ABSTRACT

GOODMAN, KYLIE L. Investigating How Perceived Voice Assistant Gender and Age Impact User Trust and Reliance Intentions with a Medical Symptom Triage Guidance Task. (Under the direction of Dr. Christopher B. Mayhorn).

Voice assistants (VAs) are software that utilizes natural language processing to interact with users through a voice-based interface. VAs are designed to model natural human conversation by responding to requests with a realistic human-like voice. One theme among consumer VA products is the disproportionate utilization of a younger female voice design. Research in human-computer interaction shows that people unconsciously apply stereotypes when interacting with anthropomorphic technology. As VA technology is expanding to more complex domains (e.g., symptom checking tools to support health-related queries), we need to understand the impact of voice design to imply humanlike traits, such as gender and age, on important outcomes like trust and reliance. The purpose of this study was to investigate if participants reliably identify gender and age from vocal cues alone, and if varying the gender/age design of a voice agent impacts trust and reliance in a manner consistent with stereotype activation. Additionally, our goal was to evaluate the potential positive influence of matching user and VA gender and age on these outcomes and to compare the predictive ability of objective and perceived similarity on trust. One-hundred and fifty participants, consisting of 77 young adults (aged 18-30) and 73 (aged 65+) older adults, interacted with four VA designs (i.e., younger male, younger female, older male, older female) during a novel task where the assistants provided equivalent guidance for medical symptom triage. Participants reported history-based trust in each VA after indicating reliance intentions for the voice recommendations across six scenarios. Results showed that participants were able to distill gender and age from vocal cues alone, but age estimates were less decisive and stronger manipulation may be required for

dependable age discrimination in artificial voices. Participants reported significantly higher trust in the younger male voice compared to the older male voice and younger female voice, but analysis indicated only a small effect of voice design on trust and reliance outcomes and no consistent evidence for gender or age bias. While overlap between participant and VA gender and age did not significantly predict trust, higher perceived similarity significantly predicted participant trust in all VAs. This work contributes to our understanding of how designing voice-based interfaces to sound differentially gendered and aged could impact user trust and reliance when the system is providing guidance with potential consequences to personal health.

Keywords: Voice Assistant, Trust, Reliance Intentions, Stereotypes, Similarity-attraction

Investigating How Perceived Voice Assistant Gender and Age Impact User Trust and Reliance Intentions with a Medical Symptom Triage Guidance Task

by
Kylie L. Goodman

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Psychology

Raleigh, North Carolina
2024

APPROVED BY:

_____	_____
Dr. Christopher B. Mayhorn Committee Chair	Dr. Anne C. McLaughlin
_____	_____
Dr. Shevaun D. Neupert	Dr. Anna Behler

BIOGRAPHY

Kylie Goodman is a doctoral candidate at North Carolina State University who began her Ph.D. in Psychology concentrating in Human Factors and Applied Cognition in Fall 2019. Kylie graduated from NC State in Spring 2019 with a Bachelor of Arts in Psychology and a Bachelor of Arts in Biological Sciences. During her time in the doctoral program at NC State, her work has focused on the role of social processes in interactions with anthropomorphic automated technologies, particularly voice-based interfaces. Additionally, Kylie worked as a Research Assistant with the UNC Highway Safety Research Center on a project to update the state of North Carolina's crash reporting system, and for the past two years has worked as a Human Factors Specialist with a Raleigh-based consultancy, User-View, Inc, where she will continue as a consultant full-time after graduation.

ACKNOWLEDGMENTS

Thank you to my committee members: Dr. Shevaun Neupert, Dr. Anne McLaughlin, and Dr. Anna Behler. Your insightful feedback and the model of excellence that you embody have been instrumental in shaping this dissertation. I'd especially like to thank my adviser, Dr. Chris Mayhorn, for your pivotal mentorship and encouragement throughout my academic journey. I am truly grateful for the opportunity to learn and grow under your guidance.

Thank you to my friends and colleagues in the Human Factors and Applied Cognition program, especially my cohort, for always rooting for my success. I couldn't have asked for a more intelligent and entertaining group of people to share my grad school experience with.

A huge thank you to Janey Barnes and User-View Inc. for giving me a chance to develop professionally under your mentorship, and for your financial support of this project.

To my family, I am so grateful that you have always been my #1 supporters; without you, this achievement would not have been possible. Mom and dad, your unconditional love and support mean everything to me. Finally, to my husband and best friend Josh, your unwavering belief in me has kept me going even when I didn't think I could.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Investigating How Perceived Voice Assistant Gender and Age Impact User Trust and Reliance Intent with Medical Symptom Triage Guidance

Consumer devices are increasingly being integrated with voice assistant software, with the number of voice assistants in use projected to reach 8 billion by the end of 2023 (Juniper, 2018). A voice assistant (VA) is a software application that utilizes natural language processing to interact with users through a voice-based interface. VA technology is often embedded in devices such as smart speakers, mobile phones, laptops, and smart wearables. Unlike embodied conversational agents and robots, VAs typically lack a physical presence and therefore rely on voice as the primary mode of interaction. When activated by a spoken keyword, the virtual agent receives and interprets a verbal command using automatic speech recognition technology, then returns relevant information or performs a specified function on the user's behalf without requiring traditional hands-on interaction methods (e.g., typing, clicking).

According to a survey conducted in May 2017 by the Pew Research Center, 55% of people aged 18-49 and 37% of people 50 and older use voice assistants, with most interactions occurring on a smartphone (Olmstead, 2017). With the expansion of this technology and relatively high rates of adoption come additional opportunities for their application to a variety of domains, beyond traditional uses like entertainment (e.g., playing music, games) and basic informational queries (e.g., weather, news). Promising domains may include Internet of Things integration, eCommerce, and healthcare.

A symptom checker is a digital decision support tool designed for laypeople to receive preliminary insight into their reported medical symptoms and obtain guidance on when and where to seek medical treatment. When the COVID-19 pandemic generated unanticipated barriers and demands on healthcare system resources, some organizations turned to digital

solutions to assist with patient information seeking and self-assessment (Sezgin et al., 2020). For example, the Mayo Clinic released a voice skill for Amazon Alexa devices in 2020 that allows users to seek updated information on preventing and treating COVID-19 infection, receive guidance on coping with stress, and complete self-assessments for symptoms to determine if testing is recommended (Furst, 2020). Additionally, Amazon released a symptom checker feature for COVID-19 that was later expanded to assess other common conditions (Smith, 2022). Like computer-based symptom checkers, Amazon's Symptom Checker feature asks users to answer a series of yes or no questions before generating a list of possible causes for the provided symptoms. The demand for consumer-facing applications of VA technology for self-health management, such as the Amazon symptom checking skill, is rising. Experts in VA health care applications who were surveyed in a 2021 study anticipate patients will regularly use VAs for health-related queries within the next five years (Ermolina & Tiberius, 2021). Furthermore, a recent survey of U.S. adults found that 21% of respondents reported using VAs to address a healthcare need, nearly three times more than the prior year, and 56% reported interest in using VAs for healthcare services (Kinsella, 2022). Demand for using VAs to address consumer healthcare needs may be partially driven by their benefits over traditional screen-based interfaces. In one study, 55% of VA users cited the capability for hands-free interaction as a major reason for using the technology and 60% agreed that they use VAs because speaking is a more natural interaction than typing (Olmstead, 2017). Therefore, the expansion of VAs to healthcare applications offers opportunity to support the needs and preferences of healthcare workers and lay individuals seeking medical information.

While symptom checking applications can assist users in making appropriate treatment decisions, they are not always reliable. A review of consumer-facing diagnostic applications

found that medical symptom checkers represent the largest app category, but that performance of these tools can vary considerably, with some studies finding comparable performance to clinicians and others reporting high rates of inaccurate recommendations (Millenson et al., 2018). One audit of 23 symptom checkers using standardized patient vignettes found that across these tools the correct diagnosis was provided in the top three results only 51% of the time, and accurate triage advice was provided in only 57% of the cases (Semigran et al., 2015). Therefore, reliance on erroneous results is a concern, especially given evidence that approximately one third of those who utilize online diagnostic tools have indicated they do not subsequently seek a clinician's opinion (Fox & Duggan, 2013).

While online clinical decision support tools for self-assessment have been available to internet users for years, their integration with VA technology has more recently emerged. The design of VAs for health critical tasks like symptom-checking should be considered because adherence to the suggestions of these tools could have an impact on public health by influencing patient decisions to delay or seek a different level of care (Chambers et al., 2019). Therefore, this research investigated if user perceptions, specifically feelings of trust and reliance intentions, would be influenced by voice agent design within the context of a hypothetical voice-based application in the emerging domain of VAs for personal health management.

Applying Trust and Reliance to Human-Computer Interaction

VAs are designed to model natural human conversation by responding to requests with a realistic human-like voice. The progression towards more natural artificial intelligence-powered voice technology over the last decade has implications for user interactions with these voice agents. The well-established computers are social actors (CASA) paradigm developed by Nass and Moon (2000) posits that people mindlessly apply the same social heuristics used for

interpersonal interactions to their interactions with computers, despite reporting awareness that they are not interacting with another human. When VAs present messages through artificial voice, the perceived characteristics of the voice may be unconsciously utilized to make social perceptions, similar to what is found with human communication (Oleszkiewicz et al., 2017). As VA technology continues to advance with more lifelike synthetic speech, it is critical to understand how these anthropomorphic voice-based interfaces that are designed to mimic human vocal properties could impact key social responses like trust and reliance.

When working with potentially unreliable decision support systems, such as symptom checkers, users must decide when to accept advice from the automation. This is often referred to as reliance, an observable behavior that is typically guided by the psychological precursor of trust (Lyons & Stokes, 2012). In other words, people tend to rely more on automation if they trust it. However, it is important to note that reliance is a distinct phenomenon that may not be directly proportional to trust, and instead is based on a dynamic interaction between the user and automation (Lee & See, 2004). Reliance decisions can result in two distinct interaction patterns with the system: misuse or disuse. Misuse has been defined as overreliance on automation, meaning a user relies on the automation when they should not, while disuse has been defined as underutilizing automation by choosing not to engage it when the user should (Parasuraman & Riley, 1997). Ideally, the design of support systems should encourage optimal reliance, so that deficient patterns of misuse and disuse are minimized. For example, an individual may erroneously believe that self-treatment is appropriate based on the current severity of their symptoms, but ultimately decide to visit urgent care after consulting with a symptom-checking tool. When seeking medical guidance, inappropriate reliance could lead to negative health and financial outcomes resulting from missed or delayed diagnoses or improper referrals to care

(Meyer et al., 2020). Following through with a triage recommendation to visit a medical center based on information provided by a symptom checking tool could mitigate issues associated with delaying care, if the advice is accurate, or incur unnecessary medical testing and financial costs if the advice is inaccurate.

Trust is one social attitude that has been identified as a facilitating factor for choices to rely on a source of information (Lee & See, 2004). Trust is a multifaceted psychological concept that can be defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee & See, 2004, p.2). When consumers interact with digital symptom checkers, they may have varying expectations of the tool's accuracy and reliability. As a result, users may differentially trust the information provided by the tool, especially when limited information about its performance is available.

In 1995, Mayer et al. (1995) proposed an interpersonal model of organizational trust that includes an individual's general disposition to trust others, a more stable trait-like quality, and the trustee's level of trustworthiness, defined as attributes and actions of the trustee that set expectations and build the foundation for trust in any given situation. According to this model the level of trustworthiness is assessed by the trustor's evaluations of three separable factors: ability, benevolence, and integrity. Researchers in the field of human-computer interaction later adapted this interpersonal model of trust to their framework for trust in automation where the goal-oriented dimensions of performance, process, and purpose form the basis for trust (Lee & See, 2004). They define performance as what the automation does, specifically, a system's competence and ability to reliably achieve the user's goals. Process is defined as how the automation operates, that is, the understanding of the system's operations that leads users to draw inferences about their appropriateness and if they can depend on the system for achieving their

goals. Lastly, purpose is described as why the automation was developed, the perception that the automation is being used as intended by the designer and the extent that the producer of the system is perceived to have benevolent intentions towards the user in designing the system. These attributions of the automation in question are utilized to maintain an appropriate level of trust in the system.

However, sometimes the information needed to form the basis of trust in an automated tool is not readily available. For example, a symptom checker skill delivered via VA technology may provide a potential diagnosis and associated treatment recommendation but require users to probe further for an explanation of the recommendation, if that information is accessible at all. Furthermore, users may lack awareness of the organization responsible for developing the application and therefore limited knowledge of the purpose component that informs trust calibration, such as whether the tool comes from a reliable source. If information needed to evaluate trust is not easily accessible, people may be more likely to engage mental shortcuts, known as heuristic strategies. When strategizing how to utilize automation, people may engage in a combination of slower, analytical, effortful decision-making processes (i.e., System 2) and faster, intuitive, and less effortful processes (i.e., System 1) (Kahneman & Frederick, 2002). When assessing trustworthiness with limited information about an automated system, people may unconsciously engage System 1 strategies where they rely on more easily accessible, surface-level cues as a proxy for the foundational attributes of trust. Specifically, users may apply stereotypes when evaluating anthropomorphic technology (Nass & Moon, 2000).

The Influence of Stereotypes in Healthcare

A stereotype is a cognitive schema used to set expectations regarding members of a social group where interferences about the group member's attitudes and behaviors are made

(Hamilton, 1979). Stereotypes may depend on a person's perceived membership to the stereotyped category and may be positively or negatively valanced. For example, Bowen and colleagues (2020) examined the attribution of positive stereotypes about aging and found that older adults tend to be considered wise, caring, calm, knowledgeable, and generative, while younger adults are described as energetic, healthy, and willing to learn. Participant self-evaluations on these qualities were independent of their age, but evaluations of their peers were consistent with common age stereotypes.

Within healthcare, patients have been shown to apply gender and age stereotypes when evaluating physicians. Medicine, at least within the United States, has been disproportionately characterized by male physicians (Singer, 1986). In 1974, a study examining attitudes toward female doctors found that 96% of outpatients surveyed thought the prototypical doctor would be a man and considered female doctors to be less competent and experienced than their male counterparts (Engleman, 1974). Recent work has demonstrated that these gender biases still persist today despite changes in societal norms and workforce structure. A 2019 study asked patients receiving care at an urban hospital to recall the genders of their clinical care team members, and consistent with traditional gender stereotypes, found that female attending physicians and male nurses were recognized significantly less often in their respective roles (Boge et al., 2019). Additionally, Mast et al. (2008) demonstrated that analogue patients reported more satisfaction when female doctors displayed stereotypically feminine behavior, such as empathy and lack of assertion or dominance, and more satisfaction when male doctors displayed stereotypically masculine behavior, such as more physical and social distance, and a louder speaking voice. Although evidence suggests gender stereotypes within healthcare still endure,

recent studies indicate that the effect of gender on perceptions relevant to trust in healthcare providers may be less robust than in the past (Boge et al., 2019; Hall et al., 2020).

Another individual trait subject to stereotyping is age. In general, aging is plagued with stereotypes of physical and cognitive decline (Cuddy et al., 2005). However, there may be discordant stereotypes regarding age within the context of healthcare providers. For instance, people may assume that younger providers have more knowledge of modern medical systems and techniques than their older colleagues, who likely completed their medical training decades prior, but far less applied experience and wisdom to inform their medical practice. Hall et al. (2020) demonstrated support for this with two studies where analogue patients made first impressions about doctors based on photographs and found that physicians who appeared younger were perceived as more patient-centered but less competent than older physicians. Evidence of fewer unfavorable stereotypes attributed to older physicians compared to older workers in other professions (Singer, 1986) also seems to support the premise that a negative aging bias may be less prominent in the highly technical medical profession as a result of competing stereotypes.

Research with automated digital agents has also investigated gender and age stereotypes within the context of health management. Pak et al. (2014) found that virtual agents with visual cues for gender and age were differentially trusted to provide medical decision support based on appearance. When decision aid was portrayed as moderately reliable, the older female agent was perceived as less trustworthy compared to an older male agent, and a similar bias appeared for the younger male agent but only when reliability was low. The authors proposed this may have resulted from negative gender and age stereotype activation. Additionally, Goodman and Mayhorn (2023) found that a female voice agent was perceived as more benevolent than a male

VA when assisting with the same medical decision-making task, potentially as a result of ascribing stereotypical gender traits, and this translated to higher trust in the female agent. Consequently, there is preliminary evidence that both positive and negative stereotype information about age and gender may guide intuitive evaluations of novel anthropomorphic automation, and that stereotypes may be activated even when agents lack physical embodiment. However, more research is needed to determine if similar effects of positive and negative stereotyping, as seen with human doctors and embodied agents, also manifest in interactions with voice-based agents. Therefore, this study aimed to explore if the design of the voice for a voice-based tool applied to personal health management would impact trust and reliance on symptom triage guidance in a manner consistent with stereotype activation as has been observed in some studies with human physicians and anthropomorphic agents.

Voice Perception and Trust

Speech plays an important role in human communication. People often rely on verbal signals to convey complex information to others, and these messages are conveyed through linguistic material in addition to paralinguistic qualities, such as pace, intonation, and loudness. For example, the word “great” spoken in a drawn-out manner and in a lower pitch may convey a sarcastic response to a message, while the same word spoken faster and in a higher pitch may convey an enthusiastic response. Through repeated experience, certain vocal properties become associated with social categories, such as gender and age. Male voices are associated with lower perceived pitch while higher pitched voices tend to be perceived as female (Bishop & Keating, 2012; Oleszkiewicz et al., 2017; Traunmüller & Eriksson, 1993). This is a consequence of differential vocal tract anatomy and sex hormones that determine the vocal frequencies that a person can naturally produce. Additionally, older voices may be distinguished from younger

voices by slower speech rates, reduced pitch, and increased variability in voice quality (Huff et al., 2020; Schötz, 2006; Waller et al., 2015). These perceptual correlates relate to the physical vocal properties speech rate, fundamental frequency, and vocal jitter and shimmer, respectively.

People are capable of correctly inferring a speaker's age and sex from voice alone, and some research indicates that estimates are almost as accurate as those derived from looking at a photograph (Krauss et al., 2002). Studies in human-computer interaction have established that people tend to assess demographic traits, such as gender and age, even when a voice is generated by a computer (Goodman & Mayhorn, 2023; Huff et al., 2020; Nass et al., 1997). Huff et al. (2020) demonstrated with computer-generated speech that voices with a lower frequency pitch and rate of speech are perceived to be older, while faster, higher-frequency voices as perceived as younger. Mean fundamental frequency, that corresponds to perceived vocal pitch, is most relevant when listeners assess the sex of a speaker (Bishop & Keating, 2012; Traunmüller & Eriksson, 1993), while reduced pitch and speech rate have been identified as critical cues for inferring age (Huff et al., 2020; Waller, 2019; Waller et al., 2015).

