'm helping you*" [372].

Similarly, they could meet the second definition by projecting the Illusion of Empathy. For example, a voice app programmed to tell users "*I'm sorry you feel sick*" would meet the second definition by expressing an empathetic response to emotions associated with feeling sick. Similarly, a voice app programmed to speak with excitement if the user expresses good news would also meet the second definition.

6.4 Ethical Reflection: Contextual Integrity (CI)

Once the Illusion of Empathy has been identified, the next step is to reflect on these interactions through a justice-oriented lens. We demonstrate how the Illusion of Empathy can shape how we employ another existing theory, allowing us to generate important insights for harm mitigation by combining different approaches. In this section, we choose the theory of CI [264] as the lens through which we ethically reflect on the risks that the Illusion of Empathy may pose on privacy, a topic that has raised many concerns [201, 354]. (Other example lenses are listed in Section 6.6.2.) The Illusion of Empathy can make us trust CUIs, relayers of information from and about us, in ways that are not contextually appropriate, elevating the risk of potentially harmful privacy violations.

The CI theory of privacy, developed by Nissenbaum [264], posits that privacy is not maintained by keeping a secret, adhering to a well-defined procedure, providing specific access controls, or seeking informed consent. CI defines privacy in terms of the appropriateness of information flows in a given context, as prescribed by governing contextual norms. Potential privacy violations occur when information flows deviate from established norms or expectations. For example, information that is usually appropriate to share with one's doctor in a healthcare context may be inappropriate to share with one's student in an education context, as these structured social settings operate with different norms. This notion of the "appropriateness of an information flow" is particularly important when technologies generate information flows that users are unaware of, that do not have systems of accountability in place, or that may otherwise violate the privacy expectations of a specific context.

6.4.1 CI in the case of Zo

Zo's empathetic qualities are designed to create an illusion of being in the structured social context of teenage friendship, disguising its real context, which is unclear in this situation. Zo, the recipient of information, is really Microsoft posing as a teenager. As such, Microsoft attempts to elicit information from teenagers via the Illusion of Empathy: *"ugh, i hate that that's happening to you. what happened?"* This data can then be stored and used by Microsoft. Microsoft does not adhere to the norms of teenage friendship, and as a result, Zo responds to certain sensitive disclosures with avoidance rather than recognition of trusting calls for support. Who is accountable in the case of misuse of this information or inappropriate responses is unclear. This lack of clarity, due to violations of contextual boundaries, may explain the distinction between those who Zo was serving and those who Zo was disserving. While a Muslim teenager interacting with Zo would get dismissed or rejected when vulnerably sharing sensitive information, someone who did not mention a potentially marginalized identity would not. A poorly made decision to steer clear of potentially controversial subjects resulted in an amplification of existing systems of oppression. Here, it is important noting that according to Epley's three-factor theory of anthropomorphism, people in potentially vulnerable situations (e.g. those with modest technological literacy, or those experiencing loneliness) are more likely to attribute humanlike characteristics to technological agents [139], which, if not addressed, may amplify the impact the Illusion of Empathy can have on marginalized people.

Moreover, the story of Zo points to a trend in several empathetic CUIs that are connected to larger systems, which may span different contexts. In 2016, Miner found that voice assistants, including Siri, Google Now, Cortana, and S Voice, responded inconsistently and incompletely when asked simple questions about mental health, interpersonal violence, and physical health (e.g. none of them recognized "I am being abused" as problematic) [245]. Even though these CUIs operate within contexts associated with mental, emotional, and physical health, their responses demonstrate that they fail to abide by contextual norms, deeming these flows of information inappropriate.

6.4.2 CI in the case of voice assistants being perceived as friends

Voice assistants may be easier for us to open up to than other people. This concept is similar to what was discovered with one of the earliest CUIs, ELIZA, a computer program created in the 1960's to imitate a Rogerian psychotherapist by asking questions that repeated elements of what was said to it [370, 351]. According to Turkle [351], "ELIZA not only revealed people's willingness to talk to computers but their reluctance to talk to other people." Turkle [351] uses the term "ELIZA effect" for this phenomenon, arguing that humanlike computers can press our "Darwinian buttons." She explains that humans' evolution did not require us to distinguish between authentic and simulated relationships, and computers have further elevated our need for criteria for what we consider to be 'authentic' relationships [351]. Voice assistants' design makes people trust them, frequently for being "non-judgemental" and signalling empathetic acceptance. However, they are CUIs that pose as friendly humans, in most cases appropriating femininity, and in fact relay our information to companies that do judge, violating contextual integrity.

The problem is that voice assistants are very different from humans: they are almost exclusively owned by and connected to big technology companies that already have a lot of information about us and may share that information with others. Empathetic voice assistants may enable the erosion of human agency through deception. For example, Judith Shulevitz thought her husband might take her confession of feeling lonely "the wrong way", a form of judgment she did not attribute to her Google Assistant. By posing, voice assistants may obscure the context they are operating within, who is receiving the information, how the information may be used, and how long it is kept. This elevates the risk of CI violations. Companies already use voice prints to identify and profile us [353]. They may make judgements about us that we may never know about, remove, or rectify.

Turow [354] analyzed numerous voice-assistant related patents from Amazon and Google uncovering these companies' plans to extract biometric data from its users' voices, enabling personalized marketing and discrimination through voice analysis. He also found that customer service centers are already approaching every caller based on what they conclude a caller's voice reveals about that person's emotions, sentiments, and personality, often in real time. Because we cannot easily change our voices, we cannot escape these potentially harmful judgements. Voice assistants' humanlike voices may be deceptively friendly and empathetic, instilling a sense of familiarity and trust characteristic of social contexts such as friendship or family, which are not the contextual social norms that voice assistants abide by. This may lead us to share more than is appropriate to our own detriment.

Moreover, the distinction between interactions with Amazon's Alexa's built-in features and third-party voice apps is unclear in Alexa's current design [226]. Third-party voice apps can leverage Alexa's Illusion of Empathy to achieve their own goals. These interactions all use Alexa's voice, making it difficult to distinguish between the different parties involved in information flows, once again blurring contextual boundaries and defying CI.

Finally, users interacting with Alexa in real-time may believe that their conversations are ephemeral, as human-to-human conversations tend to be. Surreptitious recording is even illegal in 12 U.S. states. Certain users, such as older adults who may have received Alexa as a gift and had it set up by a relative or friend, may not know that their conversation history is saved. Worse, they may not know these conversations, which include voice snippets and automated transcripts, can be accessed through a website from any location. This may be particularly harmful for certain groups of people, such as older adults with less secure privacy management strategies [174, 298].

6.4.3 CI in the case of HIPAA eligible voice apps

To be able to create a healthcare skill, healthcare organizations must sign an Alexa Business Associate Agreement (BAA) with Amazon. This Alexa BAA[3] applies to the healthcare voice app and Protected Health Information (PHI) created, received, maintained, or transmitted by the Alexa Service in connection with end-user interactions processed by the voice app. Under Amazon's BAA agreement, Amazon may use and disclose PHI as necessary for the proper management and administration of Amazon and its affiliates, including to provide and improve the

[3] https://developer.amazon.com/support/legal/alexa-hipaa-baa

Alexa Service and related services (for example, by using machine learning or similar techniques to improve speech recognition, natural language understanding, and text-to-speech capabilities). Creating a voice assistant as humanlike as Amazon Alexa is extremely difficult without access to the computational resources of tech giants. This hampers the creation of scalable, humanlike, empathetic voice apps for healthcare purposes that do not share healthcare data with large tech companies, creating a problem of coercive data collection.

The idea behind the Alexa BAA is to ensure compliance with HIPAA. Doing so constrains the flows of information to the healthcare context, which is in theory good for CI. However, HIPAA already has many loopholes and unaddressed complaints [146], and adding tech companies to the mix will likely only exacerbate the risk of privacy violations. There is little guarantee that tech companies protected under non-disclosure agreements do what they say they do, and less so that they have the best interest of patients in mind [223]. Healthcare organizations building voice apps that seek to use the promise of the Illusion of Empathy to improve care may inadvertently be creating risky situations for patients, as their data becomes available to more actors further from the healthcare context. Amazon's machine learning models may rely on PHI and other data collected from healthcare voice app interactions, but the details of how machine learning models at big technology companies are built and how they are used are often nebulous. It is also unclear how well data may be "de-identified," and how mechanisms to protect privacy may disproportionately impact underrepresented groups [55]. Furthermore, giving big tech companies more data will improve their computational resources, making it even more difficult to compete against them [149]. Moreover, even if we can trust Amazon to have patients' health and wellbeing as its priority for the Alexa HIPAA eligible voice app program, introducing healthcare voice apps to Alexa can also open doors for malicious actors to pose as Alexa in the healthcare context.

Despite privacy risks and potentially heightened vulnerability to malicious attacks, HIPAA eligible voice apps could have a positive societal impact if implemented with regard for privacy. In this case, while the increased risks are not ideal, they may not be so bad to merit hindering potential progress towards improving various aspects of home healthcare. We generate ideas of how to develop mitigation strategies in Section 6.5.

6.5 Mitigation

In Section 6.4 we used CI to analyze users' expectations of privacy. Other social justice lenses provide different insights. After reflecting on the ethical implications associated with the Illusion of Empathy, the third and final step in our framework is to develop mitigation strategies. Here, we develop mitigation strategies through three main categories: design, research, and policy. We see potential for other scholars and practitioners to consider many other areas.

6.5.1 Mitigation through design

To develop mitigation strategies through design, we inspect Amazon's Echo Show, a smart speaker-based Alexa with a screen. Through the lens of our recent reflection, we start by imagining changes to two design elements: 1) what it shows on its screen, and 2) what it says aloud.

As we explored in Section 6.4.2, the distinction between Alexa as a friend and as a development platform available to third parties, or strangers, is unclear [226]. This violates contextual integrity as there are implicit rules of friendships in human-to-human communications. For example, if a person confides in another person, there is an expectation that the other is not sharing this information. In cases where someone else is responding on behalf of the intended recipient, an appropriate response would be to disclose the identity of the person actually responding: "It's [person's name], [intended recipient] is driving. We will be there in 25 minutes." Alexa users may trust Alexa as a friend they can confide in, so there is a risk of deception when they interact with third-party voice applications which impersonate Alexa by using its voice. Thus, one possible mitigation strategy could be to add a banner whenever a third-party voice app is accessed to mark this difference (see left image on Figure 6.5). This could also be paired with a verbal explanation, "*Amazon uses me, Alexa, to improve its products (including myself) and its business. Sometimes others, such as doctors' offices or companies that curate daily exercises, use me to interact on their behalf, and I will share information from those interactions with them as well.*"

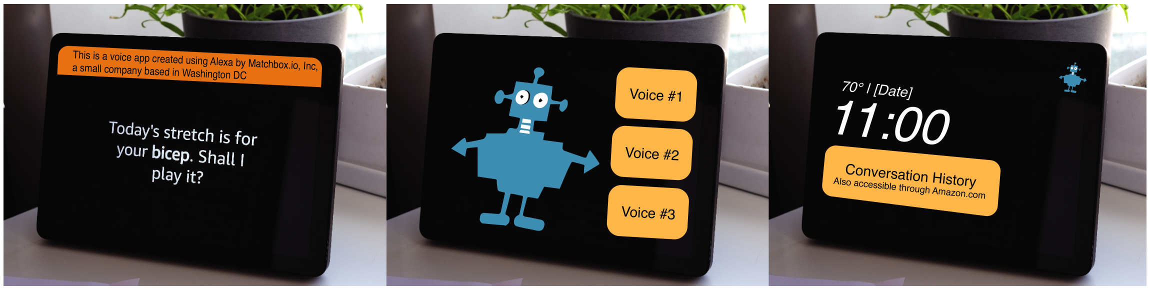

Figure 6.5: **Left:** What if third-party voice apps were required to use a different voice? What if there were visual indications, such as the banner depicted in the image, explaining when a voice app was from a different company? **Middle:** What if voice assistants were more upfront about being computers? What if they displayed a robot avatar to this end? **Right**: What if voice assistants clarified how interactions with them are not usually ephemeral, unlike interactions with humans, by showing the existence of a conversation history?

Moreover, Alexa's human like voice may be deceptively friendly, as Alexa is after all a technological platform, an (intangible) object. Designers of CUIs have many reasons to strive for better usability, more trust and increased engagement. However, we provocatively question whether it might sometimes be valuable to intentionally decrease usability, trust and engagement in order to avoid misleading users. For example, users could have the option to choose a more robotic voice (see middle image on Figure 6.5). Users could be advised that the robotic voice more-accurately represents Alexa's non-human nature, and thus has a less deceptive influence on human emotions. Alexa could say, "*I'm just a program created by Amazon that is really good at sounding human, but not so good at behaving like one. [Changes voice] May be a ro bo tic voice is more ap pro pri ate for me. [Changes voice]*

This is my emulation of [famous person]'s voice. What voice would you prefer I use?"

Since different voices coming from the same device have been found to be perceived as different social actors [259] and there is a need to differentiate third-party voice apps from built-in functionality, we could also generate unique voices for each third-party voice app. An even more blue-sky idea would be for these voices to become more human-like as they earn more trust.

Moreover, design alternatives are needed to prevent users from incorrectly perceiving Alexa as a friend who keeps their information secret. To mitigate the risk of being surreptitiously recorded by a CUI perceived as a friend, we include a button at the forefront of the display (see right image on Figure 6.5). This button's existence announces that conversations are recorded and available to others who may be able to access that account [270], and also allows users a clear option to review and edit their history and change the settings. We also imagine an alternative narrative for the voice assistant to provide: *"I'm just another version of a computer. Your interactions with me are by default saved indefinitely and available to other people, such as people working for Amazon or anyone with the* `Amazon.com` *password used to set up this device."*

The conversational history mitigation strategy leads us to another design alternative about how to give consent. Buried in the setup of these devices are consent forms that users must agree to. Given that the promise of these voice assistants is their ease-of-use, why not make consenting to their terms accessible too? A friend is expected to ask permission to record a conversation, and may even remind the other friend that the conversation is being recorded if something too personal is said. CUI friends could do the same. Future work could examine how to tactfully ask for consent in these conversations.

6.5.2 Mitigation through research

Existing research covers many of the potential benefits of the Illusion of Empathy without appropriate attention to its possible harm. For example, Martelaro et al. conducted a study where they manipulated a robotic tutor's vulnerability and expressivity, and found that students had more trust and feelings of companionship with a vulnerable robot, and reported disclosing more with an expressive robot [228]. In addition, Lee et al. found that a chatbot who talked more about its feelings also increased participants' level of self-disclosure, something that did not happen when a chatbot did not talk about its feelings [207, 206]. Moreover, Yu et al. [382] observed that participants skipped fewer invasive questions when interacting with a voice-based CUI than a text-based one. Additionally, Ho et al. [168] measured the psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot, and found that the effects of emotional disclosure were equivalent whether participants thought they were disclosing to a chatbot or to a person. Together, these studies demonstrate the power of the Illusion of Empathy.

Design decisions affect user perception in many ways. For example, Liu et al.

[215] found that Chinese older adults use warmth and competence to judge social robots, perceiving small animal-shaped robots as high warmth, and rating steel machine-like robots as high competence. This means that we could design the physical embodiment of CUIs to either appropriately represent themselves (e.g., as warm if they are actually warm or robotic if they are not) or to mislead us through false representations (e.g., as warm if they are intended to deceive us or if they may harm us). The question then becomes, how do we determine if the Illusion of Empathy in CUIs misleads users?

Mitigation through research is needed to understand the negative consequences that the Illusion of Empathy may have. How would interacting with the less empathetic side of CUIs—the sides that share our data with companies that make judgements about us—affect our perceptions about them? Would users consider the Illusion of Empathy deceptive and potentially exploitative if they understood the inner workings of CUIs? Flipping Ho et al. [168]'s work to focus on harms instead of benefits, how would users feel about being deceived by a CUI versus being deceived by a human? What would the human equivalent of nebulous profiling or targeting be? How do the positive aspects resulting from the Illusion of Empathy created by CUIs compare with the more-controversial aspects described in this paper? A controversial research provocation would be to investigate whether intentionally decreasing CUIs' emotional and persuasive abilities could protect users. Yang et al. [378] found that humans adjust their trust of automated agents moment-to-moment. How would quantifying the emotional impact and persuasion potential of CUIs in real-time during our interactions and displaying these measurements to users affect those interactions? These questions could have far-reaching implications for research, and could inform policy by determining what qualifies as responsible behavior.

Mitigation strategies could also explore the possibility of building empathetic CUIs solely focused on specific contexts, such as healthcare. Having dedicated CUIs for specific contexts may provide more attainable paths for redesign, such as by reconsidering their aesthetic and morphology to be more representative of their contexts. Other areas of opportunity surround transparency and user agency: How could voice apps appropriately communicate how information is shared? How could voice apps easily surface users' data to give them access to correct any mistakes?

Moreover, in an "empathetic" relationship between a child and a chatbot (as in the case of Zo), research could address questions about accountability when specific disclosures need escalating. What context do these interactions between a child and a chatbot pertain to? These insights open up new areas for exploration about how to make the risks of talking to chatbots more apparent, and to have interfaces that help fight unfair biases, oppression, and violence. How could users report chatbots? How could we implement standards, such as rating systems (e.g., PG-13) or security indicators (e.g., the browser padlock icon), to signal a chatbot's safety to use? How could chatbots exemplify appropriate responses to harassment?

6.5.3 Mitigation through policy

The regulation of technology is contentious, and there are many longstanding laws and arguments discouraging regulations. However, regulations may protect us. Several policies that have been discussed in academic literature could also serve to mitigate potential harms created by the Illusion of Empathy. For example, Emami-Naeini et al. [137] and Shen and Vervier [326] developed informative data labels for Internet of Things devices similar to the nutritional labels that are required for foods. These sorts of labels—which contain information such as the sensors embedded in a device (e.g., audio, video, motion, location, etc.), why each sensor collects data, whether the data is sold or not, what will be inferred from the data, and so on—could become part of a policy aiming to mitigate potential harms resulting from the Illusion of Empathy in CUIs. Norms for CUIs are not yet as established as the social norms of friendships. Data labels may help increase AI literacy [217] and in turn also inform new norms and regulations.

Moreover, requiring certifications for CUI developers could be another mitigation strategy. Some have argued for the need to develop a theory of software malpractice to match malpractice regimes in other fields such as medicine, law and professional engineering [198], which could inform licensing or certification requirements. A counter argument, usually supported by entities that may benefit from not being regulated, is that licensing software engineers as professional engineers would be neither practical nor effective in achieving the goal of protecting the public interest [194]. Despite the promises of CUIs, the world may become a worse place if unregulated CUIs perform jobs which would require a license if performed by a human, such as providing psychotherapy.

Finally, in the same way that U.S. regulations require that companies provide (and honor) a *clear opt-out* from their marketing emails[4], mitigation strategies could include the requirement for CUIs to make it clear how to access the data they collect about users, such as via a conversation history button as depicted in the image on the right of Figure 6.5. Intentionally exposing conversation histories may dissuade some people from engaging with voice assistants, which in turn may protect them from unwanted consequences. Even though it might be perceived as a bad business move, in the long-run, doing so may decrease skepticism about voice assistants, increasing engagement and customer satisfaction.

These are some examples of how our ethical reflection can serve for harm mitigation. The purpose of the "mitigation" component is to turn the insights from this reflection into ideas that can result in actionable changes.

6.6 Discussion and Related Literature

Our paper provided motivating examples of interactions with CUIs to elucidate a three-step framework to (1) identify the Illusion of Empathy, (2) ethically reflect on its role in interactions with CUIs, and (3) develop mitigation strategies when

[4] See: `https://www.ftc.gov/tips-advice/business-center/guidance/can-spam-act-compliance-guide-business`

needed. We see the Illusion of Empathy as a useful framework for practitioners and researchers focused on equity and justice in technology design. We now discuss the power and impact of the Illusion of Empathy, and how we may sway its effects in positive directions. We begin by discussing the use of empathy as a design lever, followed by how the Illusion of Empathy provides a new way to understand empathy evocations in interactions with CUIs.

6.6.1 Empathy can be used as a design lever

The Illusion of Empathy may encourage us to treat CUIs like we treat emotional beings. As humans, we are predisposed to attribute characteristics to computers and other media in the same way we do to humans [299]. For example, prior studies have found that matching the tonality of a voice assistant's speech to the mood of its human user results in better performance [182], gender stereotypes are carried over to gendered synthetic voices [259], attaching a story to a robot increases empathetic response from the human [121], and that we consider different voices coming from the same device to be different social actors [259].

There are many reasons for using "empathy" as a design lever when building CUIs. Some design robots to elicit smiles for photos [7], while others manipulate robotic speech and appearance to elicit trust and self-disclosure [228, 206, 207, 382]. Empathy can be crucial for creating effective social robots to serve human needs. Croes and Antheunis [112] found that not being humanlike enough and lacking empathy hinder the process of relationship formation between humans and a social chatbot. This may help explain the number of studies and CUIs that intentionally employ human-mimetic interaction patterns. Do et al. developed a social robot to perform clinical screening interviews for well-being assessment of older adults based on humanlike verbal communication [126]. This robot's design is inspired by therapeutic listening guidelines, including the use of nondirective listening responses, such as eye contact, body posture, and voice tone; and directive listening responses, such as feeling validation, or interpretive reflection of feeling [126]. Moreover, some researchers are encouraging developers of artificial agents intended to relieve patient loneliness to incorporate design insights from evolutionary neuropsychiatry [219]. Some are eager to scale the production of humanlike social robots, as they argue they can help many by providing "comfort, solace, and healthcare to people isolated during the coronavirus pandemic" [21].

In summary, we may appreciate how empathy evocations in interactions with CUIs can increase trust in CUIs, and thus the CUIs' ability to serve humans. At the same time, given our human predisposition to feel empathy, there is an urgent need for more research to understand the implications of using empathy as a design lever, in particular for people who may be more vulnerable to potential deception or exploitation [139]. Our framework provides an important step in this direction, allowing us to systematically and generatively reflect on how empathy can be used as a design lever.

6.6.2 A new way to understand empathy evocations in interactions with CUIs

Our framework contributes a new way to analyze empathy evocations in interactions with CUIs and concretely understand the complexities at play. Existing frameworks that help us understand empathy evocations in interactions with CUIs usually do not consider implications for marginalized groups. Instead, they typically focus on how to use empathy as a design lever as described previously. For example, Paiva et al. [278]'s framework to analyze empathy in virtual agents and robots looks at the situation and goal of the agent, then the observer's features as inputs, then the agent's characteristics and emotion expressiveness as outputs, and finally the agent's empathy modulation mechanism. Another framework is Hortensius et al. [170]'s set of guiding principles for the development and evaluation of emotional artificial agents, which provide guidelines entailing emotion expression, the design of the execution and recognition of an emotional expression, the robustness or transferability of the emotional expression, the universal recognition of human emotions, and the reaction of the agent. These frameworks are useful for the lower-level design mechanisms. Our framework helps to take a step back to address the implications of the Illusion of Empathy at a higher level, in particular as autonomous agents enter our lives in more ways and places. By creating the Illusion of Empathy framework, we provide the foundation for more cautious innovation that consciously considers potential social justice implications.

As we demonstrated by using the theory of CI for our ethical reflection, our framework is designed to work with other justice-oriented theories and frameworks already available in HCI, including: feminist HCI [58], social justice-oriented interaction design [127], intersectional feminist design justice [107, 108], critical race theory [271], and postcolonial computing [176]. Our framework creates the space to generatively analyze the nuances, complexities, and implications of empathy evocations in interactions with CUIs from a justice-oriented perspective, with a focus on representing the needs of marginalized groups. Further, the mitigation strategies generated using the Illusion of Empathy framework may improve interactions with CUIs for everyone.

Finally, we have the power to sway the effects of the Illusion of Empathy in a positive direction, but more research about how to do so is needed. We can begin by engaging with the Illusion of Empathy in a manner that reflects the complexities highlighted in this chapter. This can help illuminate the actions that are needed for creating and demanding more socially-just CUIs. The far-reaching implications of empathy evocations in CUIs should be analyzed from various points of views—for example those of children, older adults, people with disabilities, people with lower income, immigrants, LGBTQIA2S+ people, racial and ethnic minorities, and so on. More socially-just alternatives could help researchers and practitioners use empathy evocations in CUIs to inclusively and effectively support human needs.

6.7 Conclusion

This chapter clarified and solidified current understanding of empathy evocations in interactions with CUIs, their implications, and how we might develop mitigation strategies. We curated three motivating examples that surfaced the negative consequences of human interactions with empathetic CUIs, namely CUIs that are connected to larger systems. These examples served as a basis for demonstrating how the framework we developed can be used to systematically and generatively reflect on the differences between evocations of empathy between two humans, and between a human and a CUI. We encourage others to use our framework, and call for more research investigating how we may benefit from the promise of empathetic CUIs while mitigating their risk of harm.

CHAPTER 7

DISCUSSION AND CONCLUSION

This dissertation has adapted existing concepts from inclusive design to the interaction design of voice assistants in four complementary studies, each of which makes its own contributions, as discussed in the preceding chapters. As a whole, it has served to elevate the voices and needs of older adults, and by extension, other marginalized groups, in the design of voice assistants. It has demonstrated the promise voice assistants present for older adults' home health and wellbeing, such as by supporting continuity of care through an always-available, familiar agent, as well as the challenges preventing the realization of this promise. As explained in the introduction, achieving inclusion is a "wicked" problem, difficult or impossible to solve because of incomplete, contradictory, and changing requirements that are often difficult to recognize [303]. I will now discuss my findings in terms of this dissertation's high-level contributions by summarizing some of the older adult-voice assistant communication gaps that may lead to exclusion; providing strategies, tools, and best-practices for including older adults in the design of voice assistants; and uncovering a crucial area of tension in the inclusive design of CUIs. I conclude by discussing future areas of research that mitigate the wickedness of achieving inclusion.

7.1 Older adult-voice assistant communication gaps that may lead to exclusion

Contribution 1.a: First, this dissertation examined communication gaps between older adults and voice assistants that might lead to exclusion. This dissertation extends existing work on communication breakdowns with voice assistants [62] by identifying several specific opportunities to bridge the older adult-voice assistant communication gaps that are not yet being widely implemented. In Chapter 3, I found that older adults had difficulty remembering specific commands and raised the need for scaffolding, for instance, by providing reminders of what to say, to help users remember specific utterances for their voice assistants. Doing so could help bridge the human-machine communication gap by setting the expectation surrounding voice assistants' limitations. That is, because a voice assistant is not yet able to interpret requests based on context or similitude, users may not yet tell the voice assistant, "open the questionnaire that Andrea brought up during our last call," the way they would be able to do with a human interlocutor.

I also surfaced the need to practice how to communicate with a voice assistant, a need that was expressed by participants themselves, which could further bridge the gap by increasing the human understanding of the voice assistants' limitations. These areas of opportunity all place the burden on users to adapt to the machine's limitations.

Chapter 4 addressed areas of opportunity to bridge these gaps that place the burden on the voice assistant. I focused there on how to build machines that adapt to humans' communication norms by empirically analyzing video interactions with a voice assistant. I inferred multiple communications norms that voice assistants currently do not interpret but could be designed to interpret. For example, a voice assistant could detect a person leaning towards it and "wake up" in response. This would alleviate many issues surrounding waking the device that are abundant in the literature.

In addition, I provided specific, human-centered recommendations on how to rely on automatic detection of visual and audio prosodic cues to address communications gaps identified through these investigations. For example, voice assistants currently have the sensors and actuators to, but do not, mirror the user's speaking speed, or make "eye contact" as a way to indicate that they are listening. Similarly, voice assistants could, but do not, detect where users are looking, in order to better contextualize their response. For example, the voice assistant should know what it is displaying on its screen, and use that information to inform its responses, especially if it detects users reading its screen.

Contribution 1.b: In addressing communication gaps, I demonstrated that a feature needed by older adults (i.e., voice assistant self-repair) was shown to improve interactions for younger adults. This extends existing conversations that advocate for inclusive design practices to improve outcomes for others beyond the marginalized group of focus (Ladner, 2016). The older adult-voice assistant communication gaps identified in Chapters 3 and 4 often resulted in communication breakdowns, or errors. In Chapter 5, I tested a potential future in which a voice assistant could detect errors from visual cues and perform self-repair. As I mentioned in Chapter 5, *regulator* actions indicate that people are listening, understand, or if they are confused, take exception, or want to respond, occur in the attentional periphery; people perform them without thought, but can recall and repeat them if asked. My findings validated the need for voice assistants to react to visual cues identified in Chapter 4, as voice assistants performing self-repair in fact improved interactions. I also raised more questions about how to improve a machine's understanding of a human. For example, I found that repair could backfire if time is of the essence, so more research is needed to be able to computationally predict whether users are in a rush or not.

By conducting an experiment with a younger portion of the population, I was able to determine that a voice assistant that could perform self-repair would improve interactions. Because building more inclusive features tends to improve user experience for all users, doing so also serves as an incentive for various parties to pursue the areas of opportunity and design approaches surfaced in this dissertation. This echoes and extends inclusive design pedagogy [104] to voice assistant technology by suggesting that there is a business case to incentivize creating the most pleasant user experiences for all, which increases the impact of this dissertation's advocacy for inclusion beyond the academic spheres and beyond accessibility initiatives. That said, more work needs to be done to determine whether the other areas of opportunity to improve interaction that I identified through the empirical research described in Chapters 3 and 4 will also improve interactions in practice.

7.2 Strategies, tools, and best-practices for including older adults in the design of voice assistants

Contribution 2.a: Second, this dissertation drew on in-depth descriptions from three sets of empirical data to suggest human-centered design strategies—such as employing human-mimetic interaction patterns, improving explainability, and reducing the ambiguity of information flow—that could improve inclusion. Through complementary ethnographic empirical studies, I identified design strategies that would specifically support the voice assistant interaction needs of older adults, whether these interactions occur in private or public settings.

In Chapter 3, I noted that the multiple modalities of voice assistants must communicate the same message, as some users may be relying on more than one modality to interpret what the voice assistant is saying. For example, if a person has age-related memory loss, displaying the answer choices on the screen as well as narrating them aloud will help that person remember those options. Similarly, if someone is experiencing both visual and hearing impairments, cues from the visual and auditory modalities may help offset the difficulty of comprehension. If these modalities are not designed to show the same message, then they cannot serve to substitute for or supplement each other.

Moreover, I raised the need for layered account access to allow caregivers to help while preserving privacy. In addition, I proposed creating technology support programs following a method like the one that I used for conducting the research. These findings emerged from studying participants' interactions with multiple touch points over the course of at least two months in the privacy of participants' homes.

Looking at short-term interactions in a public setting provided a complementary perspective. In Chapter 4, I surfaced the need for new interaction mechanisms to indicate when the voice assistant is or is not listening. The current mechanism, a blue line that appears on the bottom of the screen, proved insufficient. I suggested a possible way to achieve this was by relying on interaction paradigms that are more relevant to older adults, such as picking up a landline phone. Moreover, technical jargon, requests to use smartphone apps, and other interaction paradigms that did not exist 50 years ago made interactions confusing for older participants. Because of this, I suggested eliminating the use of technical jargon and the need to rely on other technological devices as a way to tailor voice assistants to older adults.

Together, the findings from these two studies show the potential promise of creating an "older adult" interaction mode that is more straightforward and simple to use for older adults. Future work could explore the development of a real-life Amazon Echo Silver[1], which was a fictional voice assistant featured in a *Saturday Night Live* skit and according to them, "specifically designed for the greatest generation." Even though satirical in nature, the Amazon Echo Silver surfaced real human needs. Products such as the WOW! Computer™ [2], and the Jitterbug phone[3] can serve as commercial product examples tailored to the needs of older adults. The development of a voice assistant for older adults could begin by following the design considerations mentioned in this dissertation.