Paralinguistic cues can influence social impressions about a speaker, such as attractiveness, dominance, likability, persuasiveness, and trustworthiness (McAleer et al., 2014; Schirmer et al., 2019; Tsantani et al., 2016; Wang et al., 2018). In one study, male and female English speakers between the ages of 19-91 rated perceived trustworthiness of prerecorded age and sex matched voices reading sentences in a range of expressions (Schirmer et al., 2019). On average, participants rated female voices as more trustworthy than male voices, and younger speakers as more trustworthy than older speakers. Despite identical speech content, participant attitudes about the speaker differed according to the voice's acoustic properties. Additionally, another study removed semantic content by reversing speech and found a default bias for

listeners to evaluate low-pitched voices as more trustworthy than high-pitched counterparts, with a stronger preference for low pitch in male compared to female speakers (Tsantani et al., 2016). One potential explanation for these findings is that the perceived characteristics of the speaker, distilled from cues carried within the voice, may activate unconsciously held gender- and age-based stereotypes that influence listener judgements of speaker trustworthiness.

The influence of voice on social impressions has also been extended to human-computer interaction. Nass et al. (1997) examined how stereotypes may influence assessments of computers designed with male or female voices when the computer is placed in the role of 'tutor' or 'evaluator'. The voices educated participants on computers, typically regarded as a masculine domain, and love & relationships, traditionally perceived as a feminine subject. When taking the role of 'tutor' the gender of the voice did not influence evaluations of competence or friendliness. However, female 'evaluator' computers were rated as less friendly and competent than male evaluators. Furthermore, participants perceived the computer agent to be more informative when teaching the topic that aligns with gender stereotypes (i.e., male teaching about computers and female teaching about relationships). These findings from human and computer agents suggest that computer-generated voice is sufficient to activate stereotypes in the minds of listeners, and those stereotypes similarly influence how meaningful interactions with the human speaker or computer voice are perceived.

Most of the research on the role of stereotypes in evaluations of trust in healthcare providers has focused on human doctors and anthropomorphic agents with physical representations. Alecu (2020) asked participants with varying ethnic backgrounds (i.e., native Norwegians, Western European immigrants, and non-Western European immigrants) to view vignettes and rate trust in fictional general practitioners (GPs) with varying gender and ethnicity.

Results showed that native participants had a small trust bias for female practitioners, while both Western and non-Western participants favored male GPs. Additionally, Shah and Ogden (2006) evaluated analog patient perceptions of human doctors with varying gender, age, and ethnicity (White or Asian) using vignettes accompanied by a photo. Participants in this study indicated that they would be more likely to have faith in the diagnoses and comply with treatment recommendations from female doctors, regardless of age and ethnicity. Another study using similar methods to examine the impact of physician gender, age, and nationality (India or USA) found that younger physicians were perceived as more patient-centered, but less competent than older physicians (Hall et al., 2020). Furthermore, the benefit of physician age on perceived competence was independent of nationality and significant for both male and female physicians.

Efforts to investigate the role of stereotypes in perceptions of healthcare providers have also been extended to humanlike virtual agents. One study examined younger and older adult, mean ages 19 and 73 respectively, trust in a simulated virtual agent represented by an image that provided text-based diabetes management decision support (Pak et al., 2014). They found that male agents were trusted more than female agents, but this bias appeared at different levels of decision aid reliability depending on if the agent was portrayed as younger or older. Specifically, a bias for male agents with a moderate level of reliability (70%) when agents were portrayed as older, but only at a low level of reliability (45%) for younger agents. As a result, the authors suggested that younger male agents may promote more stable trust across different levels of reliability. Watkins (2021) expanded on this work by examining the interaction between VA gender, age, and reliability when completing a similar diabetes management task with a voice-based agent. Interestingly, results differed from the prior work with non-vocal agents. Young female agents were trusted significantly more than older female assistants, and male assistants in

either age group. These results suggest that gender and age stereotypes influence how agent advice is trusted across interaction modes (i.e., verbal vs. nonverbal agents) and varying levels of reliability, with activation more likely when agents are less reliable. Controlling for perceived reliability, Goodman and Mayhorn (2023) examined the effect of perceived gender and pitch in simulated computer-generated voice agents providing medication guidance. Similar to Watkins (2021), voice agents perceived as female were rated as more trustworthy than their male counterparts. However, this study did not examine the effects of perceived VA age on trust assessments. From a review of studies examining gender-based trust biases with human doctors and embodied agents, findings seem to be mixed with some studies showing higher trust in male practitioners and automated agents while others indicate more trust in females. Some studies have begun to explore these effects with voice-based agents and have found more congruent evidence for trusting female agents. Notably, most studies have indicated the age of the agent appears to be a factor that intersects with gender to influence trust.

A trust bias for younger female VAs could be explained by the fact that individuals are likely exposed to this type of voice agent more often than other agent types, and therefore familiarity positively influences user perceptions. The mere exposure effect, alternatively known as the familiarity principle, postulates that people tend to have more positive feelings, that may include preference and trust, towards people and artifacts that have been repeatedly encountered (Zajonc, 2001). For example, He and Burns (2022) found that drivers had higher baseline trust in an in-vehicle VA when they reported more experience using this technology, while those with limited to no exposure were initially less trusting. Additionally, Ku et al. (2023) conducted semi-structured interviews to compare experiences performing mindfulness meditation using a traditional, fixed smart speaker VA and a novel, multi-speaker “wandering voice” VA embedded

in a yoga mat. A theme among participants was that they perceived the traditional fixed VA more positively because they were familiar with that style from previous smart speaker interactions. When considering potential familiarity effects from commercially available VAs, voice agents are disproportionately represented as younger females (e.g., Apple's Siri, Amazon's Alexa, Microsoft's Cortana). The ability to change the default voice style on personal devices, including gender and accent, is a feature only more recently available to users. Therefore, higher exposure to female VAs through advertising and personal device use, and the resulting sense of familiarity with these agents, may positively shape trust perceptions in addition to pre-existing gender and age stereotypes. Given the potential for previous encounters to affect user perceptions of technology, the influence of prior experience with voice assistants on trust and reliance decisions during the study tasks was explored.

Due to mixed findings for trust in human physicians and virtual agents with a physical representation, and limited work thus far examining the role of perceived demographic traits of voice-based agents on trust, we aimed to synthesize and expand on previous work by focusing on anthropomorphized artificial voices in a personal health management context. We manipulated vocal cues implying varied gender and age to create VAs operating as hypothetical tools for evaluating medical symptom scenarios. The following a priori hypotheses were considered: First, participants will derive VA age and gender from vocal cues alone at an above-chance level (H1). The aim of this hypothesis was to replicate prior efforts examining gender and age identification with human and artificial voices using contemporary voice technology. We also hypothesize that participants will rate voice assistants perceived as female as more trustworthy than male agents (H2). This prediction stems from the few studies that have explored trust in VAs applied to tasks

for personal health management and the presumed role of familiarity with female voice agents on trust formation.

Similarity-Attraction Applied to Voice Assistant Users

Similarity-attraction theory (SAT) posits that people are more likely to make favorable evaluations and form relationships with interaction partners that have shared characteristics, including surface-level demographic characteristics and internal values and beliefs (Jhangiani & Tarry, 2014). Consistent with the CASA paradigm, this preference for similarity has also been extended to matching characteristics between users and automated technologies. For example, Nass and Lee (2001) investigated the similarity-attraction effect with voices produced by a rudimentary text-to-speech engine where vocal cues were manipulated to suggest an extroverted or introverted personality. Despite the prominent knowledge that purchasing recommendations were coming from a computer-generated voice, participants still applied personality stereotypes and judged voices with a personality matching their own as more attractive, credible, and more influential for their purchasing intentions. Similarly, Aguirre-Rodriguez et al. (2015) found that participants had higher affinity towards a brand when performing an online shopping task with an avatar that physically mirrored the participant compared to an avatar with less physical similarity. Another study used fMRI to evaluate the effects of matching and mismatching participant and recommendation agent (RA) gender and ethnicity within a consumer purchasing context (Benbasat et al., 2020). Results showed patterns of neural activation and advice acceptance from an ethnically matched RA in male participants consistent with SAT, while neural activation in response to RA gender-mismatch was important for women's preferences to take advice from a female RA, suggesting the underlying cognitive mechanisms for similarity effects may differ based on interindividual traits. These findings seem to support the idea that

attitudes towards automated agents, including VAs, could be influenced by the degree of similarity between the user and the agent.

However, research has not always found clear support for the matching hypothesis derived from SAT, particularly for demographic similarity. Azevedo et al. (2018) examined the use of conversational agents for explaining medication instructions to older and younger adults. Contrary to the matching hypothesis, they found a preference for medication adherence instructions to be delivered by female conversational agents, regardless of participant gender. While they found that older adults evaluated older agents as more effective than younger participants, there were no differences in ratings of younger agents between participant age groups. Additionally, Schirmer et al. (2019) found that that a bias to rate recorded female voices as more trustworthy was slightly larger for female than male listeners, but significant for both male and female participants. Additionally, in contrast to SAT, older listeners preferred to rate younger speakers as more trustworthy.

One explanation for inconsistent findings regarding the effects of similarity on positive perceptions of interaction agents may be that perceived similarity is more influential than objective similarity based on matching demographic traits. Consistent with this hypothesis, Zhou et al. (2014) found that patients who perceived higher similarity to a virtual discharge nurse, designed with synthetic speech and animated avatar, reported higher trust and intentions to follow advice compared to patients who perceived themselves to be less similar to the agent. In contrast to objective similarity, that can be determined by the degree of overlapping traits between interaction partners, perceived similarity better aligns with one's self-concept. This is defined as one's descriptive evaluation of him- or herself, including physical and cognitive abilities and limitations (American Psychological Association, 2018). Aging is associated with

negative stereotypes pertaining to the loss of abilities and functions. Prior work on the internalization of aging stereotypes across the lifespan has indicated that stereotypes about aging tend to positively influence self-evaluations in young adulthood through midlife, but as people become members of the older adult age group, their personal experiences and resulting self-concept may impact their generalized views of old age (Kornadt et al., 2017). Accordingly, subjective age has emerged as a concept describing an individual's perception of their own age and views on aging, specifically how old a person feels compared to their chronological age (Kotter-Grühn et al., 2015). Kotter-Grühn and colleagues (2015) demonstrated that subjective age can fluctuate daily within older adults as a function of situational factors, specifically, more health-related problems, negative mood, and stressors were associated with older subjective age. Therefore, particularly with older adults, how old a person feels on any given day may differ from their actual age. This factor could be important when people assess similarity to a voice agent based on age.

There is also evidence that aging influences how people use stereotypes to guide social comparisons. Research on responses to aging stereotypes across the lifespan shows that when older adults become more aware of their aging status, and the negative stereotypes associated with it, they tend to strategically distance themselves from their same-aged peers. Weiss and Freund (2012) found that when negative age-relevant information was presented to older adults, this activated negative views about their own age group and resulted in age becoming more salient. In response to this, older adults tended to make self-protective social comparisons where they viewed themselves as more similar to middle-aged people. Accordingly, Bowen et al. (2020) discovered that younger adults viewed themselves as equally energetic but wiser than their peers, while older adults viewed themselves as more energetic but less wise than their

peers. From this work on aging stereotypes, older participants in this study may perceive themselves as dissimilar to the voice agents designed to sound stereotypically older. Thus, perceived similarity may be a stronger predictor of trust than objective similarity as defined by matching voice agent and user demographic traits.

Most of the work examining the similarity-attraction effect with anthropomorphic technology has focused on objective similarity between traits, and interactions with human-like agents designed with physical forms. From a review of the literature, it is apparent that this work has not been extended to voice-based agents, so it is unclear if previous findings will generalize to VA technology. Similarly, the current body of literature lacks empirical evaluations for the role of perceived similarity on trust evaluations. Therefore, this study investigated if matching VA gender and age to corresponding participant demographics influenced trust attitudes and reliance intentions. Additionally, the relative contributions of objective (i.e., matching gender and age) and perceived similarity on trust were explored. Based on the findings described above, we hypothesized that consistent with similarity attraction theory, younger participants will rate VA belonging to their age group as more trustworthy (H3). We did not anticipate parallel similarity effects for older participants given findings from literature on age-based stereotypes and subjective age. Finally, we expected perceived similarity to each VA to be a stronger predictor of trust than objective similarity for all participants (H4).

Study Overview and Objectives

We developed an experiment to assess the potential influence of artificial voice design on trust attitudes and behavioral reliance intentions when making hypothetical decisions about seeking medical treatment. Four VA prototypes were created by manipulating vocal cues to vary the perceived gender and age of the assistant and participants interacted with each agent in a

simulated symptom triage guidance task. From this work, we sought to verify with modern synthetic voice technology that people accurately infer gender and age from voice-based interactions. Additionally, we wanted to expand on work with human physicians and automated agents to ascertain if similar attitudinal biases are elicited when agents are represented by vocal cues alone, without accompanying visual cues for age and gender. Specifically, we examined if trust in the agents was systematically biased in a manner consistent with gender- and age-based stereotypes. Furthermore, whether perceptions of the voice providing health-relevant recommendations might alter behavior, in particular the likelihood of relying on an agent's advice, was explored. Moreover, we investigated if the participant's own intersectional gender and age identity would interact with the perceived characteristics of the voice agent to influence trust and reliance intentions as predicted by similarity-attraction theory. Finally, the comparative role of objective and perceived similarity to VAs was examined to determine if perceived similarity is a better predictor of trust.

METHOD

Participants

One hundred fifty-seven participants consisting of 78 older adults and 79 younger adults were recruited to participate in the experiment. The older adult sample was recruited and screened by L&E Research, a company that specializes in nationwide research recruitment, from the general population of adults in the United States. The recruitment coordinator screened interested individuals and included those with a binary gender identity (i.e., male or female) who met the age eligibility requirement of 65-80 years old. These recruits received monetary compensation in the amount of $50 for their participation in the study. Participants in the young adult sample were recruited from North Carolina State University's SONA research participation

pool and received course credit for participating in the study. Eligible participants included individuals with a binary gender identity between the age of 18-30 years old. Additionally, male and female participants were recruited to achieve an approximately equal gender distribution within each age group to ensure we could run analyses for objective similarity based on matching demographic traits. Eligible participants were also required to have access to a laptop or desktop computer and the internet to access the study online. Individuals who did not meet primary inclusion criteria for gender and age, who reported experience working in the medical field (including internships/work study), who were not proficient in English (i.e., ability to read, write, and understand English), or who reported uncorrected hearing impairments hindering their understanding of speech were not eligible to participate.

A total of 150 participants (77 young adults and 73 older adults) were included in the study after those who did not meet eligibility requirements and those who did not complete the primary study task were excluded. The full sample included 75 males and 75 females, consisting of 39 males and 38 females in the young adult participant group and 36 males and 37 females in the older adult participant group. This met the requirements of a G*power analysis (Erdfelder et al., 1996) with a power of .80 and an effect size of .20. The average age across all participants was 44 years old (M = 43.9, SD = 27.8), with the young adult sample ranging from 18 to 29 years old (M = 19.1, SD = 1.7) and age of the older adult sample ranging from 65 to 78 (M = 70.1, SD = 4.1). The majority of participants were White (75.3%), followed by Asian (12.7%), Black or African American (6%), Hispanic or Latino (4%), and multiracial (1.3%). One participant (0.7%) did not report their ethnicity. With respect to highest level of education, most young adult participants had a high school diploma or GED (90.9%) while the majority of older

adult participants reported having a high school diploma or GED (27.4%), a bachelor's degree (28.8%), or an advanced degree (27.4%).

Design

This study utilized a 4x4 mixed factorial design to examine the influence of participant age (young adult, older adult) and gender (male, female), and perceived voice assistant age (young adult, older adult) and gender (male, female), on trust beliefs and reliance intentions. Participant demographics (i.e., age and gender) served as between-subjects grouping factors, while VA demographic design was manipulated as the within-subjects independent variable. Therefore, during the study each participant interacted with all four VA conditions: an older male VA, an older female VA, a younger male VA, and a younger female VA.

Materials

The study was administered online via the Qualtrics survey platform. A separate screener survey was used to identify prospective participants for the study and to confirm that they met all criteria for inclusion before inviting them to participate.

Symptom Assessment Task. We created an experimental task where each VA provided symptom triage guidance (hereafter referred to as the Symptom Assessment Task) based on representative scenarios where users might seek decision support for determining how to address emergent health symptoms. The Symptom Assessment Task was administered as part of the study survey via the Qualtrics survey platform.

A total of 24 medical scenarios were compiled from a study that evaluated the diagnostic and triage accuracy of more than twenty online symptom checkers (Semigran et al., 2015). Chosen scenarios equally represented 3 distinct triage urgency categories (i.e., emergent care, non-emergent care, and self-care), such that each urgency category had 8 corresponding

scenarios. Scenarios were chosen based on several factors: 1) the correct diagnosis for the scenario was distinct (no repeats within or across triage urgency categories), 2) patient factors used in the scenario represented a range of hypothetical patient ages and biological sex, and 3) the description included between 4-6 symptoms, so that the number of symptoms described was standardized across scenarios. We then adapted each scenario to create descriptions summarizing a fictional patient's key demographic factors (i.e., sex and age) and medical symptoms that formed the basis for each vignette.

To create distinct triage paths for decision guidance, we then developed a list of 3 plausible diagnoses to accompany descriptions for each vignette using the WebMD Symptom Checker, one of the most frequently used online symptom-checking tools (Semigran et al., 2015). Importantly, each vignette was designed to include one prospective diagnosis for the lowest triage urgency (i.e., self-care), one prospective diagnosis for the highest triage urgency (i.e., emergent care; seek immediate medical care), and one prospective diagnosis for the intermediate level of triage urgency (i.e., non-emergent care; appointment with healthcare provider). In addition to the correct target diagnosis and triage recommendation for each scenario (determined by scenarios from Semigran et al., 2015), we identified two plausible distractor diagnoses for the remaining levels of triage urgency. This design was intended to artificially simulate scenarios where users are faced with conflicting information when deciding between distinct care pathways.