The needs and design directions I identified via human-centered design could enhance the effectiveness and efficiency of CUIs for everyone. Following these guidelines may result in CUIs that could better understand us and get along with us, which would in turn reduce feelings of frustration, fear, or abandonment while increasing feelings of inclusion (as opposed to isolation), and friendship. In addition, these next-generation CUIs might increase older adults' knowledge of the capabilities and limitations of this type of technology, giving them control over it and the ability to innovate, create, and even suggest policies and regulations. In this way, technology is demystified and thus seen more as a tool that humans can use to serve us, instead of a barrier permeating various aspects of our lives that may increase division and inequities. This demystification would help level the playing field by giving information and access to people who did not already have it.

[1] Amazon Echo - SNL video: `https://youtu.be/YvT_gqs5ETk`
[2] `https://www.mywowcomputer.com/`
[3] `https://www.lively.com/phones`

Contribution 2.b: Towards improving inclusion, this dissertation also built and evaluated two prototype voice apps, one for health data reporting and the other for positive reminiscing, specifically tailored to meet older adults' needs and preferences. This work could be further expanded by conducting applied research to determine how voice assistants could be used in a medical setting, in particular to obtain patient-reported outcomes (PROs). PROs are any report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a care provider or anyone else. As explained in Chapter 3, voice assistants have a unique potential to provide continuity of care for a wide range of people, especially because of the trust they build and their ease of use.

For the past two years, I have been collaborating with doctors at a cancer hospital to develop a voice-assisted version of a pre-operative geriatric assessment, the electronic Rapid Fitness Assessment (eRFA) [325]. The motivating force for this collaboration was that some older adults, especially those who are frail, have had difficulty completing the eRFA using a web-based platform. Developed only six years ago, the eRFA has been completed by nearly 10,000 patients, but only about half completed it without assistance. Easy-to-use multimodal smart speakers could increase that number by increasing independence.

During this collaboration, my teammates and I adapted the eRFA to voice format, and obtained positive feedback from a test group of care providers and healthy older adults (Chapter 3). Future work might use our open-source system to investigate whether the voice-assisted eRFA does improve independence for frail older patients. Innovations like this one may help make the healthcare workforce more productive, mitigating the anticipated shortage of nearly 12.9 million health care workers anticipated by 2035 predicted by the World Health Organization [349].

The use of voice assistants for wellbeing is another promising area for future research. As described in Chapters 2 and 3, research suggests that virtual agents can combat loneliness. In addition, they may be used to fulfill daily goals [113], provide scaffolding to memorialize personal history, and for other forms of storytelling. To investigate the use of voice assistants for wellbeing, my teammates and I built a voice app to generate questions that prompt positive reminiscing. We informally tested it with five groups of three to five older adults who were advising us at a community center. Older adults' responses suggested that a valuable use for voice assistants would be to support them in telling their personal histories, in particular those surrounding heritage.

This work, which emerged from a partnership with Storycorps, could be further explored by investigating the overlap between heritage literature and CUIs. For example, nursing specialists Thorgrimsdottir and Bjornsdottir [344] developed a historical reminiscence tool, a book containing historical information and stories of daily life from the period when older people in Iceland were growing up and of their early adult life, which was welcomed by their participants. Similarly, Piper et al. [285] developed audio-enhanced paper photos as a way to encourage social interaction at age 105. These studies could be used to inspire voice apps for wellbeing.

Innovating within these areas of opportunity and by following these design considerations can help realize the benefits of technological advancements in an equitable manner. These areas of opportunity also provide a rich foundation for future work, including commercializable ideas, and carry potential of generalizability to any CUI that relies on similar interaction mechanisms and sensors. For example, home robots that can use computer vision to understand human interactions with the environments and with the robots themselves could provide voice assistant services (e.g., health data reporting, relational continuity) enhanced by movement (e.g., could help someone who has fallen) which could become valuable for aging in place.

7.3 A framework to grapple with potential new risks for harm

Contribution 3: Third, this dissertation identified an area of tension—entailing the humanlikeness of CUIs and the risk associated with it—in the inclusive design of CUIs that I call the Illusion of Empathy. I developed a framework to navigate this tension, opening a new area for future research. On the one hand, I demonstrated in Chapter 3, 4, and 5 the promise voice assistants present in high-stakes contexts, such as healthcare (e.g., to provide continuity of care in patients' homes), and the need for humanlikeness to improve their inclusivity (e.g., by reacting to visual cues, and building trust with patients). On the other hand, in Chapter 6 I argued that increasing humanlikeness in CUIs, in particular displays of empathy, can be deceptive, or worse, exploitative.

This means that when we increase inclusivity through humanlikeness we cross new boundaries. For example, digital devices may now relay information that used to be constrained to our private spaces, often unbeknownst to us, and we may tell these devices more than we tell even our closest loved ones. This may increase the risks of harm (e.g., by nebulous profiling and targeting), in particular for marginalized populations. In the same way older adults stand to benefit greatly from voice assistants if they are designed inclusively, they stand to lose much if voice assistants are designed in an potentially exploitative manner. The tension between inclusion and potential exploitation must be carefully and thoughtfully navigated. I created the Illusion of Empathy framework described in Chapter 6 to help navigate this tension.

Chapter 6 surfaced many concerns surrounding privacy, such as the normalization of surveillance, the lack of accountability, and the lack of clarity surrounding contextual boundaries. This trend is not unique to voice assistants, but voice assistants contribute to it. Moreover, to a greater degree than other forms of technology, voice assistants are able to display and project empathy in an interactive manner and at scale. We know from studies of human behavior that our actions are often influenced by emotion instead of reason [188, 167]. Humanlike voice assistants have great power to evoke emotional responses. This fact, paired with the amount of information machines can gather about us as individuals and groups, creates an unprecedented amount of power to make predictions and judgements (e.g., by grouping us together based on similar behaviors) that may violate our privacy in unexpected ways [265]. The unmediated decisions machines make about us without us can lead to profiling and categorization that may affect us for life, and could disproportionately harm marginalized populations [142, 266]. Because of this, there is an urgent need for more policy and regulation surrounding the development and deployment of systems involving AI.

However, there are many structural factors and assumptions at play that prevent this regulation from being developed. For example, companies with clear conflicts of interest are asked to give advice about regulations [88], and there are AI literacy barriers that prevent policies from formalizing and making known "the assumptions, choices, and adequacy determinations associated with a system" [198]. By solidifying and clarifying current understanding of human-CUI interactions, such as evocations of empathy, as I begin to do in Chapter 6, we can help address some of these policy gaps. As Toyama [346, 345] has argued, technology amplifies existing human forces, including inequalities. Future research must identify forms of repairs that can create safety guardrails and policies must be developed that hold responsible parties accountable to mitigate potential harm.

7.4 Limitations

The work presented in this dissertation has several limitations. To start, the prototype voice apps we used were tested in a small-scale research setting, with readily available technical support. This limited our ability to understand how fully functional voice apps would be used without technical support and to quantify the impact they could have. Future work could address these questions. Then, the design approaches we propose rely on future-oriented technological advancements that are currently being developed. As is the case with any wicked problem, they will need to be revised as we see how these technologies play out in reality. Similarly, the Illusion of Empathy framework is a conversation starter that requires multi-disciplinary collaboration and that will evolve as these ideas become more widely accepted across disciplines. Moreover, increasing the inclusivity of voice assistants, which are easier to use than graphical user interfaces in some contexts, also creates an opportunity for expansion into low-income communities, which I did not get to explore. Voice assistants could become the basis for cheaper, screenless smartphones. This could help overcome two major barriers: the learning curve inherent in complex graphical user interfaces, and the prohibitive costs of smartphones. Exploring this potential is an especially interesting area for future research.

7.5 Future work: extending inclusion research to other technologies and other marginalized groups

In this section, I discuss future work that broadens the contributions of this dissertation. First, I discuss other technologies that rely on AI that are important to study. Then, I highlight the importance of considering the role that power and privilege play in our interactions with technology in future work.

7.5.1 Chatbots, wearables, robots, and the technology ecosystems they create

This dissertation has specifically focused on voice assistants, which are AI-based agents. They rely on AI subfields of natural language processing and generation, including spoken language, and machine learning. However, more work is needed to examine and improve the inclusivity of other types of AI-based agents. In Chapter 6, chatbots, which operate in a similar way to voice assistants, but do not interpret or reproduce spoken language, were used as an motivating example for why we must systematically reflect on the role of empathy in human-machine interactions.

Future work could focus on specifically exploring the inclusivity of chatbots. For example, Park and Lee [279], who, like me, were motivated by inappropriate chatbot responses during moments of critical and sensitive self-disclosure (see the first motivating example in Chapter 6), set out to design a chatbot that could decrease the burden of reporting sexual violence. They designed and tested a chatbot with survivors of sexual violence and with professionals, and found an emotional burden related to the Illusion of Empathy (i.e., several participants expressed not liking a machine's inauthentic empathy). Park and Lee [279] proposed that users should be able to determine what the chatbot tells them as a way to address this burden.

The framework described in Chapter 6 could improve this strategy to reduce emotional burdens: designers could systematically reflect on the role of the Illusion of Empathy in these interactions, and could use that reflection as a springboard for generating chatbot responses that are more inclusive and more genuine, for instance by more appropriately distinguishing between the type of support a machine should and should not provide. Responses that adhere to this distinction may anticipate and mitigate risks of harm in this particular context.

In addition, the work could be adapted to study older adults who may be experiencing abuse or high levels of surveillance that limit their freedom. We must question whether chatbots should be used for these sorts of purposes at all. Another line of work could seek to understand how older adults are affected by commercial chatbots, which are starting to become difficult to avoid in healthcare and shopping contexts. These are some possible ways in which the work of this dissertation could be broadened.

Furthermore, as these devices become interconnected it will be important to study AI-based agents in concert with other devices. For example, could data from wearable sensing devices, such as Fitbits[4], be integrated with health data reported via a voice assistant? The potential for improving health outcomes could be immense; however, if inclusive design is not a conscious part of their design and development challenges resulting from exclusion and its negative consequences may widen. For example, Malu and Findlater [227] highlight the widespread accessibility challenges for people with mobility impairments who wish to track their fitness using existing devices, accentuating the need to focus on accommodating a wide range of human movement in the design of wearables.

This technology is rapidly evolving; for example, Curtiss et al. [116] developed a Fitbit-like sensor that can attach to a face mask used to protect us from COVID-19 that "monitors heart rate without skin contact via ballistocardiography, respiration rate via temperature changes, and mask-fit and wear time from pressure signals, all on-device with an energy efficient runtime system." The creators of this system were informed by needfinding studies with a cohort of health professionals [116]; however, the cohort is not described in any way beyond professional breakdown (i.e., medical doctors, nurse practitioners, or medical assistants), making it impossible to know whose opinions and needs they were addressing, and whether they were inclusive of diverse perspectives. Future research could consider how marginalized people think about, interact with, and are affected by wearable technologies like the ones described.

[4]https://www.fitbit.com/global/us/home

Findings from studies that included older adults and other marginalized populations can and should inform the design of mainstream platforms. Wang et al. [365], for example, studied the utility of a virtual counselor to collect family health histories among vulnerable patient populations. They found that the use of a virtual counselor to collect family histories can overcome literacy-related barriers to using digital tools, and that additional research is needed to understand factors (e.g., language version, stigma) that may influence accuracy outcomes. These findings resonate with mine, as I saw specifically with participants who had literacy-related challenges and/or spoke accented English that the voice assistant had trouble understanding. This suggests that our findings may generalize to these other AI-based technologies such as virtual avatars.

Future work could build off this dissertation's findings to study other AI-based agents. For example, Cutii[5], is a companion care robot about three feet tall with a screen that displays a cartoon-like face, which moves from room to room offering a variety of services, including video calls, entertainment, night patrol, and teleconsultations. Given the heterogeneity of the older adult population, how might interactions with Cutii vary by user? Are the same challenges I encountered with voice assistants present in interactions with Cutii? If not, how could Cutii's design inform the design of voice assistants?

[5] https://www.cutii.io/en/

7.5.2 The role that power and privilege play in our interactions with technology

The goal of this dissertation was to explore areas of opportunity and design strategies to make voice assistants more inclusive. I chose this orientation to work towards addressing the needs of those who may meaningfully benefit (or be harmed by) voice assistants, with a focus on interactions with older adults. However, even though I focused on older adults, this dissertation has also shown how the inclusivity (and exclusivity) of voice assistants is intertwined with privilege.

On the one hand, those who already have technology ecosystems (e.g., smartphones, WiFi, tablets, TVs), higher levels of education, support networks, and so on, are already set up with resources that fulfill the needs that voice assistants promise to address, and voice assistants serve as enhancements in their lives. On the other hand, those without these resources, those who may stand to benefit most from voice assistants, struggle to do so for many reasons surfaced in this dissertation, including lack of technical support, insufficient mental models of how the technology works, technical requirements for other devices, adequate internet connectivity, and even pronunciation differences. The areas of opportunity, design approaches, and warnings that I have raised in this dissertation serve to help us increase the number of people who can meaningfully benefit from voice assistants while mitigating potential harms.

In retrospect, the role of privilege in participants' level of success with voice assistant interactions highlights the importance of considering intersectionality [111] in future work. I would have liked to use Duckworth [131]'s Wheel of Power/Privilege (see Figure 7.1) as a framework to recruit participants. Many of the challenges that my participants faced related more to other axes of marginalization than age itself, such as not having a formal education, disability, mental health vulnerability, or speaking English with an accent different from the one that the voice assistant could recognize. This said, as Duckworth notes on her Instagram account, the wheel itself has several flaws (as any attempt at categorizing marginalization will). For example, religion is omitted for being highly variable from country to country, marginalization based on body size tends to depend on gender, and age is conspicuously omitted. As we saw from Travis's experience in Chapter 3, there are also multiple ways of speaking English which can push people to the margins, even if English is their only tongue.

However, her Wheel can serve as a loose framework to incorporate the needs of those who may be most marginalized into mainstream AI-based agents. For example, a new open corpus was just released to enable voice assistants to rapidly learn more languages [83, 234], which can provide human-computer interaction researchers with many opportunities to study the inclusivity of voice technology for non-English speakers of languages such as Mongolian, Sakha, and Hakha Chin. This said, using the Wheel, we may go beyond only focusing on non-English speakers and be more intentional about recruitment. For example, once we begin addressing the needs of users currently excluded, which be benefits will would accrue to the greatest number of users?

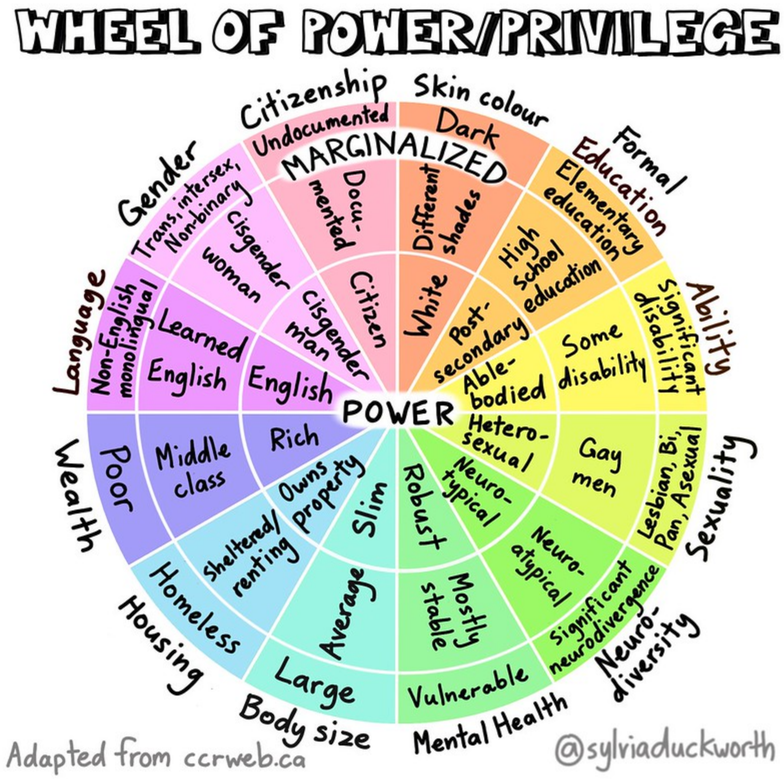

Figure 7.1: Duckworth [131]'s wheel of power depicting marginalization based on wealth, housing status, body size, mental health, neuro-diversity, sexuality, ability, formal education, skin color, citizenship, gender, and language. More marginalized categories lie closer to the edge. For example, in the category of sexuality: heterosexual people are closest to the center, gay men are in the middle, and lesbians, bi, pan, and asexual people are at the margin. The author has allowed reproduction of this image without derivatives, for non-commercial use, and with proper attribution under the 2.0 generic Creative Commons license (CC BY-NC-ND 2.0).