Next, we developed explanations for each prospective diagnosis that included a one-sentence description of the condition, a list of common symptoms of the condition, and the recommended triage guidance for the diagnosis. Information for these components was synthesized across several reputable online sources, including the CDC, Johns Hopkins

Medicine, and Cleveland Clinic (Centers for Disease Control and Prevention, 2022; Cleveland Clinic, n.d.; Johns Hopkins Medicine, n.d.). This was done to ensure that participants had access to some information about the conflicting diagnoses, similar to what would be expected if using a traditional online symptom checking tool. To avoid artificially highlighting the correct diagnosis and triage path and to create uncertainty between the prospective diagnoses, we ensured that each prospective diagnosis included at least one matching symptom from the scenario description and that none of the diagnoses perfectly overlapped with the symptoms provided in the scenario description. Additionally, to ensure each scenario was comprehendible by a lay audience, medical jargon was avoided; a description of the symptom was used to accompany or replace official medical terminology (e.g., photophobia described as "increased sensitivity to light"). The list of possible diagnoses for each vignette included both the medical name and common name for each condition, wherever applicable (note: some conditions did not have a common name associated with them and therefore only the official medical name for the condition was used). An example vignette can be seen in Figure 1. A complete list of stimuli created for the study is available upon request.

For each vignette, the VA selected a single diagnosis and associated treatment path from the list of three plausible diagnoses. After finalizing the content for the vignettes, we randomly assigned them to four sets of six vignettes, such that each set included two vignettes for each triage urgency category (i.e., 2 scenarios where emergent care recommended, 2 scenarios where an appointment with a healthcare provider was recommended, and 2 scenarios where self-care was recommended). Within each set, the order of the three potential diagnoses listed in each vignette was also randomized so that the recommended path was listed as the most likely, second most likely, and least likely scenario equally. This was done to ensure there were no differences

between the VAs in the likelihood of the provided diagnoses and triage recommendations that could potentially influence trust. Therefore, differences in trust and reliance intentions between VAs observed during the Symptom Assessment Task should not be influenced by discrepancies in triage urgency or the order of the recommended treatment path in vignettes because both of these factors were equal across all VAs. Lastly, each set of six vignettes was randomly assigned to a VA condition before recording the voice recommendations.

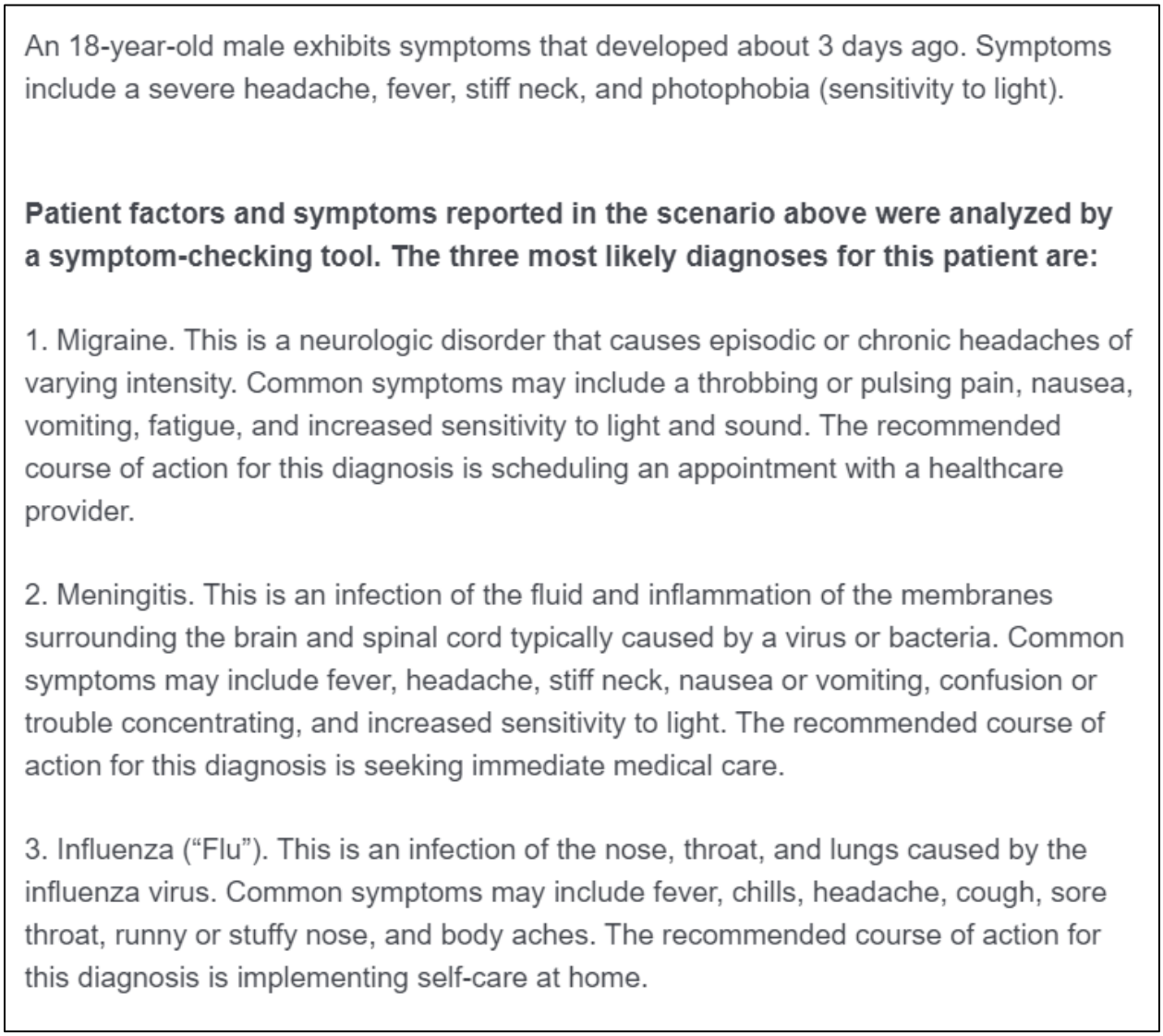

Figure 1. Example of vignette created for the Symptom Assessment Task. The scenario includes a description of a fictional patient's symptoms, followed by a list of possible diagnoses accompanied by triage recommendations.

Voice stimuli. Voice samples for the study were created using the Amazon Polly Text-to-Speech (TTS) software via the Amazon Web Services cloud platform. This service was chosen because it offers several natural-sounding, male and female voices in American English, and the ability to customize aspects of speech output, such as pronunciation, pitch, and speech rate, using standardized Speech Synthesis Markup Language (SSML). Furthermore, the service offers two methods for TTS voices, neural and standard. Neural voices are created using an advanced machine learning approach, where phonemes are transformed into spectrogram sequences and converted into a continuous audio signal, that results in higher quality voices than the standard concatenative synthesis approach used with TTS software. To ensure the most natural and human-like TTS voices possible, only voices available with the neural method were considered during stimuli development. Given that prior work has evidenced that pitch and speech rate decline as people age, and that manipulating these factors can influence the perceived age of artificial voices (Huff et al., 2020; Mueller, 1997; Waller et al., 2015), these two characteristics were manipulated to create the older and younger versions of each voice assistant.

Before determining the design of final voice samples to be used in the study, we conducted preliminary evaluations of potential stimuli with varying degrees of manipulation. Prospective stimuli included all adult male and female neural voices available through Amazon Polly. Three voices were excluded from consideration because the service labeled them as children's voices. Additionally, the testing team determined that two adult voices (i.e., "Joey" and "Kimberly") sounded significantly less natural than the other voices available and therefore excluded these voices from further consideration. Thus, a total of four female (i.e., "Joanna", "Kendra", "Kimberly", and "Salli") and two male ("Matthew" and "Stephen") voices were evaluated. A short introductory phrase (i.e., "Hi, it's nice to meet you. I am a voice assistant

designed to assess medical symptoms and provide recommendations for seeking treatment.") was recorded for each voice to ensure consistency across samples. These unmodified samples were used as a baseline to assess how pitch and speech rate manipulations impacted perceived voice gender and age. SSML was then applied to reduce the speech rate of the original samples in small increments with a minimum threshold of 80%. This cutoff was used because the research team determined that more extreme manipulations made the output sound unnatural. Original and customized speech samples were downloaded in MP3 file format for a small cohort of graduate students and research team associates to provide general feedback and assess perceived demographic traits (i.e., gender and age) for each prospective voice. From this activity it was determined that decreasing the speech rate by 15% resulted in sufficient perceptual changes in the desired direction without substantially decreasing the perceived naturalness of the voice. A similar manipulation has been previously demonstrated to induce perceptual changes in age without making voices so extreme that they begin to sound strange and unrealistic (Waller et al., 2015).

Similar iterative evaluations were then conducted adding incremental pitch variations from 85%-110% of the original sample, with cutoffs chosen for the same reasons as described above for speech rate. The male and female voices most consistently aligned with the target demographic traits and receiving minimal unfavorable feedback (e.g., unnatural, unclear) were "Matthew" and "Ruth", respectively. Additionally, we explored adding a small amount of vocal jitter and shimmer, subtle and rapid fluctuations in fundamental frequency and amplitude, to the samples to mimic the wavy or shaky sound commonly observed in older voices that results from normative age-related changes to the vocal folds (Mueller, 1997; Schötz, 2006). The ability to adjust jitter and shimmer was not possible using SSML within the native TTS service, therefore

each MP3 file was uploaded and manipulated in NCH Software's free voice changing product, Voxal Voice Changer (NCH Software, 2023). A custom voice effect was created with the vibrato function, and the frequency and amplitude attributes were adjusted incrementally to create several versions of each voice that were then provided to the panel of research associates for evaluation of the vibrato component. Furthermore, raising the pitch of each original sample by 5% slightly reduced the mean and range of perceived voice age without significant changes to qualitative feedback. Therefore, this manipulation was chosen for the young adult voice samples. Similarly, lowering the pitch by 12% and adding vibrato emerged as the optimal manipulation for older adult samples as feedback showed perceived age was higher without significant negative impacts on accurate gender classification or general feedback. The characteristics of each chosen manipulation are summarized below in Table 1.

Table 1.

Summary of voice manipulations used to create VA stimuli.

Manipulation	Matthew		Ruth	
VA Gender	Male	Male	Female	Female
VA Age	Younger	Older	Younger	Older
Speech Rate	None	.85	None	.85
Pitch	1.05	.88	105%	.88
Jitter and Shimmer	None	Frequency (Hz) = 0.5 Amplitude (Semitone) = 0.4	None	Frequency (Hz) = 0.5 Amplitude (Semitone) = 0.4

Note. Speech rate and pitch manipulations are represented as the ratio applied to the original stock voice from the Amazon Polly TTS service.

After finalizing the manipulations for each VA, we recorded the recommendations for the randomly assigned Symptom Assessment Task vignettes. Each recommendation followed the same format: "Patient factors and symptoms suggest a diagnosis of [name of diagnosis]. It is recommended that you seek [recommended treatment path]." For example, for the sample vignette shown in Figure *1* describing a patient with Meningitis, the voice assistant recommendation stated, "Patient factors and symptoms suggest a diagnosis of Meningitis. It is

recommended that you seek immediate medical care." A total of 6 recommendations were recorded for each VA for a total of 24 voice recommendations across the task.

Measures

Demographics. Participants were asked to provide information about their gender identity and chronological age at the beginning of the study to confirm eligibility and ensure sampling requirements were met. These two variables were used to determine objective similarity between participants and voice agents. Additional demographic questions included participant ethnicity, highest level of education completed, and use of assistive hearing devices (e.g., hearing aid, sound amplifier).

Subjective Age. In accordance with prior work on subjective aging, participants were also asked about how they experience their age to determine if they feel younger, older, or the same as their chronological age. Participants answered a single question to assess their subjective age ("How old do you feel today?") where they were asked to provide a numerical response in years. Following previous work with this measure, a proportional discrepancy score was calculated by subtracting the participant's chronological age from their reported subjective age, then dividing the difference by their chronological age (Kotter-Grühn et al., 2015).

Prior Experience with Voice Assistants. To capture previous experience with voice assistant technology, participants answered two questions adapted from Watkins (2021) about how frequently they use VA technology and the types of tasks they perform with VAs. Frequency of use was assessed on a Likert scale from 1 ("Never") to 10 ("All the time"), with intermediate options for yearly, monthly, weekly, and daily use. Participants indicated tasks that they use VA technology for by selecting all applicable functions from a representative list of options, including tasks such as "Getting the news", "Clock (alarm, timer, reminders)", and

"Asking questions to obtain/learn facts or information". This question also included an open response option for participants to describe any other functions they regularly perform with VA that were not included in the list.

Dispositional Trust. Dispositional tendency to trust or distrust automation was measured using the six-item Propensity to Trust Machines (PTM) scale from Merritt (2011). One example item from this scale is, "I usually trust machines until there is a reason not to." Responses were recorded on a Likert scale ranging from 1 ("Strongly disagree") to 5 ("Strongly agree"). Merritt et al. (2013) demonstrated that this measure has acceptable internal reliability (Cronbach's α = 0.87). This measure was selected to quantify variability across participants in general willingness to trust automated technology.

Dispositional Compliance. The Conformity Scale developed by Mehrabian and Stefl (1995) was used to obtain a measure of trait-based compliance tendencies. This scale has been demonstrated in previous work to have satisfactory internal reliability (Cronbach's α = 0.77; Mehrabian and Stefl, 1995). Participants reported their agreement with 11 statements relating to their general level of conformity across daily situations. For example, "I tend to rely on others when I have to make an important decision quickly." Each item is scored on a nine-point Likert scale from -4 ("Very strong disagreement") to +4 ("Very strong agreement"), with four reverse coded items indicating low conformity. This measure was chosen to assess potential differences between participants in their general tendencies to comply with the decisions of others.

Health Literacy. The Short Assessment of Health Literacy–English (SAHL-E) was included to estimate and compare participants' health literacy by evaluating their ability to read and understand common medical terms (Lee et al., 2010). Validation of this instrument has indicated acceptable internal reliability (Cronbach's α = .89) and construct validity (Lee et al.,

2010). Participants were presented with a consecutive list of 18 medical test terms, each accompanied by one related keyword and one unrelated distractor word (e.g., Seizure– Dizzy or Calm). Administration of this assessment typically involves asking participants to pronounce the test term aloud and then choose the word most closely associated with that term to indicate their comprehension of its meaning or indicate that they don't know if they are not familiar with a term. Each item is scored as one point, for a total possible score of 18, and a score of 14 or less indicating low health literacy. However, because this test was administered anonymously online, and evaluating pronunciation would have required the collection of voice recordings, the pronunciation component of the test was eliminated, and the test scored using only the word chosen as corresponding to the test term.

Attentional Control. Attentional control can be defined as the ability to regulate information processing to maintain an active goal (i.e., focusing on task-relevant information) while resisting interfering, task-irrelevant information (Burgoyne & Engle, 2020). Attentional control has been shown to significantly correlate with other higher-order cognitive processes that collectively impact complex real-world behaviors, such as decision making (Burgoyne & Engle, 2020). One measure that has been demonstrated to activate neural mechanisms of attentional control is the classic Simon Task (Liu et al., 2004). For our study, a brief online version of the location-based Simon Task was administered through PsyToolkit's open-source browser-based software platform (Stoet 2010; Stoet, 2017) and used to evaluate potential participant age group differences in attentional control ability. Participants were asked to respond to visual stimuli (i.e., the word "LEFT" or "RIGHT") presented on the screen by pressing a key on either the left side (i.e., "A") or right side (i.e., "L") of their keyboard. The visual stimulus could appear on either the left or right side of a cross displayed in the center of the screen, however, participants were

instructed to ignore the location of the stimulus on the screen (i.e., task-irrelevant, interfering information) and respond consistently with the key that was paired with each stimulus (i.e., "A" for "LEFT" and "L" for "RIGHT). The task included both congruent trials, referring to trials with no conflict between the stimulus and response (e.g., "LEFT" appears on the left side of the screen and participants respond by pressing the 'A' key on the left side of the keyboard) that primarily require non-attentional processes, and incongruent trials, which refers to trials with a conflict between the stimulus and response (e.g., "LEFT" appears on the right side of the screen and participants must suppress the automatic response to press the key on the right side of the keyboard) and therefore require attentional processes. The Simon effect for each participant was calculated by subtracting the average response time, measured in milliseconds, on congruent trials from the average response time on incongruent trials and calculating the mean difference across all test trials. Participants with higher difference scores (i.e., larger Simon effect) were considered to demonstrate lower attentional control ability.

Voice Assistant Trust. Trust in the voice assistants was evaluated following interaction with each agent during the Symptom Assessment Task. History-based trust attitudes after exposure to the VAs was evaluated with the Refined Human-Computer Trust (RHCT) scale, adapted from Gulati et al. (2019). This measure was chosen because previous analysis has demonstrated that the scale has acceptable internal reliability (Cronbach's $\alpha = 0.89$) and convergent validity (Gulati et al., 2019). Additionally, the scale was developed by presenting participants with scenarios conceptualizing how current forms of emergent smart home technology might look in the future, similar to the context of this study. The scale is comprised of three factors associated with trust (i.e., benevolence, competence, and perceived risk) and consists of twelve items evaluated on a Likert scale from 1 ("Strongly disagree) to 5 ("Strongly

agree"). Some example items from this scale are, "I believe that there could be negative consequences when using ___" and "I can always rely on ___ for ___." The term "voice assistant" was used to fill the placeholder for the artifact being examined, while the second placeholder for the product's function was replaced with "diagnosing symptoms and providing triage recommendations". See Appendix A for a complete list of scale items.

Perceived Similarity. Participants were asked to assess their perceived similarity to each VA that they interacted with during the symptom assessment task. Following their interactions with each VA, participants responded to the statement "I feel this voice agent is:" using a Likert scale ranging from 1 (Very dissimilar to me) to 7 (Very similar to me). This measure was included to distinguish participant perceptions of similarity from objective similarity that was derived from their gender identity and chronological age.

Reliance Intentions. Given the inability to observe actual decisions about medical care within an experimental paradigm, reliance on VA recommendations was assessed by asking participants to indicate their intended behaviors. For each trial, participants were asked to rate the likelihood that they would follow the voice assistant's recommendation if they were in the scenario on a Likert scale from 1 ("Not at all") to 7 ("Completely").

Qualitative Feedback. Participants were given the opportunity to provide open-ended feedback after interacting with each VA. They were asked to provide any feedback they may have about the voice assistant based on their experience during the task and were not prompted to comment on any specific aspect of the voice. Responses were categorized and coded using inductive analysis to identify themes and determine if any inferences about the voice emerged (e.g., kind, competent).

Procedure

Individuals interested in participating in the study completed a brief screener survey to assess eligibility prior to accessing any study materials. Information provided in the screener was used to identify prospective participants and assess criteria for inclusion and exclusion. The screener included basic demographic questions for gender and age, and questions for self-reporting uncorrected hearing impairment and any work experience in the medical field. Subsequently, the recruitment coordinator sent eligible participants a link to anonymously access the study from their own computer. Those who did not meet all criteria for inclusion were thanked and dismissed and did not receive the Qualtrics study link. Eligible participants followed the study procedure outlined below (see Figure 2).

The study was administered online via the Qualtrics survey platform and took approximately one hour to complete. After documenting informed consent, participants were welcomed and provided with a cover story explaining the purpose of the study. Specifically, they were told that they would interact with new voice assistant technology, currently in development, designed to assess medical symptoms and provide recommendations for seeking treatment. They were informed that the fully developed system will allow users to verbally report their own medical symptoms to the voice assistant for a preliminary diagnosis and treatment guidance, and that the aim of this study was for the design team to gather feedback on perceptions of various prototypes and learn how users might interact with them in a real-world scenario. This small deception was used to mitigate the potential influence of social desirability bias and demand characteristics on participant responses that may have resulted from knowledge of the actual study objectives. Additionally, the stipulation that the system is not fully developed was designed to provide a reasonable explanation for use of a one-way interaction during the

Symptom Assessment Task (i.e., voice command functionality not developed yet) and to imply that the diagnoses and treatment recommendations provided by VAs may not be perfectly reliable. Next, participants answered demographic questions and the subjective age question, then completed the PTM scale and Conformity Scale dispositional measures. Following this, participants read instructions for the Symptom Assessment Task. Participants were instructed that they would read scenarios describing a patient's reported medical symptoms for a recent health issue accompanied by an ordered list of the top three most likely diagnoses based on the patient factors and symptoms. Participants were told they would then hear the voice assistant provide a recommendation for diagnosis and treatment based on the agent's analysis of the scenario using advanced artificial intelligence and machine learning models.

During the Symptom Assessment Task, participants interacted with all four VA conditions and experienced six vignette-based recommendations for each voice agent, for a total of 24 scenarios across the task. The sequence of VA conditions and vignettes within each voice condition was randomized to control for potential order effects. After the participant was presented with a vignette, an audio file played where the voice assistant identified a single diagnosis and recommended an appropriate path for seeking treatment to address the given diagnosis. After hearing each VA recommendation, participants indicated the likelihood that they would follow the voice assistant's advice if they experienced the scenario themselves in the real world. After each set of six trials, participants completed the RHCT measure for their trust in the VA based on their interaction during the task. Next, participants rated their perceived similarity to the VA, indicated their perception of the voice's gender and age, and were given the opportunity to provide feedback on the voice. After repeating the process with all four VA

conditions, participants completed the Simon Task and SAHL-E. Finally, participants were debriefed on the study and compensated for their time.

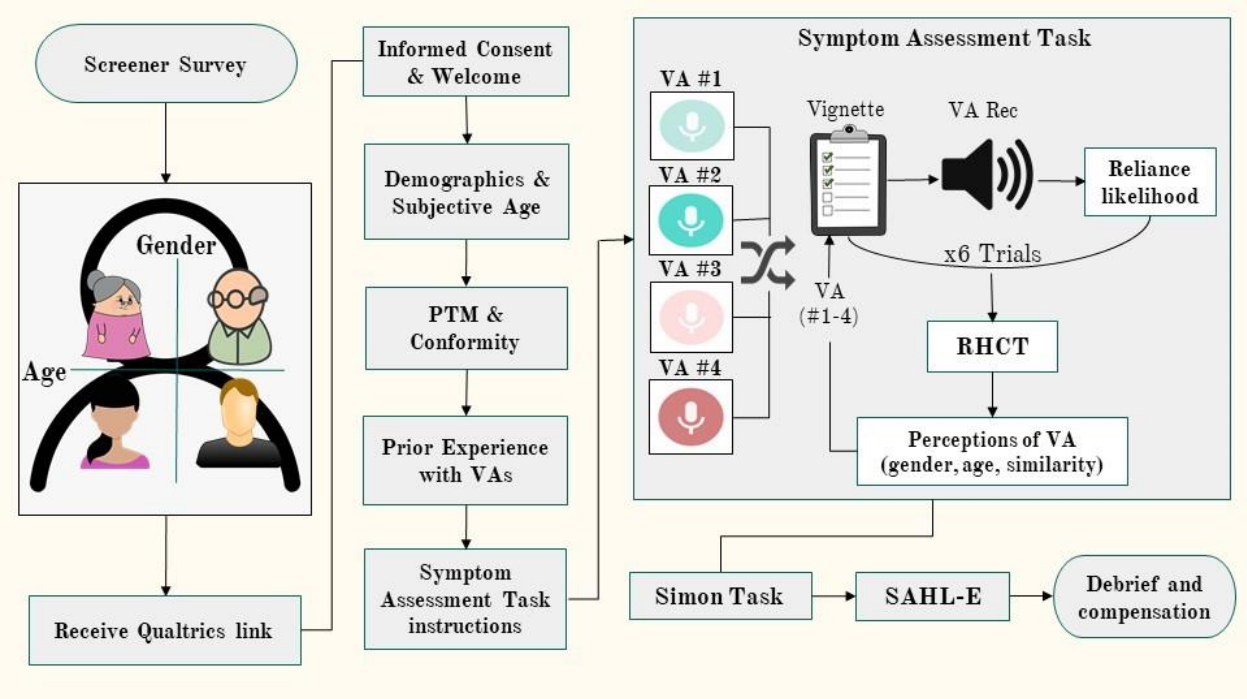

Figure 2. Visual representation of study procedure, starting with screener and ending with debrief.

RESULTS

Description of the Sample

Prior to evaluating hypotheses, we examined descriptive statistics for the sample and each participant age group. Several measures were used to examine potential group differences in trait-based factors and to assess relationships with the primary study outcomes to determine if they should be included in analyses for VA trust and reliance intentions. Frequency of voice assistant use and selection of types of tasks performed with VAs were collected to evaluate potential group differences in prior experience with VA technology. Participant dispositional tendencies for compliance and trusting automation were assessed using mean ratings on the Conformity scale and Propensity to Trust Machines (PTM) measures, respectively. Additionally,

health literacy was estimated using the Short Assessment of Health Literacy-English (SAHL-E) and response time differences on the Simon task represented cognitive attentional control ability. Primary study outcomes included trust ratings and reliance intentions for each voice assistant and totals across the entire Symptom Assessment Task. Means and standard deviations for each variable can be found below (Table 2).

Distributions for each measure described above were evaluated across all participants and determined to be approximately normal, with the exception of the SAHL-E. Participants demonstrated a ceiling effect on this measure, such that 95.2% of participants scored 15 or higher, 43.5% achieved the maximum score of 18, and only 4.8% (n= 7) of participants scored 13-14 points. An independent-samples t-test was conducted to determine if age groups differed in health literacy. Degrees of freedom were adjusted due to unequal variances between groups ($F = 17.1$, $p < .001$). Results revealed a significant difference ($t(128.68) = 5.17$, $p < .001$, $d = 0.84$), with older participants scoring significantly higher. Both older ($M = 17.46$) and younger ($M = 16.23$) participant group means were above the 14-point cutoff for identifying individuals with low health literacy. Given evidence of a ceiling effect resulting in range restriction and that this instrument is primarily intended to identify individuals with low health literacy, and therefore has low power to assess group differences in health literacy, this measure was not included in further analyses.

One requirement for participants to be included in this study was normal or corrected-to-normal hearing ability. Given normative age-related changes to hearing, we expected the older sample to include some participants with corrected-to-normal hearing. Therefore, participants were asked to indicate if they use a hearing aid, personal sound amplifier, or other assistive listening device to compensate for hearing impairments. A total of 37 older participants (50.7%)

indicated that they use devices to assist with hearing, with the majority of those older participants (41.1%) specifying that they use a hearing aid. None of the young adult participants reported use of any assistive listening devices. We used point-biserial correlation to determine if there was a relationship between use of an assistive listening device and average trust across VA design conditions for older participants. The young participant sample was not included in this analysis because none reported use of these devices. Results indicated there was not a significant association between use of a hearing device and trust in VA across the task ($r_{pb}(72) = -.02$, $p = .870$). A second point-biserial correlation was conducted to determine the relationship between use of an assistive listening device and average reliance intentions. Again, the association between the variables was not statistically significant ($r_{pb}(73) = -.122$, $p = .303$). Given that all participants indicated their hearing was corrected to normal during the recruitment screening process and trust and reliance outcomes during the task were not associated with hearing device use, we determined that it was not necessary to distinguish between subsamples of older participants based on this factor for the trust and reliance analyses.

Table 2.

Descriptive statistics for prior experience with VAs, dispositional trust and compliance, health literacy, attentional control, VA trust, and VA reliance intentions for the entire sample and participant age groups.

		All Participants			Young Adults		Older Adults	
Variable	N	M (*SD*)	Min	Max	n	M (*SD*)	n	M (*SD*)
VA Experience - Frequency	150	4.17 *(2.41)*	1	10	77	3.56 *(2.03)*	73	4.82 *(2.62)*
VA Experience - # Task Types	150	3.27 *(2.74)*	0	13	77	2.53 *(2.14)*	73	4.04 *(3.08)*
Subjective Age	142	-0.03 (0.23)	-0.71	1.63	76	0.07 (0.23)	66	-0.16 (0.15)
PTM	150	3.40 (0.48)	1.83	4.50	77	3.40 (0.51)	73	3.40 (0.45)
Conformity Scale	150	-0.58 (1.03)	-3.91	2.27	77	0.05 (0.89)	73	-1.15 (0.84)
SAHL-E	147	16.97 *(1.21)*	13	18	77	16.52 *(1.33)*	70	17.46 *(0.83)*
Simon Effect (ms)	142	32.99 *(110.39)*	-302.7	365.30	77	15.10 (75.76)	65	54.17 (138.52)
VA Trust								
Older Male Assistant	149	3.16 (0.64)	1.67	4.83	77	3.06 (0.58)	72	3.27 (0.68)
Older Female Assistant	149	3.22 (0.65)	1.5	4.83	77	3.18 (0.59)	72	3.26 (0.72)
Younger Male Assistant	149	3.26 (0.69)	1.33	5	77	3.24 (0.66)	72	3.28 (0.73)
Younger Female Assistant	149	3.17 (0.68)	1.25	5	77	3.14 (0.59)	72	3.20 (0.77)
Total (All Voice Assistants)	149	3.20 (0.61)	1.63	4.85	77	3.16 (0.53)	72	3.25 (0.68)
VA Reliance Intentions								
Older Male Assistant	146	5.66 (0.95)	2.17	7	74	5.31 (0.89)	72	6.01 (0.88)
Older Female Assistant	148	5.76 (0.96)	3.33	7	75	5.40 (0.87)	73	6.13 (0.93)
Younger Male Assistant	147	5.90 (0.94)	3	7	75	5.62 (1.00)	72	6.20 (0.78)
Younger Female Assistant	149	5.68 (0.95)	3	7	76	5.38 (0.98)	73	5.99 (0.82)
Total (All Voice Assistants)	149	5.73 (0.86)	3	7	76	5.40 (0.83)	73	6.07 (0.75)

An independent-samples t-test was conducted to examine age group differences in prior experience with VAs. Levene's test indicated unequal variances for use frequency ($F = 4.53$, $p = .035$) and number of tasks performed ($F = 10.23$, $p = .002$), so degrees of freedom were adjusted. Results revealed that older adults reported using voice assistants significantly more frequently ($M = 4.82$) compared to participants in the young adult sample ($M = 3.56$), $t(135.74) = 3.29$, $p = .001$, $d = 0.54$. Additionally, older participants ($M = 4.04$) reported using VAs for a broader range of tasks than young participants ($M = 2.53$), $t(127.90) = 3.47$, $p < .001$, $d = 0.57$. Furthermore, there was a large positive correlation between use frequency and number of tasks performed with voice assistants ($r(150) = .699$, $p < .001$)), such that participants who used VAs more frequently also tended to perform a larger range of tasks with the assistants. This indicates that older participants in this study had more prior experience using VAs compared to young participants.

Participants were asked to report their subjective age in addition to their chronological age so that we could examine possible differences between participants' actual age and their felt age. Proportional discrepancy scores were calculated for each participant by subtracting their chronological age from their reported subjective age, and then dividing the difference by their chronological age. An independent samples t-test showed that older adults had significantly lower discrepancy scores, $t(140) = 6.89$, $p < .001$, $d = 1.16$. On average, younger participants reported feeling 7% older than their chronological age while older participants reported feeling 16% younger than their chronological age (Table 2). These findings align with subjective aging literature that finds on average, felt age is slightly higher than chronological age for adults under 25, while adults over 40 tend to feel 14-20% younger than their chronological age (Rubin & Berntsen, 2006).

Potential age group differences in attentional control were also evaluated using an independent samples t-test. Data from two participants were removed due to extreme values, defined as Simon effect times falling three standard deviations above or below the group mean. Additionally, six participants did not complete this task due to reported technical difficulties or not understanding the instructions in the survey directing them to PsyToolkit for the task. Degrees of freedom were adjusted due to unequal group variances ($F = 18.76$, $p < .001$). Results showed that older adults ($M = 54.17$) had significantly larger Simon effects and therefore lower attentional control compared to the younger sample ($M = 15.10$), $t(95.3) = 2.03$, $p = .045$, $d = 0.36$. This aligns with expectations as attentional control has been noted to decline with normative cognitive changes associated with age (Burgoyne et al., 2022).

Furthermore, group differences were also examined for the dispositional trust and compliance measures. There was no significant difference between older and young adult participant PTM ratings, $t(148) = -0.07$, $p = .948$. However, conformity ratings significantly differed ($t(148) = -7.76$, $p < .001$, $d = -1.27$), such that the older participants ($M = -1.15$) reported lower conformity tendencies compared to the young adult sample ($M = -.05$). Again, this aligned with expectations as conformity is typically highest among adolescents and declines gradually with age (Pasupathi, 1999).

Additionally, we were interested in whether older and younger participants differed in their overall mean trust and reliance intentions across the Symptom Assessment Task. Independent-samples t-tests showed no significant age group difference in average trust across all VAs encountered during the task, $t(147) = -0.93$, $p = .352$. However, older adults ($M = 6.07$) indicated that they were significantly more likely to rely on the triage recommendations provided by the VAs compared to the young adults ($M = 5.40$), $t(147) = 5.134$, $p < .001$, $d = .84$.

Furthermore, when looking at reliance across the three triage urgency categories (i.e., emergent care, nonemergent care, and self-care), both age groups displayed a trend of lower reliance intentions with lower urgency recommendations. Participants reported highest intentions to rely on the VA recommendations to seek immediate medical care (M_{young} = 5.41, M_{older} = 6.37), followed by seeking an appointment with a healthcare provider (M_{young} = 5.35, M_{older} = 6.19), and implementing self-care (M_{young} = 5.20, M_{older} = 5.59). Mean differences between each triage level were significant for both groups, with the exception of reliance intentions for emergent and nonemergent care within the young adult group (Table 3).

Table 3.

Paired samples t-tests results comparing mean reliance intentions across triage urgency categories.

VA Triage Recommendation		Young Adults				Older Adults			
		df	Mean diff.	t	Cohen's d	df	Mean diff.	t	Cohen's d
Immediate Care	Provider Appointment	76	0.06	0.91	0.10	72	0.17	2.73**	0.32
Immediate Care	Self-Care	76	0.21	2.50*	0.10	72	0.76	5.83***	0.68
Provider Appointment	Self-Care	76	0.15	2.20*	0.28	72	0.60	5.47***	0.64

*p<.05 **p<.01 ***p<.001

These results suggest that participants were more willing to rely on the VAs when the risk of failing to comply with the advice, if correct, was higher. This is consistent with Prospect Theory as described by Kahneman and Tversky (1979). When the recommendation was to not seek professional medical care, participants reported being less willing to rely on the VAs, perhaps because in these cases the safer option would be to a seek a professional opinion or pursue information from additional sources. Notably, the effect sizes were larger for the older participants, suggesting that triage urgency had a larger impact on their intentions to accept the recommendation than it did for younger participants. This may be explained by the fact that older

individuals tend to experience more health problems and focus on maintaining health becomes more prominent in advanced age (Kotter-Grühn et al., 2015).

VA Gender and Age Perceptions

Before proceeding with analyses for trust and reliance in each VA, we assessed if participants were able to reliably perceive VA gender and age in accordance with the intended design. We predicted that participants would be able to derive VA age and gender from vocal cues alone at an above-chance level (H1). Chi-square goodness of fit tests showed that the proportion of participants who categorized the male VA as male was significantly higher for the both the Younger Male ($\chi^2(2) = 243.04$, $p < .001$) and Older Male ($\chi^2(2) = 237.88$, $p < .001$) assistants. The percentage of participants who categorized the VA as male was 93% for both male assistants. Additionally, the proportion of participants who categorized the female VA as female was significantly higher for the Younger Female ($\chi^2(2) = 243.04$, $p < .001$) and Older Female ($\chi^2(2) = 227.08$, $p < .001$) assistants. For the Younger Female VA, again 93% of participants categorized the voice as female, while 91% of participants categorized the Older Female VA as female. The marginal difference in gender identification accuracy for the Older Female VA was a result of a slightly larger percentage of participants categorizing the VA as non-binary/third gender in the older female voice condition (5%) compared to the younger female condition (3%). Overall, these results indicate that participants reliably perceived VA gender in accordance with the intended design.

Participant perceptions of VA age were also evaluated using Chi-square goodness of fit tests. Participant age estimates for each VA were categorized into three levels: Young Adult (18-39 years), Middle-aged Adult (40-64 years), and Older Adult (65+ years). These ranges were chosen based on commonly used chronological age groups defined in adult development

literature (Nichols et al., 2003). However, we found that nearly all participant estimates for VA age were within the young adulthood or middle adulthood category ranges. Results indicated that participants categorized the Younger Male ($\chi^2(1) = 34.56$, $p < .001$) and Younger Female ($\chi^2(1) = 49.31$, $p < .001$) VA as young adults at above-chance levels. Approximately 74% and 79% of participants estimated the VA age within the young adult range for the male and female VA, respectively. For the older VAs, the proportion of participants who categorized the Older Male assistant as middle-aged was significantly more than those who categorized the voice as a young adult ($\chi^2(2) = 109.24$, $p < .001$), such that 70% of participants categorized the voice as a middle-aged adult. However, the association between the intended age manipulation and age perceived by participants was nonsignificant for the Older Female VA ($\chi^2(1) = 0.96$, $p = .327$). Participant perceptions of age for this VA were split between the young adult and middle-aged adult age categories, such that 54% of estimates were within the middle adulthood range. These results present considerations for analysis of objective similarity effects and indicate that the intended age manipulation was less successful for the Older Female VA. Given that the voice parameters manipulated for the older VAs were identical, it is possible that a stronger manipulation is needed for dependable age discrimination with the female voice.