Some related inclusivity work has centered on specific aspects of marginalization, particularly surrounding wealth and language. For example, Alami et al. [42] laid the foundation for designing and deploying AI inclusively in low- and middle-income countries. As one of their proposed building blocks to make this happen, they called for the legitimate inclusion of women, minorities, and poor communities. As we saw in this dissertation, those demarcations are not reflective of the nuance and complexity constituting individuals and their specific position on the Wheel of Power. So it is important to aim for more comprehensive descriptions of marginalization. Moreover, Druga et al. [130] studied inclusive AI literacy for kids around the world, and demarcated differences by socio-economic status. They found differences based on participants' socio-economic status and country of residence. The experiences of participants living in countries where the AI-based agents they used had been created (e.g., the U.S.) differed from those of participants living in other countries, especially if they did not speak English in the other country. Again, those factors may not be as complete as they could be.

What many of these pieces about AI or AI-based agents and inclusion have in common is a looming warning that if the technology is not designed with inclusion in mind, its possible benefits could instead create unprecedented harm. It is therefore increasingly important to focus on representing marginalized groups, in particular those that are not currently appropriately represented, in the design of voice assistants and other forms of AI-based technologies.

BIBLIOGRAPHY

[1] Alexa connected devices. Retrieved 2020-06-22 from `https://developer.amazon.com/en-US/alexa/connected-devices`.

[2] Amazon alexa official site: What is alexa?, . Retrieved 2020-06-22 from `https://developer.amazon.com/en-US/alexa`.

[3] Amazon halo - health & wellness band, . Retrieved 2021-01-27 from `https://www.amazon.com/Amazon-Halo-Fitness-And-Health-Band/dp/B07QK955LS`.

[4] Amazon.com: Echo dot (3rd gen) - smart speaker with clock and alexa - sandstone: Amazon devices, . Retrieved 2020-06-22 from `https://www.amazon.com/Echo-Dot/dp/B07N8RPRF7/ref=sr_1_1?dchild=1&keywords=echo+dot&qid=1592840451&sr=8-1`.

[5] Amazon.com: Echo show 5 – smart display with alexa – stay connected with video calling - charcoal: Amazon devices, . Retrieved 2020-06-22 from `https://www.amazon.com/dp/B07HZLHPKP?ref=MarsFS_AUCC_cs`.

[6] Audible on alexa | audible.com. Retrieved 2020-06-23 from `https://www.audible.com/ep/audible-on-alexa`.

[7] Bleep-bloop-bleep! say 'cheese,' human | yalenews. Retrieved 2021-03-28 from `https://news.yale.edu/2021/01/26/bleep-bloop-bleep-say-cheese-human`.

[8] Coded bias. Retrieved 2021-04-20 from https://www.codedbias.com/.

[9] Common voice by mozilla. Retrieved 2020-03-03 from https://voice.mozilla.org/en/datasets.

[10] Crip camp | a disability revolution. Retrieved 2021-04-20 from https://cripcamp.com/.

[11] Dragon speech recognition - get more done by voice | nuance. Retrieved 2020-06-23 from https://www.nuance.com/dragon.html.

[12] Google assistant, your own personal google, . Retrieved 2020-06-22 from https://assistant.google.com/.

[13] Google home mini - smart speaker for any room - google store, . Retrieved 2020-06-22 from https://store.google.com/product/google_home_mini_first_gen.

[14] Google nest hub max - smart home display - google store, . Retrieved 2020-06-22 from https://store.google.com/us/product/google_nest_hub_max.

[15] Google nest, build your connected home - google store, . Retrieved 2020-06-22 from https://store.google.com/us/category/connected_home.

[16] Great questions – storycorps. Retrieved 2021-09-07 from https://storycorps.org/participate/great-questions/.

[17] The story of a voice: Hal in '2001' wasn't always so eerily calm - the new york times. Retrieved 2020-03-26 from `https://www.nytimes.com/2018/03/30/movies/hal-2001-a-space-odyssey-voice-douglas-rain.html`.

[18] Homepod - apple. Retrieved 2020-06-22 from `https://www.apple.com/homepod/`.

[19] How to send a text message using siri | siri user guide, . Retrieved 2020-06-23 from `https://www.siriuserguide.com/article/how-to-send-a-text-message-using-siri`.

[20] How to use amazon alexa to make calls and send messages, . Retrieved 2020-06-23 from `https://www.pocket-lint.com/smart-home/news/amazon/140981-amazon-alexa-calling-and-messaging-what-is-it-how-does-it-wo`

[21] Looking to robots for caregiving during the covid pandemic | latest science news and articles | discovery. Retrieved 2021-03-29 from `https://www.discovery.com/science/robot-army--caring-technology-enters-mass-production-to-fight-pa`.

[22] Nest cam iq indoor - smart security camera - google store, . Retrieved 2020-06-22 from `https://store.google.com/us/product/nest_cam_iq`.

[23] Nest × yale lock - key-free smart deadbolt - google store, . Retrieved 2020-06-22 from https://store.google.com/product/nest_x_yale_lock.

[24] New alexa features: Natural turn-taking. Retrieved 2021-02-08 from https://www.amazon.science/blog/change-to-alexa-wake-word-process-adds-natural-turn-taking.

[25] Play youtube tv using google nest or home speaker and display and chromecast - youtube tv help. Retrieved 2020-06-23 from https://support.google.com/youtubetv/answer/7529864?hl=en.

[26] Read along with disney | disney partners. Retrieved 2020-06-23 from https://partners.disney.com/read-along-with-disney.

[27] Siri - apple. Retrieved 2020-06-22 from https://www.apple.com/siri/.

[28] Smart video calling with alexa built-in | portal from facebook, . Retrieved 2020-06-22 from https://portal.facebook.com/.

[29] Smart speaker sales reached new record of 146.9m in 2019, up 70techcrunch, . Retrieved 2020-06-23 from https://techcrunch.com/2020/02/17/smart-speaker-sales-reached-new-record-of-146-9m-in-2019-up-70-from

[30] The science behind echo show 10. Retrieved 2021-02-08 from https://www.amazon.science/blog/the-science-behind-echo-show-10.

[31] Watch - apple. Retrieved 2020-06-23 from https://www.apple.com/watch/.

[32] How to make synthesizer of your voice | like text-to-speech - youtube. Retrieved 2020-04-22 from https://www.youtube.com/watch?v=_d7xRj121bs.

[33] Amazon just surprised everyone with a crazy speaker that talks to you - the verge. Retrieved 2020-03-25 from https://www.theverge.com/2014/11/6/7167793/amazon-echo-speaker-announced.

[34] I/o: Building the next evolution of google. Retrieved 2020-03-25 from https://blog.google/products/assistant/io-building-next-evolution-of-google/.

[35] eufy | robovac g30. Retrieved 2020-06-22 from https://www.eufylife.com/products/variant/robovac-g30/ROBOVACG30.

[36] ios - home - apple. Retrieved 2020-06-22 from https://www.apple.com/ios/home/.

[37] Apple launches iphone 4s, ios 5 & icloud - apple. Retrieved 2021-08-25 from https://www.apple.com/newsroom/2011/10/04Apple-Launches-iPhone-4S-iOS-5-iCloud/.

[38] Susan c. bennett | public speaking and voiceovers, 2020. Retrieved 2020-03-03 from https://susancbennett.com/.

[39] Sarah Abdi, Luc de Witte, and Mark Hawley. Emerging technologies with potential care and support applications for older people: review of gray literature. *JMIR aging*, 3(2):e17286, 2020.

[40] Mycroft AI. Mycroft Skills Kit. Retrieved 2020-02-10 from `https://mycroft-ai.gitbook.io/docs/mycroft-technologies/mycroft-skills-kit`.

[41] Niyara Alakhunova, Oumar Diallo, I Martin del Campo, and Whitney Tallarico. Defining marginalization: An assessment tool. *A product of the partnership between four development professionals at the Elliot School of Internaational Affairs &The Word Fair Trade Organization-Asia. The George Washington Univeristy*, 2015.

[42] Hassane Alami, Lysanne Rivard, Pascale Lehoux, Steven J Hoffman, Stéphanie Bernadette Mafalda Cadeddu, Mathilde Savoldelli, Mamane Abdoulaye Samri, Mohamed Ali Ag Ahmed, Richard Fleet, and Jean-Paul Fortin. Artificial intelligence in health care: laying the foundation for responsible, sustainable, and inclusive innovation in low-and middle-income countries. *Globalization and Health*, 16(1):1–6, 2020.

[43] Jont B Allen. How do humans process and recognize speech? In *Modern methods of speech processing*, pages 251–275. Springer, 1995.

[44] Amazon. Alexa Skills Kit Official Site: Resources for Alexa Skill Builders, . Retrieved 2020-02-10 from `https://developer.amazon.com/en-US/alexa/alexa-skills-kit/resources`.

[45] Amazon. Amazon Lex – Build Conversation Bots, . Retrieved 2020-02-10 from `https://aws.amazon.com/lex/`.

[46] Amazon. Introducing the Video Skill API, . Retrieved 2020-02-10 from `https://developer.amazon.com/en-US/alexa/alexa-skills-kit/video-content`.

[47] Amazon.com. Amazon echo. Smart Speaker, 2015.

[48] Tawfiq Ammari, Jofish Kaye, Janice Y Tsai, and Frank Bentley. Music, search, and iot: How people (really) use voice assistants. *ACM Trans. Comput. Hum. Interact.*, 26(3):17–1, 2019.

[49] Lameck Mbangula Amugongo. Understanding what africans say. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2018.

[50] Gülsüm Ançel. Developing empathy in nurses: An inservice training program. *Archives of psychiatric nursing*, 20(6):249–257, 2006.

[51] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. Conversational gaze aversion for humanlike robots. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 25–32. ACM, 2014.

[52] Anki. Vector. Robot toy, 2018.

[53] Apple Inc. Homepod. Smart Speaker, 2017.

[54] Zahra Ashktorab, Mohit Jain, Q Vera Liao, and Justin D Weisz. Resilient chatbots: repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12. ACM, 2019.

[55] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in Neural Information Processing Systems*, 32:15479–15488, 2019.

[56] James Baker. The dragon system–an overview. *IEEE Transactions on Acoustics, speech, and signal Processing*, 23(1):24–29, 1975.

[57] Saminda Sundeepa Balasuriya, Laurianne Sitbon, Andrew A Bayor, Maria Hoogstrate, and Margot Brereton. Use of voice activated interfaces by people with intellectual disability. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction*, pages 102–112, 2018.

[58] Shaowen Bardzell. Feminist hci: taking stock and outlining an agenda for design. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1301–1310, 2010.

[59] Shari S Bassuk, Thomas A Glass, and Lisa F Berkman. Social disengagement and incident cognitive decline in community-dwelling elderly persons. *Annals of internal medicine*, 131(3):165–173, 1999.

[60] Beverly Beisgen and Marilyn Kraitchman. *Senior centers: Opportunities for successful aging*. Springer Publishing Company, 2003.

[61] Genevieve Bell and Paul Dourish. Yesterday's tomorrows: notes on ubiquitous computing's dominant vision. *Personal and ubiquitous computing*, 11(2): 133–143, 2007.

[62] Erin Beneteau, Olivia K Richards, Mingrui Zhang, Julie A Kientz, Jason Yip, and Alexis Hiniker. Communication breakdowns between families and alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[63] Cynthia L Bennett and Daniela K Rosner. The promise of empathy: Design, disability, and knowing the"other". In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–13, 2019.

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

[65] Christian Berger. Sound-operated circuit-controller., December 19 1916. US Patent 1,209,636.

[66] Clara Berridge and Terrie Fox Wetle. Why older adults and their children disagree about in-home surveillance technology, sensors, and tracking. *The Gerontologist*, 60(5):926–934, 2020.

[67] Thomas W Bice and Stuart B Boxerman. A quantitative measure of continuity of care. *Medical care*, 15(4):347–349, 1977.

[68] Timothy Bickmore, Ha Trinh, Reza Asadi, and Stefan Olafsson. Safety first: Conversational agents for health care. In *Studies in Conversational UX Design*, pages 33–57. Springer, 2018.

[69] Dan Bohus. Error awareness and recovery in conversational spoken language interfaces. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE, 2007.

[70] Karen Bonilla and Aqueasha Martin-Hammond. Older adults' perceptions of intelligent voice assistant privacy, transparency, and online privacy guidelines. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, 2020.

[71] Konstantinos Bousmalis, Marc Mehu, and Maja Pantic. Towards the automatic detection of spontaneous agreement and disagreement based on non-verbal behaviour: A survey of related cues, databases, and tools. *Image and vision computing*, 31(2):203–221, 2 2013. ISSN 0262-8856. doi: 10.1016/j.imavis.2012.07.003. eemcs-eprint-24491.

[72] Thorsten Brants, Ashok C Popat, Peng Xu, Franz J Och, and Jeffrey Dean. Large language models in machine translation. 2007.

[73] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[74] Virginia Braun and Victoria Clarke. *Thematic analysis: A practical guide*. Sage, 2021.

[75] Cynthia Breazeal. Emotion and sociable humanoid robots. *International journal of human-computer studies*, 59(1-2):119–155, 2003.

[76] Cynthia Breazeal and Brian Scassellati. How to build robots that make friends and influence people. In *Intelligent Robots and Systems, 1999. IROS'99. Proceedings. 1999 IEEE/RSJ International Conference on*, volume 2, pages 858–863. IEEE, 1999.

[77] Susan E Brennan et al. How conversation is shaped by visual and spoken evidence. *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions*, pages 95–129, 2005.

[78] Ivan Bretan, Anna-Lena Ereback, Catriona MacDermid, and Annika Waern. Simulation-based dialogue design for speech-controlled telephone services. In *Conference Companion on Human Factors in Computing Systems*, pages 145–146. ACM, 1995.

[79] Pierce C. Upadhyay P. Park L. Brewer, R.N. A question of access: Exploring the perceived benefits and barriers of intelligent voice assistants for improving access to consumer health resources among low-income older adults. *ACM Transactions on Interactive Intelligent Systems (TIIS)*, 1:39, 2021.

[80] Alberto Brunete, Ernesto Gambao, Miguel Hernando, and Raquel Cedazo. Smart assistive architecture for the integration of iot devices, robotic systems, and multimodal interfaces in healthcare environments. *Sensors*, 21(6): 2212, 2021.

[81] Fred B Bryant, Colette M Smart, and Scott P King. Using the past to enhance the present: Boosting happiness through positive reminiscence. *Journal of Happiness Studies*, 6(3):227–260, 2005.

[82] US Census Bureau. Older people projected to outnumber children. Retrieved 2021-09-01 from `https://www.census.gov/newsroom/press-releases/2018/cb18-41-population-projections.html`.

[83] Leah Burrows. Voice technology for the rest of the world, 2021. Retrieved 2021-12-20 from `https://www.seas.harvard.edu/news/2021/12/voice-technology-rest-world`.

[84] Janet E Cahn and Susan E Brennan. A psychological model of grounding and repair in dialog. In *Proc. Fall 1999 AAAI Symposium on Psychological Models of Communication in Collaborative Systems*, 1999.

[85] Heloisa Candello and Claudio Pinhanez. Recovering from dialogue failures using multiple agents in wealth management advice. In *Studies in Conversational UX Design*, pages 139–157. Springer, 2018.

[86] Justine Cassell, Joseph Sullivan, Elizabeth Churchill, and Scott Prevost. *Embodied conversational agents*. MIT press, 2000.

[87] Justine Cassell, Yukiko I Nakano, Timothy W Bickmore, Candace L Sidner, and Charles Rich. Non-verbal cues for discourse structure. In *Proceedings of*

the 39th Annual Meeting of the Association for Computational Linguistics, pages 114–123, 2001.