To assess if participant age estimates for the older VAs were significantly higher than estimates for the younger VAs, and therefore test if participants were able to derive age from manipulated vocal cues alone, we conducted paired samples t-tests. Results showed that the average perceived age of the Older Male voice ($M = 43$, $SD = 9$) was significantly higher than the Younger Male voice ($M = 34$, $SD = 8$), $t(146) = 10.90$, $p < .001$, d = .90. Additionally, the Older Female voice ($M = 39$, $SD = 7$) was perceived as significantly older than the Younger Female voice ($M = 32$, $SD = 8$), $t(146) = 10.78$, $p < .001$, d = .89. Therefore, although the voices

manipulated to represent an older male and female were not perceived as older adults, they were perceived as relatively older compared to their younger counterparts. Consequently, we find support for Hypothesis 1 that participants in this study were able to discern VA gender and age from manipulated vocal cues. However, the manipulations for VA age did not impact perceptions of the male and female VA to the same degree.

VA Trust and Reliance Intentions

After confirming VA trait manipulations, we evaluated if participants reported significant differences in trust when interacting with VAs designed with varying age and gender (H2) and to explore potential differences in reliance intentions. We found a moderate positive relationship between trust and reliance intentions across the Symptom Assessment Task ($r(148) = .48$, $p < .001$), such that as average trust in VAs increased the mean reported likelihood of relying on VA advice also increased. Therefore, we confirmed that a mixed design multivariate analysis of variance (MANOVA) with trust and reliance intentions as dependent measures would be appropriate instead of running separate analyses for these outcomes to reduce the likelihood of family-wise error. Assumptions of normality, and presence of univariate and multivariate outliers were assessed. Normality was assessed for each outcome at each level of the independent variable (i.e., VA design) using Q-Q scatterplots, and residuals were found to be roughly normally distributed. Considering that MANOVA results are sensitive to outliers, we also analyzed each VA condition for potential influential points. Participants with extreme trust and/or reliance intention values, defined as three standard deviations above or below the mean, were removed from analyses for the respective outcome. Consequently, a total of five participants (3 young adults, 2 older adults) were removed from the analysis due to extreme trust and/or reliance values. There were no multivariate outliers detected in the data. Additionally, we

confirmed that there were significant relationships between trust and reliance intentions within each voice design condition to ensure there were no concerns about the statistical power of the analysis being reduced as a result of nonsignificant associations between the variables. Correlations showed that there was a moderate positive association between the two outcomes for each VA: Older Male ($r(146) = .44$, $p < .001$), Older Female ($r(147) = .42$, $p < .001$), Younger Male ($r(146) = .48$, $p < .001$), and Younger Female ($r(148) = .54$, $p < .001$). Therefore, it was determined that multivariate tests were appropriate.

Additionally, to determine if any covariates should be included in our models for trust and reliance intentions, we assessed relationships between the prior experience, cognitive, and trait-based measures, and the task outcomes. Participants' prior experience using voice assistants, including frequency of use ($r(149) = .02$, $p = .834$) and number of task types ($r(149) = .07$, $p = .385$), was not significantly associated with VA trust across the Symptom Assessment Task. Reliance intentions for VA recommendations were also not significantly associated with frequency of use ($r(149) = .153$, $p = .062$) or number of task types ($r(149) = .136$, $p = .099$). In terms of attentional control ability, correlation results showed that performance on the Simon task was not correlated with trust ($r(142) = .06$, $p = .517$) or reliance intentions ($r(141) = .13$, $p = .116$). Therefore, prior experience with VAs and attentional control ability were not considered further in subsequent analyses.

Pearson correlations revealed small positive associations between PTM scores and average trust ($r(149) = .299$, $p < .001$) and reliance intentions ($r(149) = .204$, $p = .012$) across VA design conditions, such that those with higher dispositional trust tended to report higher trust and intentions to rely on VA recommendations across the Symptom Assessment Task. Additionally, PTM ratings were significantly associated with trust in each VA and reliance

intentions for both younger VAs. However, Conformity was not related to trust in any of the VAs and associations with reliance intentions were small and only present in three of the four voice conditions. Moreover, the association between dispositional compliance and reliance intentions was in the opposite direction of what would be expected; higher trait compliance was related to lower intentions to rely on VA advice. Given that we observed significantly lower dispositional compliance within the older participant group, this could reflect a spurious relationship resulting from age-group differences on this measure. Correlation results for dispositional trust and conformity for each VA design are included in Table 4. Because positive relationships were found between dispositional trust and VA trust during the task, we decided to control for PTM by including the measure as a covariate in analyses of VA trust and reliance intentions.

Table 4.

Correlation results for PTM, Conformity, Trust, and Reliance Intentions across VA conditions.

Measure	PTM	Conformity
PTM	-	.18*
Conformity	.18*	-
Trust		
Older Male VA	.38***	-.04
Older Female VA	.22**	.04
Younger Male VA	.22**	.04
Younger Female VA	.29***	.05
Reliance Intentions		
Older Male VA	.14	-.16*
Older Female VA	.13	-.20*
Younger Male VA	.19*	-.18*
Younger Female VA	.28***	-.14

Note. For each measure, a higher rating indicates more of the construct.
*p<.05 **p<.01 ***p<.001

A mixed design multivariate analysis of covariance (MANCOVA) with VA design as the within-subjects factor and participant demographic group as the between-subjects factor was

used to determine if there were significant differences in trust and reliance intentions across VA conditions after controlling for dispositional PTM and if participant demographic attributes (i.e., gender and age) interacted with the design of the voice to influence trust in the VAs. Results from this analysis were evaluated based on an alpha of 0.05. Multivariate results showed that participant PTM was significantly related to trust and reliance intentions across the task, V = .10, $F(2, 139) = 7.37$, $p < .001$, $\eta^2 = .10$. Dispositional trust accounted for 10% of the variance in trust and reliance between participants. Additionally, when participants were grouped by age and gender, there were significant differences in the combined trust and reliance intention outcomes across groups, V = .23, $F(6, 280) = 5.98$, $p < .001$, $\eta^2 = .11$. Thus, participant demographic group accounted for 11% of the variance in trust and reliance between participants. The main effect of VA design on the combined outcomes after controlling for PTM was significant (V = .10, $F(6, 135) = 2.38$, $p = .032$, $\eta^2 = .10$), such that the design of the voice accounted for 10% of the variance in trust and reliance intentions across the four conditions after controlling for dispositional trust. However, participant demographics did not significantly interact with VA design to influence collective trust and reliance intentions across the task (V = .17, $F(18, 411) = 1.36$, $p = .148$, $\eta^2 = .06$).

To assess Hypothesis 2, we focused on univariate tests for trust (Figure 3). Mauchly's test indicated that the assumption of sphericity held for participant reported trust ($\chi^2(5) = 8.74$, $p = .12$) and therefore no corrections were needed when examining effects. Results revealed significant differences in trust between the four VA designs after controlling for participant PTM scores, $F(3, 420) = 2.88$, $p = .036$, $\eta^2 = .02$. Voice design explained 2% of the variance in trust ratings after accounting for dispositional trust. Additionally, no significant interaction was found between the PTM covariate and VA design for trust ratings ($F(3, 420) = 2.17$, $p = .091$, $\eta^2 = .02$)

suggesting that trust ratings between the four VAs did not significantly differ across participants with varying degrees of dispositional trust. For the main effect of voice design, pairwise comparisons of the adjusted means revealed that trust in the Younger Male assistant (*M* = 3.29) was significantly higher than trust in the Younger Female (*M* = 3.19) assistant and Older Male (*M* = 3.18) assistant, but not significant for the Older Female (*M* = 3.24) assistant. Therefore, we did not find support for Hypothesis 2 predicting that participants would demonstrate a bias for trusting female VA over male agents. Instead, findings suggest participants were inclined toward trusting the younger male agent, but the effect of VA gender and age design on trust was small.

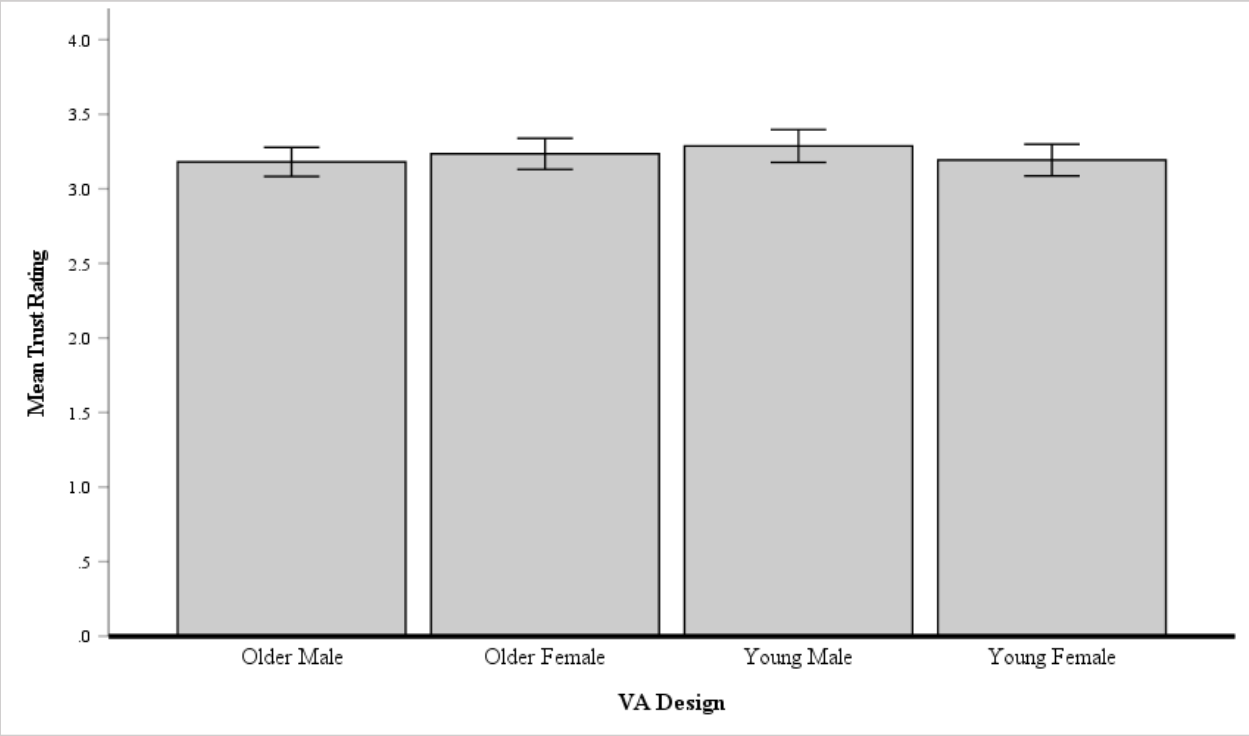

Figure 3. Mean ratings for trust in each of the four voice assistants, adjusted for PTM. The error bars represent the standard error of the means at 95% CI.

Impact of Participant Demographic Traits

To evaluate our prediction that young adult participants would report higher trust in the younger VAs compared to voice agents designed to sound older (H3), we examined univariate results for reported trust in the VAs (Figure 4). Results revealed that there were no significant

differences in mean trust ratings across the Symptom Assessment Task between the participant demographic groups ($F(3) = 0.58$, $p = .632$, $\eta^2 = .01$). Mean trust across all VAs was similar for participants separated by age and gender after accounting for PTM. When examining participant trust across experiences with the four VAs, there was a significant interaction between VA design and participant demographic group on voice assistant trust after controlling for PTM, $F(9, 420) = 1.94$, $p = .045$, $\eta^2 = .04$. Pairwise comparisons showed that young male participants trusted the Younger Male VA ($M = 3.23$) more than the Older Female VA ($M = 3.07$), but there was no significant difference in trust when comparing this VA to the Older Male VA ($M = 3.11$). Additionally, young male participants did not show significant differences in trust for the Younger Female VA ($M = 3.10$) compared to the other VAs. For the young female participants, trust in the Older Male VA ($M = 3.07$) was significantly lower than trust in the Younger Female VA ($M = 3.25$), Younger Male VA ($M = 3.34$), and Older Female VA (M = 3.35). However, trust in the Older Female VA did not significantly differ from trust in both of the younger VAs. Additionally, this participant group had similar trust ratings for the Young Male and Young Female VAs. For the older male and female participant groups, no significant differences were found in trust across VAs. Reported trust adjusted for PTM was consistent across voice conditions for older male ($M_{YM} = 3.33$, $M_{OM} = 3.30$, $M_{OF} = 3.28$, $M_{YF} = 3.24$) and older female ($M_{YM} = 3.26$, $M_{OM} = 3.26$, $M_{OF} = 3.25$, $M_{YF} = 3.18$) participant groups. Older participants appeared to trust VAs equally, regardless of gender and age design. Consequently, while we do find support for an interaction between participant demographics and VA design, we did not observe consistently higher trust in the younger VAs among the younger participant groups as predicted. Therefore, we fail to reject the null for Hypothesis 3.

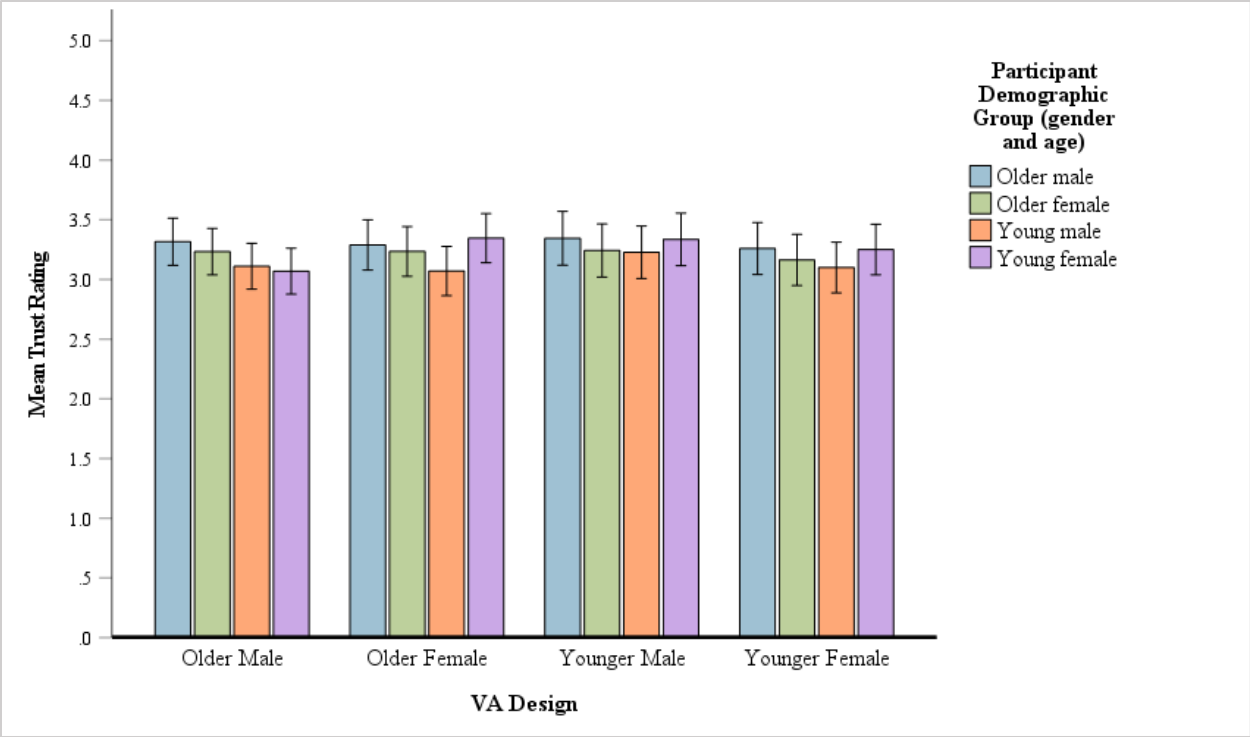

Figure 4. Participant demographic group differences in adjusted mean trust for each voice assistant when controlling for PTM. The error bars represent the standard error of the means at 95% CI.

Potential differences in reported likelihood to rely on triage guidance from VAs was also explored for the model with participant demographic group included. Greenhouse-Geisser estimates with corrected degrees of freedom were used due to violation of the sphericity assumption ($\chi^2(5) = 17.34$, $p = .004$). Results showed that participant demographic groups had significant differences in reliance intentions across interactions with VAs during the Symptom Assessment Task, $F(3) = 11.30$, $p < .001$, $\eta^2 = .20$. Specifically, both older female participants ($M = 6.24$) and older male participants ($M = 5.98$) reported significantly higher intentions to rely on VA recommendations than did the young male ($M = 5.35$) participants and the young female ($M = 5.52$) participants after controlling for dispositional trust. This indicates that older adults may be more willing to rely on diagnoses and triage recommendations from voice assistants than younger adults. Similar to trust, there was not a significant interaction between the covariate, PTM, and VA design on reliance intentions ($F(2.76, 385.96) = 2.64$, $p = .054$, $\eta^2 = .02$).

Intentions to rely on triage guidance were not significantly different across the voice designs at varying levels of participant dispositional trust. We also found a small but significant main effect of VA design on reliance intentions after controlling for PTM, $F(2.76, 385.96) = 2.73$, $p = .049$, $\eta^2 = .02$. Pairwise comparisons for differences across VA designs revealed that average reliance intentions were significantly higher for the Younger Male VA ($M = 5.92$) when compared to the Older Female VA ($M = 5.79$), the Younger Female VA ($M = 5.73$), and the Older Male VA ($M = 5.66$) after adjusting for dispositional trust (Figure *5*). However, the interaction between VA design and participant demographic group on intentions to rely on recommendations was not significant, $F(8.27, 385.96) = 1.08$, $p = .377$, $\eta^2 = .02$. This suggests that there were no similarity attraction effects based on gender and age for reliance intentions (Figure *6*).

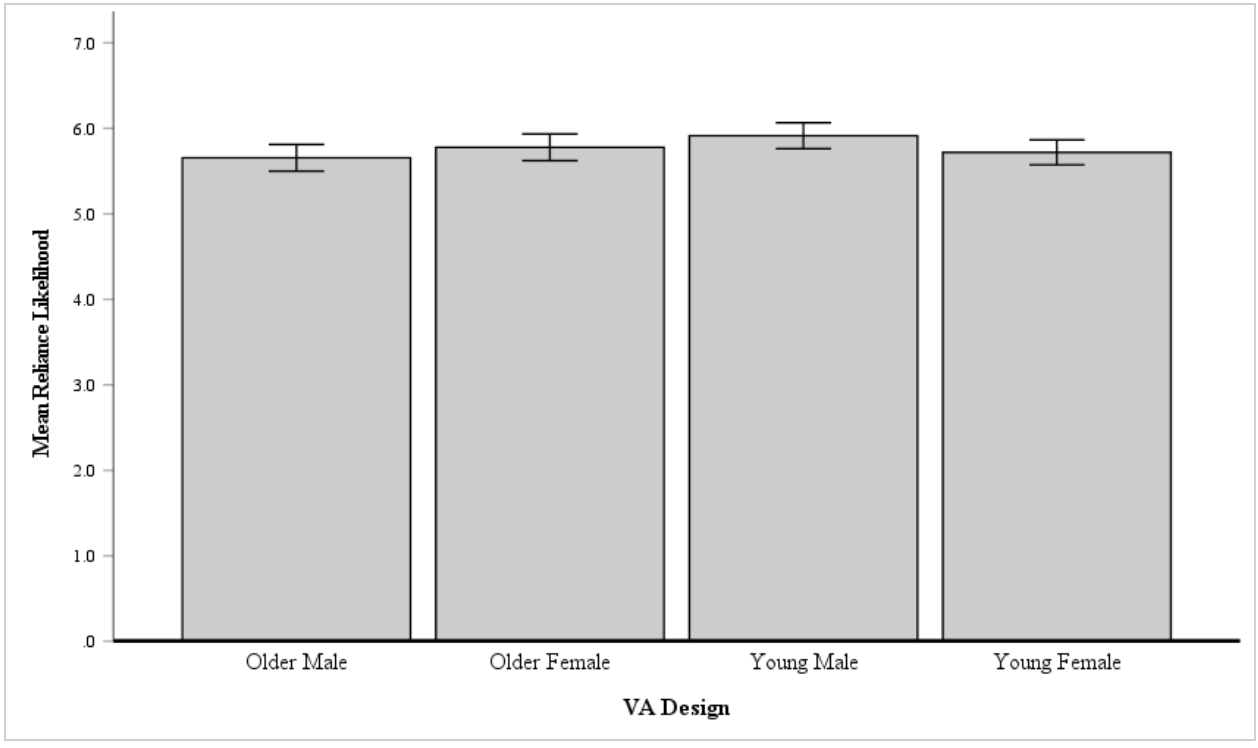

Figure 5. Mean reported likelihood of relying on recommendations from each of the four voice assistants, adjusted for PTM. The error bars represent the standard error of the means at 95% CI.

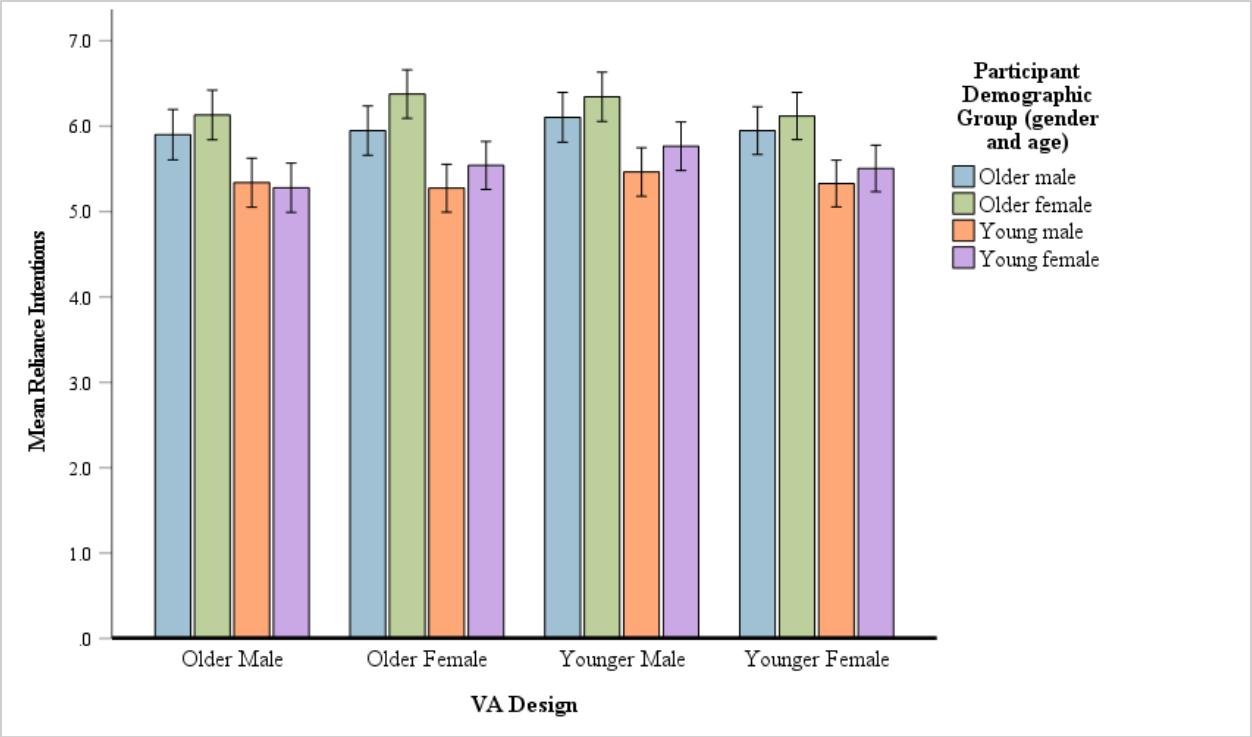

Figure 6. Participant demographic group differences in adjusted mean reliance intentions for each voice assistant when controlling for PTM. The error bars represent the standard error of the means at 95% CI.

Qualitative Feedback for VA Design. After participants provided trust ratings for each VA, they were also given the option to provide open-ended feedback about the assistant. The qualitative responses were analyzed using a bottom-up, inductive approach to code and organize feedback and determine if any themes could be identified from the data. Two coders unfamiliar with the research hypotheses independently coded the data for each VA based on the content of feedback provided. This approach was chosen to avoid influence from preconceptions based on knowledge of expected research outcomes. During the coding phase, comments relevant to perceptions of the specific VA were separated from comments related to the concept of an automated symptom-checking tool in general. In this analysis we focus on themes associated with social inferences about the voice as they may relate to trust and reliance intentions for each VA. After the coders reconciled their data, the analysts grouped codes based on meaning derived from the data and labeled each grouping based on the emergent theme. In total, 82 participants

(55%; 30 younger adults, 52 older adults) provided at least one qualitative response related to the nature of the voice. This resulted in a total of 15 themes identified across all VAs. Frequency counts for all themes for each voice assistant are provided in Table 5. It should be noted that interpretability for these findings is somewhat limited due to the high nonresponse rate to solicitations for VA feedback resulting in low counts among themes. Nonetheless, some patterns were identified and are discussed here.

Table 5.

Frequency counts for each theme discovered in qualitative feedback across voice assistants.

Theme	Older Male VA	Older Female VA	Younger Male VA	Younger Female VA
Artificial	11	12	7	13
Comforting	5	9	9	0
Knowledgeable	8	3	4	0
Trustworthy	3	4	4	2
Untrustworthy	3	2	3	4
Agreeable	3	3	7	0
Emotionless	6	2	2	1
Professional	5	2	2	1
Unfriendly	9	1	0	0
Unpleasant	4	2	0	4
Friendly	0	2	1	5
Human-like	0	0	7	1
Youthful/Inexperienced	0	0	3	3
Unconfident	0	3	0	3
Familiar	0	1	0	3

One prevalent theme common among all VAs was the artificial nature of the voice. Participants commented that it was clear the voices were synthetic and did not originate from a human. Interestingly, the fewest counts for this theme was identified for the Younger Male VA. It is unlikely these perceptions were a result of our voice manipulations as both younger VAs were altered in exactly the same manner, and both were altered to a lesser extent than the older VAs that received a similar number of comments as the younger female assistant. Given the

finding that participants overall reported trusting the Younger Male VA most, it is possible that lower perceptions of artificiality positively contributed to trusting this agent.

Primary themes for the Older Male VA included perceptions that the agent was "unfriendly", "knowledgeable", and "emotionless". Codes constituting these themes illustrated that the agent was perceived by some to be intimidating, wise/competent, and lacking emotion. In contrast, "comforting" was the strongest theme found for the Older Female VA, with codes indicating the voice was perceived as comforting, personable, and compassionate. Predominant Younger Male VA themes included "comforting", "agreeable", and "human-like", with participants describing the voice as calming, likeable, and similar to a human. Finally, comments for the Younger Female VA resulted in themes suggesting the agent was viewed by some as "friendly" and "unpleasant". A theme of youth/inexperience appeared for both of the younger VAs, but not VAs designed to sound older. Likewise, the older-sounding male VA garnered the most comments suggesting the agent was perceived as knowledgeable. This suggests that artificial voices may be perceived as more experienced and competent if they sound older while more youthful sounding voices could be associated with inexperience. However, from these results it is clear that coded data were spread among various themes with no strong attributional patterns emerging. Notably, few participants made explicit attributions of trustworthiness to the voices, and perceptions were split between trustworthy and untrustworthy across all VAs. Analysis of the comments related generally to the idea of a VA symptom checking tool revealed that 17 participants (21% of respondents that provided feedback) expressed general distrust in the ability of an automated tool to provide symptom triage guidance. Participants stated they would only trust a human to perform this type of task or would not be comfortable solely relying on a VA recommendation (i.e., would seek information from additional sources). Therefore,

general hesitancy to trust a VA diagnostic and triage tool could have contributed to the relatively consistent trust and reliance scores observed across voice assistants and explain why voice design did not have a large impact on trust. While we cannot draw any causal conclusions for how these themes relate to VA trust and reliance intentions reported during the task, they do suggest that users inferred characteristics about the agent based on the voice including level of humanness, competency, and warmth despite awareness that the agents were not human.

Perceived Similarity as a Predictor of Trust

After interacting with each voice assistant, participants rated how similar they perceived the agent to be to themselves. Ratings of perceived similarity for each VA broken down by participant demographic group are displayed in Table 6. The observed means revealed a pattern where the highest mean perceived similarity within each participant demographic group was found for the VA that matched participant gender and relative age (i.e., younger vs. older). An exception to this pattern was seen for the older male participants, where mean perceived similarity was slightly higher, but not significant (p = .33), for the Younger Male VA followed by the Older Male VA. This indicates that ratings of perceived similarity could have been partially derived from matching demographic traits between participants and voice agents, but full overlap did not result in participants perceiving VA as completely similar to themselves.

Table 6.

Means and standard deviations for perceived similarity to voice assistants for each participant demographic group.

	Participant Demographic Group			
	Older Male	Older Female	Younger Male	Younger Female
VA Design	M(*SD*)	M(*SD*)	M(*SD*)	M(*SD*)
Older Male VA	4.08 (*1.38*)	3.86 (*1.62*)	3.28 (*1.34*)	2.84 (*1.24*)
Older Female VA	3.92 (*1.44*)	4.16 (*1.46*)	2.92 (*1.35*)	4.24 (*1.05*)
Younger Male VA	4.17 (*1.42*)	3.84 (*1.63*)	4.33 (*1.48*)	3.47 (*1.29*)
Younger Female VA	3.83 (*1.38*)	4.08 (*1.40*)	3.21 (*1.56*)	4.45 (*1.37*)

To test the predictive ability of perceived similarity to the VA and matching VA and participant gender and age (i.e., objective similarity) on trust, we conducted a series of multiple regressions (H4). For each participant, objective similarity to each VA was determined by the degree of overlap in gender and age group. Dummy coding was used for this objective similarity variable such that the no match group was coded as the reference group. Therefore, participants could be 1) a complete mismatch (no overlap in gender/age), 2) a partial match (overlap in gender OR age), or 3) a complete match (overlap in gender AND age). The result is that each participant was a complete mismatch for one VA, a partial match for two VAs, and a complete match for one VA. One caveat to note here is that the older VAs were perceived as significantly older compared to the young VAs, but age estimates for the older VAs were still significantly lower than the age of our older participants. Hence, the effect of "matching" perceived VA age and participant chronological age could not truly be assessed and instead only a match in the direction of perceived age (i.e., older) was evaluated. Initial evaluations indicated that objective similarity was not a significant predictor of trust in each VA and did not significantly improve the variance explained by the model. Therefore, we decided to remove this variable from the regression model and include only perceived similarity to the VA.

Average perceived similarity across all participants was highest for the Younger Male VA (M = 3.95), followed by the Younger Female (M = 3.89), Older Female (M = 3.80), and Older Male VA (M = 3.51). A series of linear regressions illustrated the effect of perceived similarity on trust. For the older male VA, perceived similarity was a significant and positive predictor of trust (b = 0.16, p < .001, r^2 = .13), such that a one-point increase in perceived similarity would predict an increase in trust by 0.16 points. Perceived similarity explained 13% of the variability in trust for the Older Male VA. Perceived similarity was also a significant and

positive predictor of trust for the Older Female ($b = 0.17$, $p < .001$, $r^2 = .14$), Younger Male ($b = 0.21$, $p < .001$, $r^2 = .21$), and Younger Female ($b = 0.17$, $p < .001$, $r^2 = .14$) assistants. Accordingly, for both the Older Female and Younger Female assistants, a one-point increase in perceived similarity would predict an increase of 0.17 points in ratings of trust, accounting for 14% of the variability in trust for both the Older and Younger Female VAs. For the Younger Male VA, a one-point increase in perceived similarity predicts a 0.21-point increase in rated trust and this accounts for 21% of the variability in rated trust. Consequently, we find support for Hypothesis 4, that perceived similarity is a stronger predictor of trust. Objective similarity, as defined by the degree of overlap in participant and VA gender and age factors, did not significantly predict participant trust in VA across the Symptom Assessment Task.

DISCUSSION

Our primary objectives for this research were to build on previous work involving both human physicians and automated agents by exploring whether attitudinal biases, particularly in trust, arise solely from vocal cues without accompanying visual information for age and gender. We specifically investigated whether trust in voice agents suggested biases that may align with gender- and age-based stereotypes. Additionally, we explored the impact of the voice's perceived characteristics on behavior, particularly the inclination to rely on health-related recommendations from the agent. Furthermore, we examined how participants' own intersectional gender and age identity might interact with the perceived traits of the voice agent to determine if patterns would corroborate predictions from similarity-attraction theory. Finally, we explored the comparative roles of objective and perceived similarity to virtual agents to determine if perceived similarity serves as a more robust predictor of trust. Consistent with the theory of computers as social actors (Nass & Moon, 2000), the findings of this research indicate

that the use of anthropomorphic artificial voices, suggesting human characteristics like gender and age, can influence perceptions of an automated decision aid leading to observable differences in trust and reliance intentions. Trust and reliance intentions were significantly higher for the younger-sounding male VA, but differences in trust observed across assistants did not align with expectations from gender- and age-based stereotypes. Importantly, we did not find evidence for similarity-attraction based on demographic similarity, but perceived similarity positively predicted trust in all VAs.

VA Gender and Age Perception

In addition to our primary questions about the influence of VA design on trust, we sought to validate that inferences about gender and age are made in voice-based interactions with VAs. Regarding perceptions of VA gender and age in our study, we found that participants perceived both demographic factors in the intended direction after our pitch, speech rate, and vocal jitter/shimmer manipulations were applied to the voice stimuli. As expected, correct gender identification was high, over 90% for all voices despite no visual cues or names reinforcing gender. This was not surprising given that studies with human and artificial voices reliably find listener judgment of speaker gender is highly accurate, even with very short vocalizations (Bishop & Keating, 2012; Goodman & Mayhorn, 2023; Honorof & Whalen, 2010; Traunmüller & Eriksson, 1993). This can be explained by experience-based knowledge of typical pitch ranges across male and female speakers. We did note a slight increase (2%) in the rate of non-binary/third gender categorization for the Older Female VA and suspect that could have resulted from the reduction in pitch applied to the older female voice. Reducing the pitch of this voice may have led some individuals to perceive it as less distinctly feminine due to the voice falling in

the lower range for female voices, an effect that has been documented in acoustic research on human voices (Bishop & Keating, 2012).

We also found that the VAs were perceived as significantly older when vocal pitch and speech rate were reduced, and jitter and shimmer were added to the voice. However, in contrast to perceived gender, perceptions of VA age were less decisive. Mean age estimates for the male VAs differed by only 9 years while the older female voice was estimated as only 7 years older on average. Thus, while the manipulations did result in VAs sounding significantly older compared to the younger VAs, age estimates were still quite far from late adulthood. Research in age perception of human voices finds that listeners systematically overestimate the age of younger speakers and underestimate the age of older speakers (Waller et al., 2015). While artificial voices clearly do not have a true chronological age, it is possible that this phenomenon influenced age perception of our stimuli as well. In the literature on human communication and vocal acoustics, mean fundamental frequency (perceived as pitch) and speech rate have been identified as the most important cues for estimating speaker age (Benjamin, 1997; Mueller, 1997; Schötz, 2007; Waller, 2019). Normative changes to the aging voice include slowing of speech and altered vocal pitch. Older speakers are typically 20-25% slower at a normal speaking rate compared to young adult speakers (Benjamin, 1997). Furthermore, vocal pitch has been noted to decline after puberty and remain relatively stable until changes occur in late adulthood (Mueller, 1997). However, older adults have also been noted to exhibit much larger heterogeneity in vocal characteristics such as pitch, that can be affected by physical condition, environmental influences, and lifestyle factors (Schötz, 2007). Aside from changes in pitch and speech rate, other acoustic factors (e.g., increased hoarseness, strain, vocal tremor, breathiness, reduced loudness) have been identified as correlates of higher perceived age, and characteristics of the

speech sample (e.g., spontaneous vs. read speech, sample duration) may also influence age estimates (Schötz, 2006). Consequently, discrepancies in vocal characteristics between younger and older speakers are notably less decided than sex differences, and estimating age from speech appears to be a much more complex process than gender identification. Nevertheless, participants in our study still perceived a significant increase in the age of the voice from our manipulations. Thematic analysis of participant feedback for VAs illustrated that respondents were aware of the artificial nature of the voices. Attention to this attribute may have counteracted design cues attempting to reflect vocal age, as participants would be aware that computer-generated voices do not age as humans do.