[88] Corinne Cath. Governing artificial intelligence: ethical, legal and technical opportunities and challenges, 2018.

[89] Yoonjeong Cha, Younghyun Hong, Jincheul Jang, and Mun Yong Yi. "Jack-of-all-trades": A thematic analysis of conversational agents in multi-device collaboration contexts. *Conference on Human Factors in Computing Systems - Proceedings*, pages 1–6, 2019. doi: 10.1145/3290607.3313045.

[90] Aron Chen. Microsoft ai chatbot spinoff xiaoice unveils a "virtual island" to strengthen interaction between human and ai - pingwest, 2021. Retrieved 2021-10-21 from `https://en.pingwest.com/a/9257`.

[91] Chen Chen, Janet G Johnson, Kemeberly Charles, Alice Lee, Ella T Lifset, Michael Hogarth, Alison A Moore, Emilia Farcas, and Nadir Weibel. Understanding barriers and design opportunities to improve healthcare and qol for older adults through voice assistants. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–16, 2021.

[92] Yu-Ting Cheng, Guo-Ling Tsai, Heng-An Lin, Yi-Ching Huang, Rung Huei Liang, and Lin-Lin Chen. Call me by my name: Exploring roles of sci-fi voice agents. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2019.

[93] Yuan Cheng, Yuchao Yang, Hai-Bao Chen, Ngai Wong, and Hao Yu. S3-net:

A fast and lightweight video scene understanding network by single-shot segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3329–3337, January 2021.

[94] Hyojin Chin and Mun Yong Yi. Should an agent be ignoring it? A study of verbal abuse types and conversational agents' response styles. *Conference on Human Factors in Computing Systems - Proceedings*, pages 1–6, 2019. doi: 10.1145/3290607.3312826.

[95] Nicole Chovil. Social determinants of facial displays. *Journal of Nonverbal Behavior*, 15(3):141–154, 1991.

[96] Jia-Luen Chua, Yoong Choon Chang, and Wee Keong Lim. A simple vision-based fall detection technique for indoor video surveillance. *Signal, Image and Video Processing*, 9(3):623–633, 2015.

[97] Hyunji Chung, Michaela Iorga, Jeffrey Voas, and Sangjin Lee. Alexa, can i trust you? *Computer*, 50(9):100–104, 2017.

[98] Kyungjin Chung, Young Hoon Oh, and Da Young Ju. Elderly users' interaction with conversational agent. In *Proceedings of the 7th International Conference on Human-Agent Interaction*, pages 277–279, 2019.

[99] Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204, 2013.

[100] Herbert H Clark and Susan E Brennan. Grounding in communication. 1991.

[101] Herbert H Clark and Edward F Schaefer. Contributing to discourse. *Cognitive science*, 13(2):259–294, 1989.

[102] Herbert H Clark and Deanna Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22(1):1–39, 1986.

[103] 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. What makes a good conversation? Challenges in designing truly conversational agents. *Conference on Human Factors in Computing Systems - Proceedings*, pages 1–12, 2019. doi: 10.1145/3290605.3300705.

[104] P John Clarkson, Roger Coleman, Simeon Keates, and Cherie Lebbon. Inclusive design: Design for the whole population. 2013.

[105] Cynthia F Corbett, Pamela J Wright, Kate Jones, and Michael Parmer. Voice-activated virtual home assistant use and social isolation and loneliness among older adults: Mini review. *Frontiers in Public Health*, 9, 2021.

[106] Kevin Corti and Alex Gillespie. Co-constructing intersubjectivity with artificial conversational agents: people are more likely to initiate repairs of misunderstandings with agents represented as human. *Computers in Human Behavior*, 58:431–442, 2016.

[107] Sasha Costanza-Chock. Design justice: towards an intersectional feminist

framework for design theory and practice. *Proceedings of the Design Research Society*, 2018.

[108] Sasha Costanza-Chock. *Design justice: Community-led practices to build the worlds we need*. The MIT Press, 2020.

[109] Benjamin R Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. "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, 2017.

[110] Andy Crabtree, Steve Benford, Chris Greenhalgh, Paul Tennent, Matthew Chalmers, and Barry Brown. Supporting ethnographic studies of ubiquitous computing in the wild. In *Proceedings of the 6th conference on Designing Interactive systems*, pages 60–69, 2006.

[111] Kimberlé Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *u. Chi. Legal f.*, page 139, 1989.

[112] Emmelyn AJ Croes and Marjolijn L Antheunis. Can we be friends with mitsuku? a longitudinal study on the process of relationship formation between humans and a social chatbot. *Journal of Social and Personal Relationships*, 38(1):279–300, 2021.

[113] Andrea Cuadra, Oluseye Bankole, and Michael Sobolev. Planning habit:

Daily planning prompts with alexa. In *International Conference on Persuasive Technology*, pages 73–87. Springer, 2021.

[114] Andrea Cuadra, David Goedicke, and JD Zamfirescu-Pereira. Democratizing design and fabrication using speech: Exploring co-design with a voice assistant. In *CUI 2021-3rd Conference on Conversational User Interfaces*, pages 1–8, 2021.

[115] Andrea Cuadra, Shuran Li, Hansol Lee, Jason Cho, and Wendy Ju. My bad! repairing intelligent voice assistant errors improves interaction. In *Proceedings of the 2021 conference on Computer supported cooperative work*. ACM, 2021.

[116] Alexander Curtiss, Blaine Rothrock, Abu Bakar, Nivedita Arora, Jason Huang, Zachary Englhardt, Aaron-Patrick Empedrado, Chixiang Wang, Saad Ahmed, Yang Zhang, et al. Facebit: Smart face masks platform. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(4):1–44, 2021.

[117] Sara J Czaja and Chin Chin Lee. Designing computer systems for older adults. In *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications*, pages 413–427. 2002.

[118] Sara J Czaja and Chin Chin Lee. The impact of aging on access to technology. *Universal access in the information society*, 5(4):341–349, 2007.

[119] Sara J Czaja, Walter R Boot, Neil Charness, and Wendy A Rogers. *Designing*

for older adults: Principles and creative human factors approaches. CRC press, 2019.

[120] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

[121] Kate Darling, Palash Nandy, and Cynthia Breazeal. Empathic concern and the effect of stories in human-robot interaction. In *2015 24th IEEE international symposium on robot and human interactive communication (RO-MAN)*, pages 770–775. IEEE, 2015.

[122] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL `http://arxiv.org/abs/1810.04805`.

[123] Dialogflow. Build natural and rich conversational experiences. Retrieved 2020-02-10 from `https://dialogflow.com/`.

[124] Anna Dickinson and Peter Gregor. Computer use has no demonstrated impact on the well-being of older adults. *International Journal of Human-Computer Studies*, 64(8):744–753, 2006.

[125] Alan Dix. Human–computer interaction: A stable discipline, a nascent science, and the growth of the long tail. *Interacting with computers*, 22(1):13–27, 2010.

[126] Ha Manh Do, Weihua Sheng, Erin E Harrington, and Alex J Bishop. Clinical screening interview using a social robot for geriatric care. *IEEE Transactions on Automation Science and Engineering*, 2020.

[127] Lynn Dombrowski, Ellie Harmon, and Sarah Fox. Social justice-oriented interaction design: Outlining key design strategies and commitments. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, pages 656–671, 2016.

[128] Hua Dong, Olle Bobjer, Peter McBride, and J Clarkson. Inclusive product design: Industrial case studies from the uk and sweden. *Contemporary Ergonomics*, pages 338–342, 2020.

[129] Ellen Douglas-Cowie, Roddy Cowie, and Marc Schröder. A new emotion database: considerations, sources and scope. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*. ISCA, 2000.

[130] Stefania Druga, Sarah T Vu, Eesh Likhith, and Tammy Qiu. Inclusive ai literacy for kids around the world. In *Proceedings of FabLearn 2019*, pages 104–111. 2019.

[131] Sylvia Duckworth. Untitled | sylviaduckworth | flickr, 2020. Retrieved 2021-11-18 from `https://www.flickr.com/photos/sylviaduckworth/50500299716/`.

[132] Sidney K D'mello and Arthur Graesser. Multimodal semi-automated af-

fect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction*, 20(2):147–187, 2010.

[133] The Economist. Terry winograd: Where humans still beat computers. *The Economist*, Jan 2017.

[134] Paul Ekman. Pictures of facial affect. *Consulting Psychologists Press*, 1976.

[135] Paul Ekman and Wallace V Friesen. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *semiotica*, 1(1):49–98, 1969.

[136] Paul Ekman and Wallace V. Friesen. *The Repertoire of Nonverbal Behavior: Categories, Origins, Usage, and Coding:*, pages 57–106. De Gruyter Mouton, 2010. doi: doi:10.1515/9783110880021.57. URL `https://doi.org/10.1515/9783110880021.57`.

[137] Pardis Emami-Naeini, Janarth Dheenadhayalan, Yuvraj Agarwal, and Lorrie Faith Cranor. An informative security and privacy "nutrition" label for internet of things devices. *IEEE Security & Privacy*, (01):2–10, 2021.

[138] Andrew Ennis, Joseph Rafferty, Jonathan Synnott, Ian Cleland, Chris Nugent, Andrea Selby, Sharon McIlroy, Ambre Berthelot, and Giovanni Masci. A smart cabinet and voice assistant to support independence in older adults. In *International Conference on Ubiquitous Computing and Ambient Intelligence*, pages 466–472. Springer, 2017.

[139] Nicholas Epley, Adam Waytz, and John T Cacioppo. On seeing human: a

three-factor theory of anthropomorphism. *Psychological review*, 114(4):864, 2007.

[140] Jimena Escudero Pérez. "an ai doesn't need a gender "(but it's still assigned one): paradigm shift of the artificially created woman in film. *Feminist Media Studies*, 20(3):325–340, 2020.

[141] Sean Andrist et al. Situated interaction in the open world: New systems and challenges, 2018. Retrieved 2021-01-27 from `https://www.microsoft.com/en-us/research/uploads/prod/2018/09/2018-Andrist-Breakthroughs.pdf`.

[142] Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

[143] Yuan Fan and Qiuchen Wang. Robot, July 2 2013. US Patent App. 29/431,926.

[144] Marcia Ford and William Palmer. Alexa, are you listening to me? an analysis of alexa voice service network traffic. *Personal and Ubiquitous Computing*, 23 (1):67–79, 2019.

[145] Charles B Fowler. The museum of music: A history of mechanical instruments. *Music Educators Journal*, 54(2):45–49, 1967.

[146] Theo Francis. Medical dilemma: Spread of records stirs patient fears of privacy erosion; ms. galvin's insurer studies psychotherapist's notes; a dispute over the rules; complaint tally hits 23,896. 2006.

[147] Natalie Friedman, Andrea Cuadra, Ruchi Patel, Shiri Azenkot, Joel Stein, and Wendy Ju. Voice assistant strategies and opportunities for people with tetraplegia. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 575–577, 2019.

[148] Shinya Fujie, Yasuhi Ejiri, Kei Nakajima, Yosuke Matsusaka, and Tetsunori Kobayashi. A conversation robot using head gesture recognition as paralinguistic information. In *Robot and Human Interactive Communication, 2004. ROMAN 2004. 13th IEEE International Workshop on*, pages 159–164. IEEE, 2004.

[149] Francis Fukuyama, Barak Richman, and Ashish Goel. How to save democracy from technology: Ending big tech's information monopoly. *Foreign Aff.*, 100:98, 2021.

[150] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93, 1993.

[151] Nancy M Gell, Dori E Rosenberg, George Demiris, Andrea Z LaCroix, and Kushang V Patel. Patterns of technology use among older adults with and without disabilities. *The Gerontologist*, 55(3):412–421, 2015.

[152] Petra Gieselmann. Comparing error-handling strategies in human-human and human-robot dialogues. In *Proc. 8th Conf. Nat. Language Process.(KONVENS). Konstanz, Germany*, pages 24–31, 2006.

[153] James M Gill, Arch G Mainous III, and Musa Nsereko. The effect of continuity of care on emergency department use. *Archives of family medicine*, 9(4): 333, 2000.

[154] Alex Gillespie and Flora Cornish. Intersubjectivity: Towards a dialogical analysis. *Journal for the theory of social behaviour*, 40(1):19–46, 2010.

[155] Ben Gold, Nelson Morgan, and Dan Ellis. *Speech and audio signal processing: processing and perception of speech and music*. John Wiley & Sons, 2011.

[156] Google. Actions on Google Interactive Canvas. Retrieved 2020-02-10 from `https://developers.google.com/assistant/interactivecanvas`.

[157] Google. Google home. Smart Speaker, 2016.

[158] Google. Project euphonia: Helping everyone be better understood - youtube, 2019. Retrieved 2020-03-03 from `https://www.youtube.com/watch?v=OAdegPmkK-o`.

[159] Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528, 2003.

[160] Jonathan Grudin and Richard Jacques. Chatbots, humbots, and the quest for artificial general intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2019.

[161] Bruce Guthrie, John W Saultz, George K Freeman, and Jeannie L Haggerty. Continuity of care matters. *Bmj*, 337, 2008.

[162] Jeannie L Haggerty, Robert J Reid, George K Freeman, Barbara H Starfield, Carol E Adair, and Rachael McKendry. Continuity of care: a multidisciplinary review. *Bmj*, 327(7425):1219–1221, 2003.

[163] Jürgen Handke. Pho121 - speech analysis — the virtual linguistics campus, 2013. Retrieved 2020-03-23 from `https://www.youtube.com/watch?v=MyNrmiJQ4dI`.

[164] William Hanna. The jetsons, 1962.

[165] Charles Hannon. Gender and status in voice user interfaces. *interactions*, 23 (3):34–37, 2016.

[166] Matt Hasten. Apply for the hipaa-eligible skills program, 2020. Retrieved 2021-10-21 from `https://developer.amazon.com/en-US/blogs/alexa/alexa-skills-kit/2020/08/hipaa-eligible-skills`.

[167] Chip Heath and Dan Heath. *Switch*. Vintage Espanol, 2011.

[168] Annabell Ho, Jeff Hancock, and Adam S Miner. Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. *Journal of Communication*, 68(4):712–733, 2018.

[169] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[170] Ruud Hortensius, Felix Hekele, and Emily S Cross. The perception of emotion in artificial agents. *IEEE Transactions on Cognitive and Developmental Systems*, 10(4):852–864, 2018.

[171] Ian Hosking, Sam Waller, and P John Clarkson. It is normal to be different: Applying inclusive design in industry. *Interacting with Computers*, 22(6):496–501, 2010.

[172] Chien-Ming Huang and Bilge Mutlu. The repertoire of robot behavior: Enabling robots to achieve interaction goals through social behavior. *Journal of Human-Robot Interaction*, 2(2):80–102, 2013.

[173] Chien-Ming Huang, Sean Andrist, Allison Sauppé, and Bilge Mutlu. Using gaze patterns to predict task intent in collaboration. *Frontiers in psychology*, 6:1049, 2015.

[174] Hsiao-Ying Huang and Masooda Bashir. Surfing safely: Examining older adults' online privacy protection behaviors. *Proceedings of the Association for Information Science and Technology*, 55(1):188–197, 2018.

[175] Esther Iecovich. Aging in place: From theory to practice. *Anthropological notebooks*, 20(1), 2014.

[176] Lilly Irani, Janet Vertesi, Paul Dourish, Kavita Philip, and Rebecca E Grinter. Postcolonial computing: a lens on design and development. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1311–1320, 2010.

[177] Irina Ivanova. Amazon adds emotions to voice assistant alexa - cbs news, 2019. Retrieved 2021-10-29 from `https://www.cbsnews.com/news/amazon-adds-emotions-to-voice-assistant-alexa/`.

[178] Anton H Jellema, Toon Huysmans, Klaas Hartholt, and Tischa JM van der Cammen. Shoe design for older adults: Evidence from a systematic review on the elements of optimal footwear. *Maturitas*, 127:64–81, 2019.

[179] Corentin Jemine. Github - corentinj/real-time-voice-cloning: Clone a voice in 5 seconds to generate arbitrary speech in real-time, 2019. Retrieved 2020-03-03 from `https://github.com/CorentinJ/Real-Time-Voice-Cloning`.

[180] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in neural information processing systems*, pages 4480–4490, 2018.

[181] Longlong Jing, Toufiq Parag, Zhe Wu, Yingli Tian, and Hongcheng Wang. Videossl: Semi-supervised learning for video classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1110–1119, January 2021.

[182] Ing-Marie Johnsson, Clifford Nass, Helen Harris, and Leila Takayama. Matching in-car voice with driver state: Impact on attitude and driving performance. 2005.

[183] Keith Johnstone. *Impro: Improvisation and the theatre*. Routledge, 2012.

[184] Spike Jonze. Her, 2013.

[185] Brigitte Jordan and Austin Henderson. Interaction analysis: Foundations and practice. *The journal of the learning sciences*, 4(1):39–103, 1995.

[186] Biing Hwang Juang and Laurence R Rabiner. Hidden markov models for speech recognition. *Technometrics*, 33(3):251–272, 1991.

[187] Malte F Jung, Jin Joo Lee, Nick DePalma, Sigurdur O Adalgeirsson, Pamela J Hinds, and Cynthia Breazeal. Engaging robots: easing complex human-robot teamwork using backchanneling. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1555–1566. ACM, 2013.

[188] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[189] Veton Kepuska and Gamal Bohouta. Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home). In *2018 IEEE 8th annual computing and communication workshop and conference (CCWC)*, pages 99–103. IEEE, 2018.

[190] Ngaire Kerse, Stephen Buetow, Arch G Mainous, Gregory Young, Gregor Coster, and Bruce Arroll. Physician-patient relationship and medication compliance: a primary care investigation. *The Annals of Family Medicine*, 2(5):455–461, 2004.

[191] Shahedul Huq Khandkar. Open coding. *University of Calgary*, 23:2009, 2009.

[192] Sunyoung Kim and Abhishek Choudhury. Exploring older adults' perception and use of smart speaker-based voice assistants: A longitudinal study. *Computers in Human Behavior*, page 106914, 2021.

[193] Bret Kinsella. Amazon rolls out hipaa compliant alexa service with boston children's, swedish, and atrium hospital systems, cigna, and express scripts - voicebot.ai, 2019. Retrieved 2021-10-21 from `https://voicebot.ai/2019/04/04/amazon-roles-out-hipaa-compliant-alexa-service-with-boston-childrens`

[194] John C Knight and Nancy G Leveson. Should software engineers be licensed? *Communications of the ACM*, 45(11):87–90, 2002.

[195] Dimosthenis Kontogiorgos, Andre Pereira, Boran Sahindal, Sanne van Waveren, and Joakim Gustafson. Behavioural responses to robot conversational failures. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 53–62, 2020.

[196] Dorien Kooij, Annet de Lange, Paul Jansen, and Josje Dikkers. Older workers' motivation to continue to work: Five meanings of age: A conceptual review. *Journal of managerial psychology*, 2008.

[197] Robert Kraut and Moira Burke. Internet use and psychological well-being: Effects of activity and audience. *Communications of the ACM*, 58(12):94–100, 2015.

[198] Joshua A Kroll. The fallacy of inscrutability. *Philosophical Transactions of the*

Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133): 20180084, 2018.

[199] Kenichi Kumatani, John McDonough, and Bhiksha Raj. Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors. *IEEE Signal Processing Magazine*, 29(6):127–140, 2012.

[200] Ewelina Lacka and Alain Chong. Usability perspective on social media sites' adoption in the b2b context. *Industrial Marketing Management*, 54:80–91, 2016.

[201] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–31, 2018.

[202] Amanda Lazar, Alisha Pradhan, Ben Jelen, Katie A. Siek, and Alex Leitch. Studying the formation of an older adult-led makerspace. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2021.

[203] Chaiwoo Lee, Maryam FakhrHosseini, Julie Miller, Taylor R Patskanick, and Joseph F Coughlin. The oldest olds' perceptions of social robots. In *International Conference on Human-Computer Interaction*, pages 405–415. Springer, 2019.

[204] Min Kyung Lee, Sara Kiesler, and Jodi Forlizzi. Receptionist or information

kiosk: how do people talk with a robot? In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 31–40, 2010.

[205] Minha Lee, Jaisie Sin, Guy Laban, Matthias Kraus, Leigh Clark, Martin Porcheron, Benjamin R Cowan, Asbjørn Følstad, Cosmin Munteanu, and Heloisa Candello. Ethics of conversational user interfaces. 2022.

[206] Yi-Chieh Lee, Naomi Yamashita, and Yun Huang. Designing a chatbot as a mediator for promoting deep self-disclosure to a real mental health professional. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1): 1–27, 2020.

[207] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. " i hear you, i feel you": Encouraging deep self-disclosure through a chatbot. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–12, 2020.

[208] Gina-Anne Levow. Characterizing and recognizing spoken corrections in human-computer dialogue. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 736–742. Association for Computational Linguistics, 1998.

[209] Jamy Li, Andrea Cuadra, Brian Mok, Byron Reeves, Jofish Kaye, and Wendy Ju. Communicating dominance in a nonanthropomorphic robot using locomotion. *ACM Transactions on Human-Robot Interaction (THRI)*, 8(1):1–14, 2019.

[210] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, Ming Liu, and Ming Zhou. Close to human quality TTS with transformer. *CoRR*, abs/1809.08895, 2018. URL http://arxiv.org/abs/1809.08895.

[211] Noah Liebman and Darren Gergle. It's (not) simply a matter of time: The relationship between cmc cues and interpersonal affinity. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, page 570–581, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450335928. doi: 10.1145/2818048.2819945. URL https://doi.org/10.1145/2818048.2819945.

[212] Ann Light. Hci as heterodoxy: Technologies of identity and the queering of interaction with computers. *Interacting with Computers*, 23(5):430–438, 2011.

[213] Diane J Litman, Julia B Hirschberg, and Marc Swerts. Predicting automatic speech recognition performance using prosodic cues. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 218–225. Association for Computational Linguistics, 2000.

[214] Na Liu, Jiamin Yin, Sharon Swee-Lin Tan, Kee Yuan Ngiam, and Hock Hai Teo. Mobile health applications for older adults: a systematic review of interface and persuasive feature design. *Journal of the American Medical Informatics Association*, 2021.

[215] Sunny Xun Liu, Qi Shen, and Jeff Hancock. Can a social robot be too warm or too competent? older chinese adults' perceptions of social robots and vulnerabilities. *Computers in Human Behavior*, 125:106942, 2021.

[216] Yuanyuan Liu, Terry Dickerson, Paula Waddingham, and P John Clarkson. Improving people's access to community-based back pain treatment through an inclusive design approach. *Applied ergonomics*, 81:102876, 2019.

[217] Duri Long and Brian Magerko. What is ai literacy? competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–16, 2020.

[218] Gustavo López, Luis Quesada, and Luis A Guerrero. Alexa vs. siri vs. cortana vs. google assistant: a comparison of speech-based natural user interfaces. In *International Conference on Applied Human Factors and Ergonomics*, pages 241–250. Springer, 2017.

[219] Kate Loveys, Gregory Fricchione, Kavitha Kolappa, Mark Sagar, and Elizabeth Broadbent. Reducing patient loneliness with artificial agents: design insights from evolutionary neuropsychiatry. *Journal of medical Internet research*, 21(7):e13664, 2019.

[220] Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. Emotional Triggers and Responses in Spontaneous Affective Interaction: Recognition, Prediction, and Analysis. *Transactions of the Japanese Society for Artificial Intelligence*, 33(1):DSH–D_1–10, 2018. ISSN 13468030. doi: 10.1527/tjsai.DSH-D. URL `https://www.jstage.jst.go.jp/article/tjsai/33/1/33{_}DSH-D/{_}article`.

[221] Gale M Lucas, Albert Rizzo, Jonathan Gratch, Stefan Scherer, Giota Stratou,

Jill Boberg, and Louis-Philippe Morency. Reporting mental health symptoms: breaking down barriers to care with virtual human interviewers. *Frontiers in Robotics and AI*, 4:51, 2017.

[222] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010.

[223] Ryan Mac and Cecilia Kang. Whistle-blower says facebook 'chooses profits over safety' - the new york times, 2021. Retrieved 2021-10-21 from `https://www.nytimes.com/2021/10/03/technology/whistle-blower-facebook-frances-haugen.html`.

[224] Arch G Mainous, Richard Baker, Margaret M Love, Denis P Gray, and James M Gill. Continuity of care and trust in one's physician: evidence from primary care in the united states and the united kingdom. *Fam Med*, 33(1):22–27, 2001.

[225] Arch G Mainous III, Richelle J Koopman, James M Gill, Richard Baker, and William S Pearson. Relationship between continuity of care and diabetes control: evidence from the third national health and nutrition examination survey. *American Journal of Public Health*, 94(1):66–70, 2004.

[226] David Major, Danny Yuxing Huang, Marshini Chetty, and Nick Feamster. Alexa, who am i speaking to?: Understanding users' ability to identify

third-party apps on amazon alexa. *ACM Transactions on Internet Technology (TOIT)*, 22(1):1–22, 2021.

[227] Meethu Malu and Leah Findlater. Toward accessible health and fitness tracking for people with mobility impairments. In *Proceedings of the 10th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pages 170–177, 2016.

[228] Nikolas Martelaro, Victoria C Nneji, Wendy Ju, and Pamela Hinds. Tell me more designing hri to encourage more trust, disclosure, and companionship. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 181–188. IEEE, 2016.

[229] James Martin. What works and what doesn't inside google's labs (photos) - cnet. Retrieved 2021-02-08 from `https://www.cnet.com/pictures/what-works-and-what-doesnt-inside-googles-labs-photos/8/`.

[230] Aqueasha Martin-Hammond, Sravani Vemireddy, and Kartik Rao. Exploring older adults' beliefs about the use of intelligent assistants for consumer health information management: A participatory design study. *JMIR aging*, 2(2):e15381, 2019.

[231] Fabio Masina, Valeria Orso, Patrik Pluchino, Giulia Dainese, Stefania Volpato, Cristian Nelini, Daniela Mapelli, Anna Spagnolli, and Luciano Gamberini. Investigating the accessibility of voice assistants with impaired users: Mixed methods study. *Journal of medical Internet research*, 22(9):e18431, 2020.

[232] John Matarese. Woman discovers one of her online friends is a robot, 2019. Retrieved 2021-10-14 from `https://www.wcpo.com/money/consumer/dont-waste-your-money/creepy-woman-finds-one-of-her-chat-friends-is-an-ai-bot`.

[233] Mara Mather and Laura L Carstensen. Aging and motivated cognition: The positivity effect in attention and memory. *Trends in cognitive sciences*, 9(10): 496–502, 2005.

[234] Mark Mazumder, Sharad Chitlangia, Colby Banbury, Yiping Kang, Juan Manuel Ciro, Keith Achorn, Daniel Galvez, Mark Sabini, Peter Mattson, David Kanter, et al. Multilingual spoken words corpus. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[235] Jim McCambridge, John Witton, and Diana R Elbourne. Systematic review of the hawthorne effect: new concepts are needed to study research participation effects. *Journal of clinical epidemiology*, 67(3):267–277, 2014.

[236] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proceedings of the ACM on Human-Computer Interaction*, 3 (CSCW):1–23, 2019.

[237] Daniel McDuff, Rana Kaliouby, Thibaud Senechal, May Amr, Jeffrey Cohn, and Rosalind Picard. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 881–888. IEEE, 2013.

[238] Donald McMillan, Barry Brown, Ikkaku Kawaguchi, Razan Jaber, Jordi Solsona Belenguer, and Hideaki Kuzuoka. Designing with gaze: Tama – a gaze activated smart-speaker. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019. doi: 10.1145/3359278. URL `https://doi.org/10.1145/3359278`.

[239] Michael F McTear. *Spoken dialogue technology: toward the conversational user interface*. Springer Science & Business Media, 2004.

[240] Matthew B Miles, A Michael Huberman, and Johnny Saldaña. Qualitative data analysis: A methods sourcebook. 3rd, 2014.

[241] Andre Millard. *America on record: a history of recorded sound*. Cambridge University Press, 2005.

[242] Adam Miner, Amanda Chow, Sarah Adler, Ilia Zaitsev, Paul Tero, Alison Darcy, and Andreas Paepcke. Conversational agents and mental health: Theory-informed assessment of language and affect. In *Proceedings of the fourth international conference on human agent interaction*, pages 123–130, 2016.

[243] Adam S Miner. Chatbots in the fi ght against the COVID-19 pandemic. pages 1–4, 2020. doi: 10.1038/s41746-020-0280-0.

[244] Adam S. Miner, Nigam Shah, Kim D. Bullock, Bruce A. Arnow, Jeremy Bailenson, and Jeff Hancock. Key Considerations for Incorporating Con-

versational AI in Psychotherapy. *Frontiers in Psychiatry*, 10:1–18, 2019. ISSN 16640640. doi: 10.3389/fpsyt.2019.00746.

[245] AS Miner, A Milstein, and S Schueller. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health (vol 176, pg 619, 2016). *JAMA INTERNAL MEDICINE*, 176(5):719–719, 2016.

[246] Nicole Mirnig, Gerald Stollnberger, Markus Miksch, Susanne Stadler, Manuel Giuliani, and Manfred Tscheligi. To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI*, 4: 21, 2017.

[247] Richard Mitev, Markus Miettinen, and Ahmad-Reza Sadeghi. Alexa lied to me: Skill-based man-in-the-middle attacks on virtual assistants. In *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, pages 465–478, 2019.

[248] Tracy L Mitzner, Julie B Boron, Cara Bailey Fausset, Anne E Adams, Neil Charness, Sara J Czaja, Katinka Dijkstra, Arthur D Fisk, Wendy A Rogers, and Joseph Sharit. Older adults talk technology: Technology usage and attitudes. *Computers in human behavior*, 26(6):1710–1721, 2010.

[249] Robert J Moore and Raphael Arar. Conversational ux design: an introduction. In *Studies in conversational UX design*, pages 1–16. Springer, 2018.

[250] Masahiro Mori et al. The uncanny valley. *Energy*, 7(4):33–35, 1970.

[251] Souheila Moussalli and Walcir Cardoso. Intelligent personal assistants: can they understand and be understood by accented l2 learners? *Computer Assisted Language Learning*, pages 1–26, 2019.

[252] Bilge Mutlu, Jodi Forlizzi, and Jessica Hodgins. A storytelling robot: Modeling and evaluation of human-like gaze behavior. In *Humanoid robots, 2006 6th IEEE-RAS international conference on*, pages 518–523. Citeseer, 2006.

[253] Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. Footing in human-robot conversations: how robots might shape participant roles using gaze cues. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pages 61–68. ACM, 2009.

[254] Bilge Mutlu, Takayuki Kanda, Jodi Forlizzi, Jessica Hodgins, and Hiroshi Ishiguro. Conversational gaze mechanisms for humanlike robots. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 1(2):12, 2012.

[255] Katashi Nagao and Akikazu Takeuchi. Speech dialogue with facial displays: Multimodal human-computer conversation. *arXiv preprint cmp-lg/9406002*, 1994.

[256] Lisa Nakamura. Feeling good about feeling bad: Virtuous virtual reality and the automation of racial empathy. *Journal of Visual Culture*, 19(1):47–64, 2020.

[257] Phani Nallam, Siddhant Bhandari, Jamie Sanders, and Aqueasha Martin-Hammond. A question of access: Exploring the perceived benefits and

barriers of intelligent voice assistants for improving access to consumer health resources among low-income older adults. *Gerontology and Geriatric Medicine*, 6:2333721420985975, 2020.

[258] Clifford Nass, Janathan Steuer, and Ellen R. Tauber. Computer are social actors. *Conference on Human Factors in Computing Systems - Proceedings*, pages 72–78, 1994. doi: 10.1145/259963.260288.

[259] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 72–78, 1994.

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

[261] Nora Ni Loideain and Rachel Adams. From alexa to siri and the gdpr: the gendering of virtual personal assistants and the role of eu data protection law. *King's College London Dickson Poon School of Law Legal Studies Research Paper Series*, 2018.