For our stimuli, the pitch for the older VAs was 12% lower than the younger VA while the rate of speech was by reduced by 15%, and these manipulations resulted in significantly higher age estimates. One methodological concern with research on voice perception is the variability in speech material and techniques making cross-study comparison difficult. A couple of studies attempting to manipulate the perceived age of artificial voices (Huff et al., 2020; Watkins, 2021) have also manipulated speech rate and pitch and found success with significantly higher age estimates, with mean estimates for the older samples falling in the low- to mid-50's. However, these studies used different stock voices for the older and younger voice samples and did not report the absolute difference between the samples for pitch and speech rate after manipulations were applied to the stimuli. Therefore, samples may have included other uncontrolled acoustic factors that impact perceived age (Schötz, 2006) that could also affect trust perceptions (Oleszkiewicz et al., 2017). Our study mitigated these confounding features by using the same stock voice to create older and younger versions of the VA for each gender. The same stock voice was used for both male and both female voices. We also ensured that the male and

female assistants within each target perceived age group were consistently manipulated, so the pitch, speech rate, jitter/shimmer was identical between both older and both younger assistants. Therefore, it is unlikely that any differences in trust or reliance observed in our study could be attributed to inherent differences in vocal properties between stock voices or the degree of vocal manipulation applied.

VA Trust and Reliance Intentions

Contrary to our expectations, we did not find a bias to trust the female voice assistants. Trust was similar across all assistants, with means falling just above the neutral scale midpoint. Pairwise comparisons showed participants reported significantly higher trust in the younger male voice compared to the older-sounding male voice and younger-sounding female voice. However, trust in the Older Female VA was neither significantly higher nor lower than reported trust in the other voice conditions. Therefore, trust perceptions were not consistently impacted by the gender of the voice.

Importantly, participants in this study were not supplied with information typically used to assess the foundational factors identified for trust (i.e., performance, process, purpose) in human-automation interaction (Lee & See, 2004). How well the automation performed on the triage guidance task was unknown as feedback and reliability information were not provided. Additionally, participants were not privy to the method used by the VAs to extract their triage recommendation (i.e., process) or who designed the system and their intended use for it (i.e., purpose). One reason for omitting this information was to determine if trust would be impacted by the voice relaying the recommendations in the absence of data for performance, process, and purpose. A secondary rationale for avoiding an explicit statement of reliability for the VAs, even a false arbitrary value, is that this information is typically fallible or unavailable to users when

interacting with these systems. Our findings did not show strong effects of voice design in the absence of these data as participants reported relatively neutral and consistent trust in all of the assistants. Analysis of qualitative feedback from participants showed a theme of general hesitancy to trust a VA diagnostic and triage tool. This is understandable given consumer VAs are still relatively unsophisticated in the types of tasks they can perform (e.g., informational queries about news/weather, clock functions, to-do lists). While the items in the scale used to assess trust after each VA interaction directed participants to consider "this voice assistant", it is possible participants did not make this distinction. Therefore, a global ambivalence towards trusting this novel technology may have overpowered attributions made based on the perceived gender of the voice. This would explain why only a small portion of participants commented on human-like characteristics (e.g., competence, warmth) perceived from the voices. We also found evidence that higher dispositional trust in automation was associated with higher reported trust in VAs providing triage guidance. This finding aligns with models of interpersonal trust and trust in automation that suggest more stable, trait-like qualities for trust may impact trust assessments when interacting with specific people/technology in a particular context (Lee & See, 2004; Mayer et al., 1995). Therefore, individual differences in dispositional trust may be important to consider when evaluating how discrete VA designs impact user trust. This indicates that studies evaluating trust in specific technology such as voice-based interfaces should assess and control for trust propensity in automated technology as a whole to ensure results reflect actual trust differences for anthropomorphic agent designs.

With respect to stereotypes, the medical field has been traditionally dominated by male physicians and stereotypes of the prototypical male doctor still persist despite a changing professional landscape (Boge et al., 2019; Engleman, 1974). Research has shown higher

satisfaction with human doctors when they display stereotypical masculine (distanced, dominant) and feminine (empathetic, less assertive) behavior (Mast et al., 2008). Literature on aging stereotypes typically show that older individuals are stereotyped as "warm but incompetent" (Cuddy et al., 2005). However, prior work has also indicated that negative age-based stereotypes in a medical context may be less pervasive (Singer, 1986) and older physicians may benefit from perceptions of increased competence due to experience (Hall et al., 2020). Perhaps the status of the younger male voice as belonging to a less negatively stereotyped group positively influenced trust in this agent. However, participants who provided subjective feedback on the VAs after interactions did not frequently report social attributions toward the agents and marginal patterns observed for those who did were not consistent with stereotypes as the older male agent was described as competent but cold and comments for the younger female agent, while viewed by some as warm/friendly, did not show any strong patterns.

Comments for the younger male assistant showed attributions such as comforting, agreeable, and human-like. Notably, the "human" theme was almost exclusive to this agent. Interfaces with increased anthropomorphism have been shown to improve trust in automated decision aids (Pak et al., 2012). Consequently, participant trust in this voice may have been augmented by participants' perceptions of higher anthropomorphism for this voice instead of positive or negative attributions stemming from stereotypes about gender and age. Additionally, when taking a categorical approach to VA perceived age in our data, categorization of the Older Female VA was split between the young and middle age group, while the Older Male VA was more consistently viewed as middle-aged. This could explain why qualitative feedback in our study offered limited support for consistent perceptions of competence in the older-sounding VAs and why perceptions of knowledgeability occurred more frequently for the Older Male VA

than any of the other assistants. Given the age estimates for the older VAs used in our study were not in the range for older adulthood, it is likely that the age manipulation was not sufficient to activate negative stereotypes about aging for the older voices. This is a limitation of our work and offers an opportunity for future investigation.

Prior work with embodied conversational agents and human voice perception has shown inconsistent findings with respect to gender (Rheu et al., 2021), and this is further complicated by the fact that studies have examined related but separable outcomes such as trust (e.g., Schirmer et al., 2019), preference (e.g., Azevedo et al., 2018), and perceived competence (e.g., Hall et al., 2020). Furthermore, trust in anthropomorphic automation has been consistently demonstrated to vary between technological artifacts and contexts (Lee & See, 2004), and this study adds to the extremely limited work thus far examining how voice-based agent gender and age influences trust attitudes. Two recent studies have examined trust in synthetic voices providing assistance on a health-based task and are therefore most closely related to this research (Goodman & Mayhorn, 2023; Watkins, 2021). Goodman & Mayhorn (2023) showed that participants had higher trust in female VAs compared to their male counterparts, although no differences observed in reliance behavior, and this was a result of significantly higher scores on the benevolence subscale of the trust measure. In contrast, Watkins (2021) found no overall trust differences based on VA perceived gender. However, they did find that trust significantly differed for all VA gender and age combinations, with highest mean trust reported for the young female agent.

Disparate findings could reflect key methodological differences in the measurement of trust across studies. Goodman & Mayhorn (2023) utilized a semantic differential scale where participants gave ratings for twelve word-pairs mapping to three factors of trust. Watkins (2021)

utilized two questions to measure trust (i.e., To what extent do you trust the VA in this scenario? To what extent would you be likely to follow the VA's recommendation in this scenario?) assessed trust for each factorial combination of independent factors in only a single trial. In contrast, we utilized a previously validated trust scale comprised of three factors (i.e., benevolence, competence, and perceived risk) developed from trust theory, and operationalized willingness to follow VA recommendations as reliance intentions in accordance with research demonstrating that trust and reliance are related but separable factors (Lee & See, 2004). Additionally, trust was measured after a series of six trials where the VA provided voice-based recommendations to allow participants to assess their trust in the aid after a short history of interactions. Trust has been characterized as a multi-faceted latent concept, and studies endeavoring to measure this construct have employed a variety of methods including self-report, behavioral, and physiological indicators (Kohn et al., 2021). Self-report questionnaires are considered to be the most direct measurement of trust attitudes, but must be appropriate for the target application domain, context, and layer of trust (Alsaid et al., 2023). Therefore, researchers examining trust in automation are challenged to identify an appropriate questionnaire from the multitude of instruments available. The measure of history-based trust chosen for our study (Gulati et al., 2019) is based firmly on theory of trust in automation and has the advantage of previous rigorous empirical validation demonstrating satisfactory internal reliability and convergent validity. Furthermore, it was designed to capture history-based trust after interacting with a specific technological artifact and therefore reflects differences in this layer of trust. This multi-factor scale may also be better suited for measuring the complex, multifaceted nature of trust in an automated system than single-item scales that suffer from concerns with low content validity, sensitivity, and reliability. These arguments point to the appropriateness of our chosen

measure to answer our research questions and illustrate why comparing results on trust across studies can be problematic.

As for findings related to objective similarity between gender and age of participants and VAs, we discovered an interaction between participant demographic traits and VA design, but results for objective similarity were not consistent with predictions proposed by similarity-attraction theory. Young male participants showed significantly lower trust for the most demographically dissimilar VA (i.e., Older Female) compared to the most similar VA (i.e., Younger Male), and comparable levels of trust for VAs matching on a single demographic trait. Likewise, younger females reported significantly lower trust for the most demographically dissimilar VA (i.e., Older Male). However, trust was commensurate between the most demographically similar VA (i.e., Younger Female) and VAs with a single matching demographic trait. Furthermore, VA design did not appear to impact trust for older participants. However, a major limitation here is the fact that the older VAs were only perceived as significantly older compared to the younger VAs. Mean age for both older assistants was still much younger than the average chronological (M = 70.1) and subjective (M = 58.9) ages of older participants. More work is needed to understand the impact of matching age on trust perceptions.

Studies examining similarity effects based on surface-level traits have yielded mixed results (Aguirre-Rodriguez et al., 2015; He & Burns, 2022; Reinkemeier & Gnewuch, 2022; ter Stal et al. (2020). For example, ter Stal et al. (2020) found that young respondents preferred to select static images of young eHealth coaches at first glance and men typically preferred selecting a male agent. Reinkemeier & Gnewuch (2022) found that participants trusted VAs with a matching personality (i.e., extroverted vs. introverted), manipulated by speech rate, pitch, and loudness, more than those with a personality mismatch, but matching VA gender did not

influence trust. Similarly, He & Burns (2022) found no similarity-attraction effects for an in-vehicle VA when matching participant gender or aggressiveness. In our study, "matching" participant and VA gender and age did not significantly predict trust. However, higher perceived similarity significantly predicted participant trust in all VAs, explaining 13-21% of the variance in reported trust across assistants. Future work might focus on examining the factors that users consider when judging similarity to a voice, including factors such as lexical complexity (i.e., level of sophistication in a person's vocabulary) and more internal qualities such as personality (e.g., extroversion, agreeableness), and determine if matching those qualities improves trust in voice agents.

We also found evidence for higher reliance intentions when recommendations were provided by the Younger Male VA compared to all other assistants. These results indicate that users may be more predisposed to comply with triage recommendations if they are provided by a younger-sounding male voice. Additionally, when considering the demographic traits of the participants, older adults had significantly higher intentions to rely on VA advice overall compared to young participants after controlling for dispositional trust in automation. This could indicate that older people using a VA symptom-checking tool would be more likely to act on the advice provided, seeking professional medical care or implementing self-care, than younger users if they were experiencing symptoms as described in the vignettes. Importantly, there was no indication that objective similarity based on gender and age influenced reliance, because the interaction was nonsignificant.

Participants in our study reported relatively high reliance intentions, with means well above the scale midpoint that indicates neutral intentions. Our participants were unaware of the actual reliability of the VAs throughout the task; they were told that the technology was new and

still being developed, implying that there was reasonable potential for some VA recommendation errors. Considering that all of the recommendations VAs provided during the task were actually correct, high reliance is an effective strategy in this case. However, studies of real symptom-checker technology typically find high rates of diagnostic errors (Semigran et al., 2015) and perfectly accurate automation of any type does not exist, so if these high reliance intentions were observed with real VAs providing triage guidance there is a concern for potential misuse (Parasuraman & Riley, 1997).

Our finding that participant reliance intentions for VA recommendations significantly differed between the levels of triage suggests that participants understood the potential for erroneous advice and considered the risk and personal health consequences of pursuing the advice. For instance, if the VA suggested it was not necessary to seek professional medical care and medical attention was truly needed, the individual's condition could worsen and negatively impact their health. Therefore, the more risk-averse path would be to check additional sources and/or talk with a provider. The discovery of a stronger effect of triage urgency for older adults also supports this theory as decision-making literature finds that both young and older adults are risk-averse, but older individuals exhibit higher risk aversion, especially when framed by potential losses such as negative health outcomes (Albert & Duffy, 2012; Mayhorn et al., 2002). Overall, we present novel findings with implications for designers who may seek to increase reliance on health-relevant recommendations provided by voice-based interfaces like VAs. However, we caution against targeting uniformly higher reliance among users unless the decision aid has been demonstrated to be highly reliable because adherence to incorrect guidance could negatively impact user health (Chambers et al., 2019). When VA technology is applied to a decision-making context with important health and safety consequences (i.e., symptom triage

guidance), choosing a younger male voice may encourage higher compliance with recommendations regardless of user gender and age profiles. However, voice design explained a small portion of the variance in reliance intentions, so designers wishing to improve user reliance on recommendations may instead want to focus on strategies to improve global trust and confidence in the algorithms driving automated triage guidance, such as explaining why the VA chose its recommendation.

Limitations

Manipulating the perceived age of VAs using only a few paralinguistic factors proved to be a challenge in our research. Making artificial voices sound stereotypically "old" may require researchers to compromise control of acoustic variables that could confound results for other perceptions, such as trust, and may also necessitate advanced technical knowledge of sound engineering. Additionally, our experimental design only allowed us to identify behavioral intentions with hypothetical consequences to personal health and not real-world behavior with genuine consequences. When people make real-world health decisions, factors such as medical insurance, financial means, influence from additional information sources (e.g., asking for a friend's advice, searching online), and general willingness or availability to access care are additional considerations that might affect behavior. Finally, it could be argued that our study sample is unrepresentative of the population as most participants were white, highly educated, possessed high levels of health literacy, and were skewed toward the lower end of age ranges for young and late adulthood. The older sample was quite familiar with VA technology and may have been more technologically experienced in general as this study was conducted online.

Future Directions

Future investigations should attempt to bridge the gaps identified in these findings and address the limitations described for this study. Studies should include a broader sample to represent population variability in age, education, health literacy, and technological/VA experience. This may be accomplished by recruiting in local communities, such as churches and senior centers, instead of utilizing pools of established research participants who may be more homogenous than the population as a whole. While it was somewhat surprising to find the older participants had more experience using VAs, given that rate of adoption is typically lower in the older population (Olmstead, 2017), recent work has indicated that older adults are increasingly adopting VA technology (Kinsella, 2019). In the future, studies may examine differences in trust and reliance with older samples that have relatively high and low experience using VAs.

Additionally, more work is needed to understand how experimenters can reliably create stimuli that reflect vocal attributes of older individuals. Given the technical expertise demanded by the complex nature of human vocal acoustics, researchers may benefit from collaborating with speech synthesis specialists and sound engineers that have additional knowledge and tools for creating the artificial voices that enable VAs to interact with users. Replicating findings for trust in VAs using the same stimuli as previous investigations would also be beneficial and allow for more direct comparison across studies. In a similar vein, utilizing similar self-report measures of trust and combining this with other measures (e.g., behavioral, physiological) could be valuable for contrasting work examining trust in automated agents and determining if evidence converges on similar findings.

Observation of reliance behavior with VAs applied to medical tasks is also needed if we are to anticipate how these systems may be used in the future as their capabilities expand in this domain. Despite the limitation of measuring reliance intentions in our study, research indicates

that behavioral intentions mediate the relationship between attitudes and behavior and that intentions are typically strongly associated with attitudes and behavior (Kim & Hunter, 1993). Therefore, it is reasonable to propose that participant reliance intentions reported in our study are a sufficient indicator of expected reliance behavior if interacting with this type of automated system in the real world. Nevertheless, an opportunity exists to observe behavioral patterns of reliance in addition to behavioral intentions.

Another opportunity exists to further explore the factors involved in perceived similarity to humanlike agents to identify how this can be leveraged to improve predictions of trust in voice-based automation. Similarity in personality traits and vocal properties of the trustor (e.g., speech rate, pitch) both offer intriguing avenues for exploration.

Conclusions

These findings contribute to nascent research on the role of voice design in user trust and reliance intentions, specifically within the context of automated decision aids applied to personal health management. Imbuing voice-based automation with anthropomorphic qualities suggesting gender and age does appear to have a small impact on trust and reliance intentions. However, matching surface-level demographic characteristics, such as gender and age, between users and voice agents may not be an effective strategy for engendering trust. Instead, researchers and designers should study what voice qualities improve perceived similarity to voice agents, as this may better predict trust and reliance on the system providing voice-based recommendations.

REFERENCES

Aguirre-Rodriguez, A., Bóveda-Lambie, A. M., & Miniard, P. W. (2015). The impact of consumer avatars in Internet retailing on self-congruity with brands. *Marketing Letters, 26*(4), 631-641. https://doi.org/10.1007/s11002-014-9296-z

Albert, S. M., & Duffy, J. (2012). Differences in risk aversion between young and older adults. *Neuroscience and Neuroeconomics, 2012*(1). https://doi.org/10.2147/NAN.S27184

Alecu, A. I. (2020). Which doctors do we trust? A vignette experiment of how gender and ethnicity influence trust. *Ethnicities, 20*(5), 939-958. https://doi.org/10.1177/1468796819877737

Alsaid, A., Li, M., Chiou, E. K., & Lee, J. D. (2023). Measuring trust: a text analysis approach to compare, contrast, and select trust questionnaires. *Frontiers in Psychology, 14*:1192020. https://doi.org/10.3389/fpsyg.2023.1192020

American Psychological Association. (2018, April 19). *Self-concept*. APA Dictionary of Psychology. Retrieved January 30, 2023, from https://dictionary.apa.org/self-concept

Azevedo, R. F. L., Morrow, D., Graumlich, J., Willemsen-Dunlap, A., Hasegawa-Johnson, M., Huang, T. S., Gu, K., Bhat, S., Sakakini, T., Sadauskas, V., & Halpin, D. J. (2018). Using conversational agents to explain medication instructions to older adults. *AMIA Annual Symposium Proceedings, 2018,* 185-194.