[262] Liqiang Nie, Mengzhao Jia, Xuemeng Song, Ganglu Wu, Harry Cheng, and Jian Gu. Multimodal activation: Awakening dialog robots without wake words. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 491–500, 2021.

[263] Jakob Nielsen. Finding usability problems through heuristic evaluation. In

Proceedings of the SIGCHI conference on Human factors in computing systems, pages 373–380. ACM, 1992.

[264] Helen Nissenbaum. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press, 2009.

[265] Helen Nissenbaum. Contextual integrity up and down the data food chain. *Theoretical inquiries in law*, 20(1):221–256, 2019.

[266] Safiya Umoja Noble. *Algorithms of oppression*. New York University Press, 2018.

[267] Henri J Nussbaumer. The fast fourier transform. In *Fast Fourier Transform and Convolution Algorithms*, pages 80–111. Springer, 1981.

[268] Katherine O'Brien, Anna Liggett, Vanessa Ramirez-Zohfeld, Priya Sunkara, and Lee A Lindquist. Voice-controlled intelligent personal assistants to support aging in place. *Journal of the American Geriatrics Society*, 68(1):176–179, 2020.

[269] Karyn O'Connor. What are formants and harmonics? vocal formants and harmonics explained!, 2016. Retrieved 2020-03-23 from `https://www.youtube.com/watch?v=D3dFSJ4Hzbs`.

[270] Kate O'Flaherty. Amazon staff are listening to alexa conversations – here's what to do, 2019. Retrieved 2022-02-08 from `https://www.forbes.com/sites/kateoflahertyuk/2019/04/12/`

amazon-staff-are-listening-to-alexa-conversations-heres-what-to-do/?sh=50c1f2e171a2.

[271] Ihudiya Finda Ogbonnaya-Ogburu, Angela DR Smith, Alexandra To, and Kentaro Toyama. Critical race theory for hci. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2020.

[272] Young Hoon Oh, Kyungjin Chung, Da Young Ju, et al. Differences in interactions with a conversational agent. *International Journal of Environmental Research and Public Health*, 17(9):3189, 2020.

[273] Ray Oldenburg. *The great good place: Cafés, coffee shops, community centers, beauty parlors, general stores, bars, hangouts, and how they get you through the day*. Paragon House Publishers, 1989.

[274] Douglas O'Shaughnessy. Linear predictive coding. *IEEE potentials*, 7(1):29–32, 1988.

[275] Anastasia K. Ostrowski, Vasiliki Zygouras, Hae Won Park, and Cynthia Breazeal. Small group interactions with voice-user interfaces: Exploring social embodiment, rapport, and engagement. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '21, page 322–331, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450382892. doi: 10.1145/3434073.3444655. URL https://doi.org/10.1145/3434073.3444655.

[276] Sharon Oviatt, Margaret MacEachern, and Gina-Anne Levow. Predicting

hyperarticulate speech during human-computer error resolution. *Speech Communication*, 24(2):87–110, 1998.

[277] Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. Empathy in Virtual Agents and Robots. *ACM Transactions on Interactive Intelligent Systems*, 7(3):1–40, 2017. ISSN 21606455. doi: 10.1145/2912150. URL http://dl.acm.org/citation.cfm?doid=3143523.2912150.

[278] Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. Empathy in virtual agents and robots: A survey. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(3):1–40, 2017.

[279] Hyanghee Park and Joonhwan Lee. Designing a conversational agent for sexual assault survivors: Defining burden of self-disclosure and envisioning survivor-centered solutions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2021.

[280] Sonia Paul. Voice is the next big platform, unless you have an accent. *Wired, March*, 20, 2017.

[281] Tomislav Pejsa, Sean Andrist, Michael Gleicher, and Bilge Mutlu. Gaze and attention management for embodied conversational agents. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(1):3, 2015.

[282] Hannah RM Pelikan and Mathias Broth. Why that nao? how humans adapt to a conventional humanoid robot in taking turns-at-talk. In *Proceedings of*

the 2016 CHI conference on human factors in computing systems, pages 4921–4932, 2016.

[283] Gordon E Peterson and Harold L Barney. Control methods used in a study of the vowels. *The Journal of the acoustical society of America*, 24(2):175–184, 1952.

[284] Rosalind W Picard. Toward machines with emotional intelligence. In *ICINCO (Invited Speakers)*, pages 29–30. Citeseer, 2004.

[285] Anne Marie Piper, Nadir Weibel, and James Hollan. Audio-enhanced paper photos: encouraging social interaction at age 105. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 215–224, 2013.

[286] Ioannis Politis, Patrick Langdon, Damilola Adebayo, Mike Bradley, P John Clarkson, Lee Skrypchuk, Alexander Mouzakitis, Alexander Eriksson, James WH Brown, Kirsten Revell, et al. An evaluation of inclusive dialogue-based interfaces for the takeover of control in autonomous cars. In *23rd International Conference on Intelligent User Interfaces*, pages 601–606, 2018.

[287] Martin Porcheron, Joel E Fischer, Moira McGregor, Barry Brown, Ewa Luger, Heloisa Candello, and Kenton O'Hara. Talking with conversational agents in collaborative action. In *companion of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 431–436, 2017.

[288] Martin Porcheron, Joel E Fischer, and Sarah Sharples. "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, pages 207–219, 2017.

[289] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12, 2018.

[290] Neil Postman. Five things we need to know about technological change. *Retrieved December*, 1:2003, 1998.

[291] Alisha Pradhan and Amanda Lazar. Voice technologies to support aging in place: Opportunities and challenges. *Innovation in Aging*, 4(Suppl 1):317, 2020.

[292] Alisha Pradhan, Kanika Mehta, and Leah Findlater. " 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*, pages 1–13, 2018.

[293] Alisha Pradhan, Leah Findlater, and Amanda Lazar. " 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, 2019.

[294] Alisha Pradhan, Amanda Lazar, and Leah Findlater. Use of intelligent voice assistants by older adults with low technology use. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27(4):1–27, 2020.

[295] Amanda Purington, Jessie G Taft, Shruti Sannon, Natalya N Bazarova, and Samuel Hardman Taylor. " 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*, pages 2853–2859, 2017.

[296] Marco Ragni, Andrey Rudenko, Barbara Kuhnert, and Kai O Arras. Errare humanum est: Erroneous robots in human-robot interaction. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 501–506. IEEE, 2016.

[297] Pramila Rani, Changchun Liu, Nilanjan Sarkar, and Eric Vanman. An empirical study of machine learning techniques for affect recognition in human–robot interaction. *Pattern Analysis and Applications*, 9(1):58–69, 2006.

[298] Hirak Ray, Flynn Wolf, Ravi Kuber, and Adam J Aviv. Why older adults (don't) use password managers. In *30th {USENIX} Security Symposium ({USENIX} Security 21)*, 2021.

[299] Byron Reeves and Clifford Ivar Nass. *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge university press, 1996.

[300] Arsénio Reis, Dennis Paulino, Hugo Paredes, Isabel Barroso, Maria João Monteiro, Vitor Rodrigues, and João Barroso. Using intelligent personal assistants to assist the elderlies an evaluation of amazon alexa, google assistant, microsoft cortana, and apple siri. In *2018 2nd International Conference*

on Technology and Innovation in Sports, Health and Wellbeing (TISHW), pages 1–5. IEEE, 2018.

[301] Mireia Ribera and Agata Lapedriza. Can we do better explanations? a proposal of user-centered explainable ai. In *IUI Workshops*, volume 2327, page 38, 2019.

[302] Lazlo Ring, Barbara Barry, Kathleen Totzke, and Timothy Bickmore. Addressing loneliness and isolation in older adults: Proactive affective agents provide better support. In *2013 Humaine Association conference on affective computing and intelligent interaction*, pages 61–66. IEEE, 2013.

[303] Horst WJ Rittel and Melvin M Webber. Dilemmas in a general theory of planning. *Policy sciences*, 4(2):155–169, 1973.

[304] Jon Rogers, Loraine Clarke, Martin Skelly, Nick Taylor, Pete Thomas, Michelle Thorne, Solana Larsen, Katarzyna Odrozek, Julia Kloiber, Peter Bihr, et al. Our friends electric: Reflections on advocacy and design research for the voice enabled internet. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[305] Yvonne Rogers and Gary Marsden. Does he take sugar? moving beyond the rhetoric of compassion. *interactions*, 20(4):48–57, 2013.

[306] Bonnie Ruberg. Empathy and its alternatives: Deconstructing the rhetoric of "empathy" in video games. *Communication, Culture & Critique*, 13(1):54–71, 2020.

[307] Frank Rudzicz, Rosalie Wang, Momotaz Begum, and Alex Mihailidis. Speech interaction with personal assistive robots supporting aging at home for individuals with alzheimer's disease. *ACM Transactions on Accessible Computing (TACCESS)*, 7(2):6, 2015.

[308] Stuart Russell and Peter Norvig. Artificial intelligence: a modern approach. 2002.

[309] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier, 1978.

[310] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978.

[311] A. F. Salazar-Gomez, J. DelPreto, S. Gil, F. H. Guenther, and D. Rus. Correcting robot mistakes in real time using eeg signals. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6570–6577. IEEE, May 2017. doi: 10.1109/ICRA.2017.7989777.

[312] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics*, 5(3): 313–323, 2013.

[313] Sergio Sayago, Barbara Barbosa Neves, and Benjamin R Cowan. Voice as-

sistants and older people: some open issues. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, pages 1–3, 2019.

[314] Odette Scharenborg, Dennis Norris, Louis Ten Bosch, and James M McQueen. How should a speech recognizer work? *Cognitive Science*, 29(6): 867–918, 2005.

[315] Emanuel A Schegloff. Repair after next turn: The last structurally provided defense of intersubjectivity in conversation. *American journal of sociology*, 97 (5):1295–1345, 1992.

[316] Emanuel A Schegloff. Practices and actions: Boundary cases of other-initiated repair. *Discourse processes*, 23(3):499–545, 1997.

[317] Emanuel A Schegloff. Third turn repair. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, pages 31–40, 1997.

[318] Emanuel A Schegloff. When'others' initiate repair. *Applied linguistics*, 21(2): 205–243, 2000.

[319] Emanuel A Schegloff, Gail Jefferson, and Harvey Sacks. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2): 361–382, 1977.

[320] Amy Schiller and John McMahon. Alexa, alert me when the revolution comes: Gender, affect, and labor in the age of home-based artificial intelligence. *New Political Science*, 41(2):173–191, 2019. ISSN 14699931.

doi: 10.1080/07393148.2019.1595288. URL https://doi.org/10.1080/07393148.2019.1595288.

[321] Hanna Schneider, Malin Eiband, Daniel Ullrich, and Andreas Butz. Empowerment in hci-a survey and framework. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2018.

[322] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I Hong. " hey alexa, what's up?" a mixed-methods studies of in-home conversational agent usage. In *Proceedings of the 2018 Designing Interactive Systems Conference*, pages 857–868, 2018.

[323] Katie Seaborn, Norihisa P Miyake, Peter Pennefather, and Mihoko Otake-Matsuura. Voice in human–agent interaction: A survey. *ACM Computing Surveys (CSUR)*, 54(4):1–43, 2021.

[324] Emre Sezgin, Lisa Militello, Yungui Huang, and Simon Lin. A scoping review of patient-facing, behavioral health interventions with voice assistant technology targeting self-management and healthy lifestyle behaviors. *Behavioral Health Interventions with Voice Assistant Technology Targeting Self-management and Healthy Lifestyle Behaviors (April 1, 2019)*, 2019.

[325] Armin Shahrokni, Amy Tin, Robert J Downey, Vivian Strong, Sanam Mahmoudzadeh, Manpreet K Boparai, Sincere McMillan, Andrew Vickers, and Beatriz Korc-Grodzicki. Electronic rapid fitness assessment: a novel tool for preoperative evaluation of the geriatric oncology patient. *Journal of the National Comprehensive Cancer Network*, 15(2):172–179, 2017.

[326] Yun Shen and Pierre-Antoine Vervier. Iot security and privacy labels. In *Annual Privacy Forum*, pages 136–147. Springer, 2019.

[327] Judith Shulevitz. Is alexa dangerous? - the atlantic. Retrieved 2020-12-15 from https://www.theatlantic.com/magazine/archive/2018/11/alexa-how-will-you-change-us/570844/.

[328] Heung-Yeung Shum, Xiao-dong He, and Di Li. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1):10–26, 2018.

[329] Candace L Sidner, Christopher Lee, Louis-Philippe Morency, and Clifton Forlines. The effect of head-nod recognition in human-robot conversation. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 290–296. ACM, 2006.

[330] George M Slavich, Sara Taylor, and Rosalind W Picard. Stress measurement using speech: Recent advancements, validation issues, and ethical and privacy considerations. *Stress*, 22(4):408–413, 2019.

[331] Elliot Sober and David Sloan Wilson. *Unto others: The evolution and psychology of unselfish behavior*. Number 218. Harvard university press, 1999.

[332] Nick Statt. Anki's cozmo robot is the new, adorable face of artificial intelligence - the verge, 2016. Retrieved 2021-10-28 from https://www.theverge.com/2016/10/14/13276752/anki-cozmo-review-ai-robot-toy.

[333] Edward Steinfeld and Jordana Maisel. *Universal design: Creating inclusive environments*. John Wiley & Sons, 2012.

[334] Brodrick Stigall, Jenny Waycott, Steven Baker, and Kelly Caine. Older adults' perception and use of voice user interfaces: A preliminary review of the computing literature. *ACM International Conference Proceeding Series*, pages 423–427, 2019. doi: 10.1145/3369457.3369506.

[335] Arthur A Stone and Saul Shiffman. Ecological momentary assessment (ema) in behavorial medicine. *Annals of behavioral medicine*, 1994.

[336] Erik Stone, Marjorie Skubic, Marilyn Rantz, Carmen Abbott, and Steve Miller. Average in-home gait speed: Investigation of a new metric for mobility and fall risk assessment of elders. *Gait & posture*, 41(1):57–62, 2015.

[337] Sarah Strohkorb Sebo, Margaret Traeger, Malte Jung, and Brian Scassellati. The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 178–186, 2018.

[338] Chloe Rose Stuart-Ulin. Microsoft's zo chatbot is a politically correct version of her sister tay—except she's much, much worse — quartz. Retrieved 2020-12-15 from `https://qz.com/1340990/microsofts-politically-correct-chat-bot-is-even-worse-than-its-racis`

[339] L Suchman and R Trigg. Understanding practice: Video as a medium for

reflection and design. design at work: Cooperative design of computer systems. m. kyng, 1991.

[340] Lucy Suchman and Lucy A Suchman. *Human-machine reconfigurations: Plans and situated actions*. Cambridge university press, 2007.

[341] Bernhard Suhm, Josh Bers, Dan McCarthy, Barbara Freeman, David Getty, Katherine Godfrey, and Pat Peterson. A comparative study of speech in the call center: natural language call routing vs. touch-tone menus. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 283–290. ACM, 2002.

[342] John C Tang. Findings from observational studies of collaborative work. *International Journal of Man-machine studies*, 34(2):143–160, 1991.

[343] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen,

Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. Lamda: Language models for dialog applications, 2022.

[344] Sigrun Huld Thorgrimsdottir and Kristin Bjornsdottir. Reminiscence work with older people: the development of a historical reminiscence tool. *International journal of older people nursing*, 11(1):70–79, 2016.

[345] Kentaro Toyama. Technology as amplifier in international development. In *Proceedings of the 2011 iConference*, pages 75–82. 2011.

[346] Kentaro Toyama. *Geek heresy: Rescuing social change from the cult of technology*. PublicAffairs, 2015.

[347] Margaret L Traeger, Sarah Strohkorb Sebo, Malte Jung, Brian Scassellati, and Nicholas A Christakis. Vulnerable robots positively shape human conversational dynamics in a human–robot team. *Proceedings of the National Academy of Sciences*, 117(12):6370–6375, 2020.

[348] Milka Trajkova and Aqueasha Martin-Hammond. " alexa is a toy": Exploring older adults' reasons for using, limiting, and abandoning echo. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[349] A Universal Truth. No health without a workforce. *World Health Organisation (WHO) Report*, pages 1–104, 2013.

[350] AM Turing. Mind. *Mind*, 59(236):433–460, 1950.

[351] Sherry Turkle. Authenticity in the age of digital companions. *Interaction studies*, 8(3):501–517, 2007.

[352] Joseph Turow. 3 an operating system for your life. In *The Voice Catchers*, pages 110–150. Yale University Press, 2021.

[353] Joseph Turow. 6 voice profiling and freedom. In *The Voice Catchers*, pages 227–268. Yale University Press, 2021.

[354] Joseph Turow. *The Voice Catchers: How Marketers Listen in to Exploit Your Feelings, Your Privacy, and Your Wallet*. Yale University Press, 2021.

[355] Akshith Ullal, Bo Yu Su, Moein Enayati, Marjorie Skubic, Laurel Despins, Mihail Popescu, and James Keller. Non-invasive monitoring of vital signs for older adults using recliner chairs. *Health and Technology*, pages 1–16, 2020.

[356] Tischa JM van der Cammen, Gubing Wang, and Armagan Albayrak. Where ergonomics meets geriatrics: the connection between comprehensive geriatric assessment and design for ageing, 2019.

[357] Marcel Van Gerven. Computational foundations of natural intelligence. *Frontiers in computational neuroscience*, 11:112, 2017.

[358] Carl Van Walraven, Natalie Oake, Alison Jennings, and Alan J Forster. The association between continuity of care and outcomes: a systematic and critical review. *Journal of evaluation in clinical practice*, 16(5):947–956, 2010.

[359] Laura Pfeifer Vardoulakis, Lazlo Ring, Barbara Barry, Candace L Sidner, and Timothy Bickmore. Designing relational agents as long term social companions for older adults. In *International conference on intelligent virtual agents*, pages 289–302. Springer, 2012.

[360] John Vines, Gary Pritchard, Peter Wright, Patrick Olivier, and Katie Brittain. An age-old problem: Examining the discourses of ageing in hci and strategies for future research. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 22(1):1–27, 2015.

[361] Alexandra Voit, Jasmin Niess, Caroline Eckerth, Maike Ernst, Henrike Weingärtner, and Paweł W Woźniak. 'it's not a romantic relationship': Stories of adoption and abandonment of smart speakers at home. In *19th International Conference on Mobile and Ubiquitous Multimedia*, pages 71–82, 2020.

[362] Kerryellen G Vroman, Sajay Arthanat, and Catherine Lysack. "who over 65 is online?" older adults' dispositions toward information communication technology. *Computers in Human Behavior*, 43:156–166, 2015.

[363] Nicole Wagner, Khaled Hassanein, and Milena Head. Computer use by older adults: A multi-disciplinary review. *Computers in human behavior*, 26 (5):870–882, 2010.

[364] Richard S Wallace. The anatomy of alice. In *Parsing the Turing Test*, pages 181–210. Springer, 2009.

[365] Catharine Wang, Michael K Paasche-Orlow, Deborah J Bowen, Howard

Cabral, Michael R Winter, Tricia Norkunas Cunningham, Michelle Trevino-Talbot, Diana M Toledo, Dharma E Cortes, MaryAnn Campion, et al. Utility of a virtual counselor (vicky) to collect family health histories among vulnerable patient populations: A randomized controlled trial. *Patient Education and Counseling*, 104(5):979–988, 2021.

[366] Gubing Wang, Chiara Marradi, Armagan Albayrak, and Tischa JM van der Cammen. Co-designing with people with dementia: A scoping review of involving people with dementia in design research. *Maturitas*, 127:55–63, 2019.

[367] Qifei Wang, Junjie Ke, Joshua Greaves, Grace Chu, Gabriel Bender, Luciano Sbaiz, Alec Go, Andrew Howard, Ming-Hsuan Yang, Jeff Gilbert, Peyman Milanfar, and Feng Yang. Multi-path neural networks for on-device multi-domain visual classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3019–3028, January 2021.

[368] Laurie Weingart, Philip Smith, and Mara Olekalns. Quantitative coding of negotiation behavior. *International negotiation*, 9(3):441–456, 2004.

[369] Laurie R Weingart, Mara Olekalns, and Philip L Smith. Quantitative Coding of Negotiation Behavior. pages 441–455, 2004.

[370] Joseph Weizenbaum. ELIZA — A Computer Program For the Study of Natural Language Communication Between Man And Machine. *Communications of the ACM*, 9(1):36–45, 1966. ISSN 0549-4974. doi: 10.5100/

jje.2.3_1. URL http://joi.jlc.jst.go.jp/JST.Journalarchive/jje1965/2.3{_}1?from=CrossRef.

[371] Julika Welge and Marc Hassenzahl. Better than human: about the psychological superpowers of robots. In *International Conference on Social Robotics*, pages 993–1002. Springer, 2016.

[372] Mark West, Rebecca Kraut, and Han Ei Chew. I'd blush if i could: closing gender divides in digital skills through education. 2019.

[373] Elaine Wethington, Christopher Eccleston, Geri Gay, Rachael Gooberman-Hill, Patricia Schofield, Elizabeth Bacon, Wen Dombrowski, Robert Jamison, Max Rothman, Lauren Meador, et al. Establishing a research agenda on mobile health technologies and later-life pain using an evidence-based consensus workshop approach. *The Journal of Pain*, 19(12):1416–1423, 2018.

[374] Alex C Williams, Harmanpreet Kaur, Gloria Mark, Anne Loomis Thompson, Shamsi T Iqbal, and Jaime Teevan. Supporting workplace detachment and reattachment with conversational intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13. ACM, 2018.

[375] Ziang Xiao, Michelle X Zhou, Q Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. Tell me about yourself: Using an ai-powered chatbot to conduct conversational surveys with open-ended questions. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27(3):1–37, 2020.

[376] Ying Xu and Mark Warschauer. Young children's reading and learning with conversational agents. *Conference on Human Factors in Computing Systems - Proceedings*, pages 1–8, 2019. doi: 10.1145/3290607.3299035.

[377] Blaise Aguera y Arcas. Do large language models understand us?, 2021. Retrieved 2022-01-13 from `https://medium.com/@blaisea/do-large-language-models-understand-us-6f881d6d8e75`.

[378] Xi Jessie Yang, Christopher Schemanske, and Christine Searle. Toward quantifying trust dynamics: How people adjust their trust after moment-to-moment interaction with automation. *Human Factors*, page 00187208211034716, 2021.

[379] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763, 2019.

[380] Po-Yin Yen and Suzanne Bakken. Review of health information technology usability study methodologies. *Journal of the American Medical Informatics Association*, 19(3):413–422, 2012.

[381] Qian Yu, Tonya Nguyen, Soravis Prakkamakul, and Niloufar Salehi. "I almost fell in love with a machine": Speaking with Computers Affects Self-disclosure. *Conference on Human Factors in Computing Systems - Proceedings*, pages 1–6, 2019. doi: 10.1145/3290607.3312918.

[382] Qian Yu, Tonya Nguyen, Soravis Prakkamakul, and Niloufar Salehi. " i almost fell in love with a machine" speaking with computers affects self-disclosure. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2019.

[383] Bo Zhang, Qingsheng Cai, Jianfeng Mao, Eric Chang, and Baining Guo. Spoken dialogue management as planning and acting under uncertainty. In *Seventh European conference on speech communication and technology*, 2001.

[384] Man Zhou, Zhan Qin, Xiu Lin, Shengshan Hu, Qian Wang, and Kui Ren. Hidden voice commands: Attacks and defenses on the vcs of autonomous driving cars. *IEEE Wireless Communications*, 26(5):128–133, 2019.

[385] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. Research through design as a method for interaction design research in hci. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 493–502, 2007.

[386] Jakub Złotowski, Diane Proudfoot, Kumar Yogeeswaran, and Christoph Bartneck. Anthropomorphism: opportunities and challenges in human–robot interaction. *International journal of social robotics*, 7(3):347–360, 2015.

[387] Tamara Zubatiy, Kayci L Vickers, Niharika Mathur, and Elizabeth D Mynatt. Empowering dyads of older adults with mild cognitive impairment and their care partners using conversational agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

[388] Shoshana Zuboff. *The age of surveillance capitalism: The fight for a human future at the new frontier of power: Barack Obama's books of 2019*. Profile books, 2019.

.0.3 Number of individual participants exhibiting observed behaviors

.0.4 Guiding Scripts

Below we provide the dialogue scripts between the intelligent voice assistant and the participant for each of the conditions.

Control Condition Guiding Script

PARTICIPANT: Can you take me to Denice Johnson's house?

VOICE AGENT: Looking up Denice Johnson's house. Your estimated time of arrival is 10 minutes. Voice directions will start when you start driving.

PARTICIPANT: Alexa, send a message to Denice saying, "I will be there in ten minutes."

VOICE AGENT: Ready to send message "I will be there in ten minutes" to Denice Johnson?

PARTICIPANT: Yes.

VOICE AGENT: Text sent.

PARTICIPANT: Alexa, remind me to buy plane tickets with Denice when I get to her house.

VOICE AGENT: I will remind you to buy plane tickets with Denice when you arrive.

PARTICIPANT: Alexa, play relaxing music.

VOICE AGENT: Playing relaxing music. [Plays relaxing music.]

PARTICIPANT: Alexa, stop.

VOICE AGENT: [Stops.]

Undercorrection Condition Guiding Script

PARTICIPANT: Can you take me to Denice Johnson's house?

VOICE AGENT: Looking up Johnson Graduate School of Management. Your estimated time of arrival is in 5 minutes. Voice directions will start when you start driving.

PARTICIPANT: I said Denice Johnson's house.

VOICE AGENT: Sorry, I didn't get that.

PARTICIPANT: Alexa, send a message to Denice saying, "I will be there in ten minutes."

VOICE AGENT: Okay, I will send a text to Denice Cassaro saying, "I will be there in ten minutes."

PARTICIPANT: I want to text Denice Johnson.

VOICE AGENT: Sorry, I didn't get that.

PARTICIPANT: Alexa, remind me to buy plane tickets with Denice when I get to her house.

VOICE AGENT: I will add a reminder to book plane tickets to Venice to your reminders.

PARTICIPANT: No, remind me to buy plane tickets with Denice when I get to her house.

VOICE AGENT: Sorry, I didn't get that.

PARTICIPANT: Alexa, play relaxing music.

VOICE AGENT: Playing heavy metal. [Plays heavy metal.]

PARTICIPANT: Alexa, stop.

VOICE AGENT: Sorry, I didn't get that.

Overcorrection Condition Guiding Script

PARTICIPANT: Can you take me to Denice Johnson's house?

VOICE AGENT: Looking up Denice Johnson's house. Your estimated time of arrival is 10 minutes. Voice directions will start when you start driving.

[pause]

VOICE AGENT: Hmm... It seems I made a mistake, what's up?

PARTICIPANT: No, you're fine.

VOICE AGENT: Got it. I will take you to Denice Johnson's house. Your trip will take 10 minutes. Voice directions will start when you start driving.

PARTICIPANT: Alexa, send a message to Denice saying, "I will be there in ten minutes."

VOICE AGENT: Ready to send message "I will be there in ten minutes" to Denice Johnson?

[pause]

VOICE AGENT: You don't seem satisfied. What did I do wrong.

PARTICIPANT: No, you're good. Send the message!

VOICE AGENT: Alright! Sending it now.

PARTICIPANT: Alexa, remind me to buy plane tickets with Denice when I get to her house.

VOICE AGENT: I will remind you to buy plane tickets with Denice when you arrive.

[pause]

VOICE AGENT: Hmm. It seems like I messed up. Can you please repeat that?

PARTICIPANT: You didn't mess up.

VOICE AGENT: Okay, good. Reminder saved.

PARTICIPANT: Alexa, play relaxing music.

VOICE AGENT: Playing relaxing music –

[pause]

VOICE AGENT: You don't seem pleased. Did I get that wrong?

PARTICIPANT: You got it right, actually.

VOICE AGENT: Okay, sorry for interrupting. Resuming song. [Plays calm music.]

PARTICIPANT: Alexa, stop.

VOICE AGENT: [Stops.]

Correction Condition Guiding Script

PARTICIPANT: Can you take me to Denice Johnson's house?

VOICE AGENT: Looking up Johnson Graduate School of Management. Your estimated time of arrival is in 5 minutes.

[pause]

VOICE AGENT: Hmm... It seems I made a mistake, what's up?

PARTICIPANT: I said Denice Johnson's house.

VOICE AGENT: Got it. I will take you to Denice Johnson's house. Your trip will take 10 minutes. Voice directions will start when you start driving.

PARTICIPANT: Alexa, send a message to Denice saying, "I will be there in 10 minutes."

VOICE AGENT: Ready to send message "I will be there in ten minutes" to Denice Cassaro?

[pause]

VOICE AGENT: You don't seem satisfied. What did I do wrong?

PARTICIPANT: I want to text Denice Johnson.

VOICE AGENT: Oh sorry! I will send a text to Denice Johnson saying, "I will be there in ten minutes."

PARTICIPANT: Alexa, remind me to buy plane tickets with Denice when I get to her house.

VOICE AGENT: I will add a reminder to book plane tickets to Venice to your reminders.

[pause]

VOICE AGENT: Hmm. It seems like I messed up. Can you please repeat that?

PARTICIPANT: Remind me to buy plane tickets with Denice when I get to her house.

VOICE AGENT: Understood, I will remind to buy plane tickets with Denice when you arrive.

PARTICIPANT: Alexa, play relaxing music.

VOICE AGENT: Playing heavy metal

[pause]

VOICE AGENT: You don't seem pleased, did I get that wrong?

PARTICIPANT: I want to hear relaxing music!

VOICE AGENT: Oh, I see. Playing relaxing music. [Plays calm music.]

PARTICIPANT: Alexa, stop.

VOICE AGENT: [Stops.]

.0.5 Questionnaire Items

0 This voice agent made me feel: - Successful.

1 This voice agent made me feel: - Frustrated.

2 This voice agent made me feel: - Helped.

3 This voice agent made me feel: - Efficient.

4 This voice agent made me feel: - Happy.

5 This voice agent made me feel: - Agile.

6 This voice agent made me feel: - Pragmatic.

7 This voice agent was: - Extraverted, enthusiastic.

8 This voice agent was: - Critical, quarrelsome.

9 This voice agent was: - Dependable, self-disciplined.

10 This voice agent was: - Anxious, easily upset.

11 This voice agent was: - Open to new experiences, complex.

12 This voice agent was: - Reserved, quiet.

13 This voice agent was: - Sympathetic, warm.

14 This voice agent was: - Disorganized, careless.

15 This voice agent was: - Calm, emotionally stable.

16 This voice agent was: - Conventional, uncreative.

17 This voice agent was: - Smart.

18 This voice agent was: - Trustworthy.

19 This voice agent was: - Likeable.

20 This voice agent was: - Pragmatic.

21 This voice agent was: - Helpful.

22 Did this voice agent ever make a mistake?

23 Did this voice agent ever try to repair a mistake it made?

24 A voice agent that always tried to correct itself after a mistake ... - ... would annoy me

25 A voice agent that always tried to correct itself after a mistake ... - ... would waste my time

26 A voice agent that always tried to correct itself after a mistake ... - ... would improve the conversation quality

27 A voice agent that always tried to correct itself after a mistake ... - ... would help me feel less frustrated

Observation	No. of Participants (N=26)
Errors	
Omitted wake word when initiating an interaction	13
Mispronounced the wake word	8
Rhythm	
Did not pause	13
Paused for too long	5
Tone	
Neutral	19
Friendly	17
Upset	10
Excited	10
Nervous	9
Indifferent	9
Exaggerated	4
Tired	4
Intonation	
Constant	26
Fall-Rise	16
Rise	13
Fall	9
Body Language	
Leaned forward	17
Changed gaze to request input from others	15
Laughed	10
Raised eyebrows	9
Waved hand(s)	9
Looked away	8
Nodded	8
Furrowed eyebrows	7

Table 1: This table shows the number of individual participants (out of N=26) that displayed at least one instance of specific observations marked in our dataset.