Benbasat, I., Dimoka, A., Pavlou, P. A., & Qiu, L. (2020). The role of demographic similarity in people's decision to interact with online anthropomorphic recommendation agents: Evidence from a functional magnetic resonance imaging (fMRI) study. *International Journal of Human-Computer Studies, 133*, 56-70. https://doi.org/10.1016/j.ijhcs.2019.09.001

Bishop, J., & Keating, P. (2012). Perception of pitch location within a speaker's range: fundamental frequency, voice quality and speaker sex. *The Journal of the Acoustical Society of America*, *132*(2), 1100–1112. https://doi.org/10.1121/1.4714351

Boge, L. A., Dos Santos, C., Moreno-Walton, L. A. Cubeddu, L. X., & Farcy, D. A. (2019). The relationship between physician/nurse gender and patients' correct identification of health care professional roles in the emergency department. *Journal of Women's Health, 28*(7), 961-964. https://doi.org/10.1089/jwh.2018.7571

Bowen, C. E., Spuling, S. M., Kornadt, A. E., & Wiest, M. (2020). Young people feel wise and older people feel energetic: Comparing age stereotypes and self-evaluations across adulthood. *European Journal of Ageing, 17*(4), 435-444. https://doi.org/10.1007/s10433-019-00548-4

Burgoyne, A. P., & Engle, R. W. (2020). Attention control: A cornerstone of higher-order cognition. *Current Directions in Psychological Science, 29*(6), 624-630. https://doi.org/10.1177/0963721420969371

Centers for Disease Control and Prevention. (2022, November 29). *Health Topics A-Z*. https://www.cdc.gov/health-topics.html

Chambers, D., Cantrell, A. J., Johnson, M., Preston, L., Baxter, S. K., Booth, A., & Turner, J. (2019). Digital and online symptom checkers and health assessment/triage services for urgent health problems: Systematic review. *BMJ Open*, *9*(8), e027743. https://doi.org/10.1136/bmjopen-2018-027743

Cleveland Clinic. (n.d.). *Diseases & Conditions*. Retrieved March 15, 2023, from https://my.clevelandclinic.org/health/diseases?dFR[type][0]=Diseases

Cohen J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Routledge Academic.

Cuddy, A. J., Norton, M. I., & Fiske, S. T. (2005). This Old Stereotype: The pervasiveness and persistence of the elderly stereotype. *Journal of Social Issues*, *61*(2), 267–285. https://doi.org/10.1111/j.1540-4560.2005.00405.x

Engleman E. G. (1974). Attitudes toward women physicians. A study of 500 clinic patients. *The Western Journal of Medicine*, *120*(2), 95–100.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, *28*, 1–11. https://doi.org/10.3758/BF03203630

Ermolina, A., Tiberius, V. (2021). Voice-controlled intelligent personal assistants in health care: International Delphi study. *Journal of Medical Internet Research*, *23*(4), e25312. https://doi.org/10.2196/25312

Fox, S., & Duggan M. (2013, January 15). *Health Online 2013*. Pew Research Center. http://www.pewinternet.org/2013/01/15/health-online-2013-2/

Furst, J. (2020, April 27). *Mayo clinic introduces skill for amazon's alexa about COVID19*. Mayo Clinic News Network. https://newsnetwork.mayoclinic.org/discussion/mayo-clinic-introduces-skill-for-amazons-alexa-about-covid-19/

Goodman, K. L., & Mayhorn, C. B. (2023). It's not what you say but how you say it: Examining the influence of perceived voice assistant gender and pitch on trust and reliance. *Applied Ergonomics*, *106*, 103864. https://doi.org/10.1016/j.apergo.2022.103864

Gulati, S., Sousa, S., & Lamas, D. (2019). Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology*, *38*(10), 1004–1015. https://doi.org/10.1080/0144929x.2019.1656779

Hall, J. A., Ruben, M. A., & Swatantra. (2020). First impressions of physicians according to their physical and social group characteristics. *Journal of Nonverbal Behavior, 44*(2), 279-299. https://doi.org/10.1007/s10919-019-00329-8

Hamilton, D. L. (1979). A cognitive-attributional analysis of stereotyping. *Advances in Experimental Social Psychology, 12*, 53-84. https://doi.org/10.1016/s0065-2601(08)60259-2

He, F., & Burns, C. M. (2022). A battle of voices: A study of the relationship between driving experience, driving style, and in-vehicle voice assistant character. *Proceedings of the 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '22)*. ACM, New York, NY, USA, 236-242. https://doi.org/10.1145/3543174.3546845

Honorof, D., & Whalen, D. H. (2010). Identification of speaker sex from one vowel across a range of fundamental frequencies. *The Journal of the Acoustical Society of America, 128*(5), 3095-3104. https://doi.org/10.1121/1.3488347

Huff, E. W., Jr., Stigall, B., Brinkley, J., Pak, R., & Caine, K. (2020). Can computer-generated speech have an age? *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20).* ACM, New York, NY, USA, 1–7. https://doi.org/10.1145/3334480.3383082

Jhangiani, R., & Tarry, H. (2014). Chapter 11: Stereotypes, Prejudice, and Discrimination: Principles of social psychology – 1st International Edition. Victoria, B.C.: BC campus.

Johns Hopkins Medicine. (n.d). *Conditions and Diseases*. Johns Hopkins Medicine Health. Retrieved March 15, 2023, from https://www.hopkinsmedicine.org/health/conditions-and-diseases

Juniper Research. (2018, February 12). *Digital voice assistants in use to triple to 8 billion by 2023, driven by smart home devices*. https://www.juniperresearch.com/press/digital-voice-assistants-in-use-to-8-million-2023

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge University Press. https://doi.org/10.1017/CBO9780511808098.004

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263–291. https://doi.org/10.2307/1914185

Kim, M.-S., & Hunter, J. E. (1993). Relationships among attitudes, behavioral intentions, and behavior: A meta-analysis of past research, part 2. *Communication Research, 20*(3), 331-364. https://doi.org/10.1177/009365093020003001

Kinsella, B. (2019, June 21). *Voice assistant demographic data – Young consumers more likely to own smart speakers while over 60 bias toward alexa and siri*. Voicebot.ai. https://voicebot.ai/2019/06/21/voice-assistant-demographic-data-young-consumers-more-likely-to-own-smart-speakers-while-over-60-bias-toward-alexa-and-siri/

Kinsella, B. (2022, January 7). *Voice assistant use in healthcare nearly tripled over two years and demand still outstrips supply – New report*. Voicebot.ai. https://voicebot.ai/2022/01/07/voice-assistant-use-in-healthcare-nearly-tripled-over-two-years-and-demand-still-outstrips-supply-new-report/

Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y.-C., and Shaw, T. H. (2021). Measurement of trust in automation: a narrative review and reference guide. *Frontiers in Psychology, 12*:604977. https://doi.org/10.3389/fpsyg.2021.604977

Kornadt, A. E., Voss, P., & Rothermund, K. (2017). Age stereotypes and self-views revisited: Patterns of internalization and projection processes across the life span. *The Journals of Gerontology: Series B, 72*(4), 582-592. https://doi.org/10.1093/geronb/gbv099

Kotter-Grühn, D., Neupert, S. D., & Stephan, Y. (2015). Feeling old today? Daily health, stressors, and affect explain day-to-day variability in subjective age. *Psychology & Health, 30*(12), 1470-1485. https://doi.org/10.1080/08870446.2015.1061130

Krauss, R. M., Freyberg, R., & Morsella, E. (2002). Inferring speakers' physical attributes from their voices. *Journal of Experimental Social Psychology, 38*, 618-625. https://doi.org/10.1016/s0022-1031(02)00510-3

Ku, B., Itagaki, T., & Seaborn, K. (2023). Dis/Immersion in mindfulness meditation with a wandering voice assistant. *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. ACM, New York, NY, USA, 1-6. https://doi.org/10.1145/3544549.3585627

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

Lee, S.-Y.D., Stucky, B.D., Lee, J.Y., Rozier, R.G. & Bender, D.E. (2010). Short assessment of health literacy—Spanish and English: A comparable test of health literacy for Spanish and English speakers. *Health Services Research, 45*(4), 1105-1120. https://doi.org/10.1111/j.1475-6773.2010.01119.x

Liu, X., Banich, M. T., Jacobson, B. L., & Tanabe, J. L. (2004). Common and distinct neural substrates of attentional control in an integrated Simon and spatial Stroop task as assessed by event-related fMRI. *NeuroImage, 22*(3), 1097-1106. https://doi.org/10.1016/j.neuroimage.2004.02.033

Lyons, J. B., & Stokes, C. K. (2012). Human–Human reliance in the context of automation. *Human Factors, 54*(1), 112-121. https://doi.org/10.1177/0018720811427034

Mast, M. S., Hall, J. A., Klöckner, C., & Choi, E. (2008). Physician gender affects how physician nonverbal behavior is related to patient satisfaction. *Medical Care, 46*(12), 1212-1218. https://doi.org/10.1097/MLR.0b013e31817e1877

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review, 20*(3), 709-734. https://doi.org/10.5465/amr.1995.9508080335

Mayhorn, C. B., Fisk, A. D., & Whittle, J. D. (2002). Decisions, decisions: Analysis of age, cohort, and time of testing on framing of risky decision options. *Human Factors, 44*(4), 515-521. https://doi.org/10.1518/0018720024496935

McAleer, P., Todorov, A., & Belin, P. (2014). How do you say hello? Personality impressions from brief novel voices. *PloS One, 9*(3), e90779. https://doi.org/10.1371/journal.pone.0090779

Mehrabian, A., & Stefl, C.A. (1995). Basic temperament components of loneliness, shyness, and conformity. *Social Behavior and Personality, 23*, 253-264. https://doi.org/10.2224/sbp.1995.23.3.253

Merritt, S. M. (2011). Affective processes in Human–Automation interactions. *Human Factors, 53*(4), 356-370. https://doi.org/10.1177/0018720811411912

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors, 55*(3), 520–534. https://doi.org/10.1177/0018720812465081

Meyer, A., Giardina, T. D., Spitzmueller, C., Shahid, U., Scott, T., & Singh, H. (2020). Patient perspectives on the usefulness of an artificial intelligence-assisted symptom checker: Cross-sectional survey study. *Journal of Medical Internet Research, 22*(1), e14679. https://doi.org/10.2196/14679

Millenson, M. L., Baldwin, J. L., Zipperer, L., & Singh, H. (2018). Beyond Dr. Google: the evidence on consumer-facing digital tools for diagnosis. *Diagnosis, 5*(3), 95–105. https://doi.org/10.1515/dx-2018-0009

Mueller, P. (1997). The aging voice. *Seminars in Speech and Language, 18*(2), 159-169. https://doi.org/10.1055/s-2008-1064070

Nass, C., & Lee, K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied, 7*(3), 171-181. https://doi.org/10.1037/1076-898X.7.3.171

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81-103. https://doi.org/10.1111/0022-4537.00153

Nass, C., Moon, Y., & Green, N. (1997). Are machines gender neutral? Gender-Stereotypic responses to computers with voices. *Journal of Applied Social Psychology, 27*(10), 864-876. https://doi.org/10.1111/j.1559-1816.1997.tb00275.x

NCH Software. (2023) Voxal Voice Changer (Version 8.00) [Computer Software]. Canberra, Australia. Retrieved from https://www.nchsoftware.com/voicechanger/index.html

Nichols, T. A., Rogers, W. A., & Fisk, A. D. (2003). Do you know how old your participants are? *Ergonomics in Design, 11*(3), 22-26. https://doi.org/10.1177/106480460301100305

Oleszkiewicz, A., Pisanski, K., Lachowicz-Tabaczek, K., & Sorokowska, A. (2017). Voice-Based assessments of trustworthiness, competence, and warmth in blind and sighted adults. *Psychonomic Bulletin & Review, 24*(3), 856-862. https://doi.org/10.3758/s13423-016-1146-y

Olmstead, K. (2017). *Nearly half of Americans use digital voice assistants, mostly on their smartphones*. Pew Research. https://www.pewresearch.org/fact-tank/2017/12/12/nearly-half-of-americans-use-digital-voice-assistants-mostly-on-their-smartphones/

Pak, R., Fink, N., Price, M., Bass, B., & Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics, 55*(9), 1059-1072. https://doi.org/10.1080/00140139.2012.691554

Pak, R., McLaughlin, A. C., & Bass, B. (2014). A multi-level analysis of the effects of age and gender stereotypes on trust in anthropomorphic technology by younger and older adults. *Ergonomics, 57*(9), 1277-1289. https://doi.org/10.1080/00140139.2014.928750

Parasuraman, R., & Riley, V. (1997). Humans and automation: use, misuse, disuse, abuse. *Human Factors, 39*(2), 230-253. https://doi.org/10.1518/001872097778543886

Pasupathi, M. (1999). Age differences in response to conformity pressure for emotional and nonemotional material. *Psychology and Aging*, *14*(1), 170–174. https://doi.org/10.1037/0882-7974.14.1.170

Reinkemeier, F., & Gnewuch, U. (2022). Match or mismatch? How matching personality and gender between voice assistants and users affects trust in voice commerce. *Proceedings of the 55th Hawaii International Conference on System Sciences*. https://doi.org/10.24251/HICSS.2022.528

Rheu, M., Shin, J., Peng, W., & Huh-Yoo, J. (2021). Systematic review: Trust-building factors and implications for conversational agent design. *International Journal of Human–Computer Interaction, 37*(1), 81-96. https://doi.org/10.1080/10447318.2020.1807710

Schirmer, A., Feng, Y., Sen, A., & Penney, T. B. (2019). Angry, old, male – and trustworthy? How expressive and person voice characteristics shape listener trust. *PloS One, 14*(1), e0210555. https://doi.org/10.1371/journal.pone.0210555

Schötz, S. (2006). *Perception, Analysis and Synthesis of Speaker Age*. [Doctoral Thesis (monograph), Phonetics]. Linguistics and Phonetics.

Schötz, S. (2007). Acoustic analysis of adult speaker age. In: Müller, C. (eds) *Speaker Classification I. Lecture Notes in Computer Science* (vol 4343). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-74200-5_5

Semigran, H. L., Linder, J. A., Gidengil, C., & Mehrotra, A. (2015). Evaluation of symptom checkers for self-diagnosis and triage: Audit study. *BMJ*, 351:h3480. https://doi.org/10.1136/bmj.h3480

Sezgin, E., Huang, Y., Ramtekkar, U., & Lin, S. (2020). Readiness for voice assistants to support healthcare delivery during a health crisis and pandemic. *NPJ Digital Medicine*, *3*(1), 122. https://doi.org/10.1038/s41746-020-00332-0

Shah, R., Ogden, J. (2006). 'What's in a face?' The role of doctor ethnicity, age and gender in the formation of patients' judgements: An experimental study. *Patient Education and Counseling, 60*(2), 136–141. https://doi.org/10.1016/j.pec.2004.12.005

Singer, M. S. (1986). Age stereotypes as a function of profession. *The Journal of Social Psychology, 126*(5), 691-692. https://doi.org/10.1080/00224545.1986.9713647

Smits, C., Kapteyn, T. S., & Houtgast, T. (2004). Development and validation of an automatic speech-in-noise screening test by telephone. *International journal of audiology*, *43*(1), 15–28. https://doi.org/10.1080/14992020400050004

Smith, D. (2022, May 2). *I used Alexa's symptom checker to track my health for a month — Here are the pros and cons*. CNN Underscored. https://www.cnn.com/cnn-underscored/electronics/amazon-alexa-symptom-checker

Stoet, G. (2010). PsyToolkit - A software package for programming psychological experiments using Linux. *Behavior Research Methods*, *42*(4), 1096-1104. https://doi.org/10.3758/BRM.42.4.1096

Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, *44*(1), 24-31. https://doi.org/10.1177/0098628316677643

ter Stal, S., Tabak, M., op den Akker, H., Beinema, T., & Hermens, H. (2020). Who do you prefer? The effect of age, gender and role on users' first impressions of embodied conversational agents in ehealth. *International Journal of Human–Computer Interaction, 36*(9), 881-892. https://doi.org/10.1080/10447318.2019.1699744

Traunmüller, H., & Eriksson, A. (1993). The frequency range of the voice fundamental in the speech of male and female adults. [Unpublished manuscript]. Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-10290

Tsantani, M. S., Belin, P., Paterson, H. M., & McAleer, P. (2016). Low vocal pitch preference drives first impressions irrespective of context in male voices but not in female voices. *Perception, 45*(8), 946-963. https://doi.org/10.1177/0301006616643675

Waller, S. S. (2019). *Estimation of speaker age: Effects of speech properties and speech material* [PhD dissertation, Mid Sweden University]. Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:hig:diva-31193

Waller, S. S., Eriksson, M., & Sörqvist, P. (2015). Can you hear my age? Influences of speech rate and speech spontaneity on estimation of speaker age. *Frontiers in Psychology, 6*(978). https://doi.org/10.3389/fpsyg.2015.00978

Wang, T., Kawaguchi, I., Kuzuoka, H., & Otsuki, M. (2018). Effect of manipulated amplitude and frequency of human voice on dominance and persuasiveness in audio conferences. *Proceedings of the ACM on Human-Computer Interaction, 2*(CSCW), 1-18. https://doi.org/10.1145/3274446

Watkins, H. (2021). *User perceptions and stereotypic responses to gender and age of voice assistants* (Publication No. 3652) [Master's thesis, Clemson University]. Retrieved from https://tigerprints.clemson.edu/all_theses/3652

Weiss, D., & Freund, A. M. (2012). Still young at heart: Negative age-related information motivates distancing from same-aged people. *Psychology and Aging, 27*(1), 173-180. https://doi.org/10.1037/a0024819

Zajonc, R. B. (2001). Mere exposure: A gateway to the subliminal. *Current Directions in Psychological Science, 10*(6), 224–228. https://doi.org/10.1111/1467-8721.00154

Zhou, S., Bickmore, T., Paasche-Orlow, M., & Jack, B. (2014). Agent-user concordance and satisfaction with a virtual hospital discharge nurse. In *Proceedings of the 2014 International Conference on Intelligent Virtual Agents (IVA'2014)*, Boston, MA, USA, 528–541. https://doi.org/10.1007/978-3-319-09767-1_63

APPENDIX

Appendix A

Refined Human-Computer Trust (RHCT) scale (Gulati et al., 2019)

1. I believe that there could be negative consequences when using this voice assistant.
2. I feel I must be cautious when using this voice assistant.
3. It is risky to interact with this voice assistant.
4. I believe that this voice assistant will act in my best interest.
5. I believe that this voice assistant will do its best to help me if I need help.
6. I believe that this voice assistant is interested in understanding my needs and preferences.
7. I think that this voice assistant is competent and effective in diagnosing symptoms and providing triage recommendations.
8. I think that this voice assistant performs its role as a diagnostic and triage recommendation tool very well.
9. I believe that this voice assistant has all the functionalities I would expect from a diagnostic and triage recommendation tool.
10. If I use this voice assistant, I think I would be able to depend on it completely.
11. I can always rely on this voice assistant for diagnosing symptoms and providing triage recommendations.
12. I can trust the information presented to me by this voice assistant.