Creepy Assistant: Development and Validation of a Scale to Measure the Perceived Creepiness of Voice Assistants

Rachel Phinnemore
University of Toronto, Canada
rphinnemore@dgp.toronto.edu

Mohi Reza
University of Toronto, Canada
mohireza@cs.toronto.edu

Blaine Lewis
University of Toronto, Canada
blaine@dgp.toronto.edu

Karthik Mahadevan
University of Toronto, Canada
karthikm@dgp.toronto.edu

Bryan Wang
University of Toronto, Canada
bryanw@dgp.toronto.edu

Michelle Annett
MishMashMakers, Canada
michelle@mishmashmakers.com

Daniel Wigdor
University of Toronto, Canada
daniel@dgp.toronto.edu

ABSTRACT

Voice assistants have afforded users rich interaction opportunities to access information and issue commands in a variety of contexts. However, some users feel uneasy or creeped out by voice assistants, leading to a decreased desire to use them. As there has yet to be a comprehensive understanding of the factors that cause users to perceive voice assistants as being creepy, this research developed an empirical scale to measure the creepiness inherent in various voice assistants. Utilizing prior scale creation methodologies, a 7-item Perceived Creepiness of Voice Assistants Scale (PCAS) was created and validated. The scale measures how creepy a new voice assistant would be for users of voice assistants. The scale was developed to ensure that researchers and designers can evaluate the next generation of voice assistants before such voice assistants are released to the wider public.

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**.

KEYWORDS

creepiness; voice assistants; evaluation; empirical scale; questionnaire; perceived creepiness of voice assistants

ACM Reference Format:
Rachel Phinnemore, Mohi Reza, Blaine Lewis, Karthik Mahadevan, Bryan Wang, Michelle Annett, and Daniel Wigdor. 2023. Creepy Assistant: Development and Validation of a Scale to Measure the Perceived Creepiness of Voice Assistants. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3544548.3581346

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

CHI '23, April 23–28, 2023, Hamburg, Germany

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3581346

1 INTRODUCTION

From HAL 900 in Space Odessey to J.A.R.V.I.S in Iron Man, voice assistants have long captured popular imagination, and raised public concern due to their perception of creepiness. Today, voice assistants are widely embedded in smartphones, smart speakers, cars, and Internet of Things devices. In 2020, there were 4.2 billion voice assistant-enabled devices in use, and this number is projected to grow to 8.4 billion by 2024 [39]. Voice assistants enable handsfree interaction, which has made them valuable in several contexts such as while cooking [74] or driving [41, 61]. They also lessen the cognitive load needed to track daily tasks and make it easier for users to retrieve information [68]. Beyond increased efficiency, voice assistants have also been found to reduce depression, stimulate positive emotions, and generate greater interest in engaging in physical activity [53].

The benefits that voice assistants provide to users will be curtailed, however, if they, like other new technologies (e.g., robots or self-driving cars), lead users to perceive them as being creepy. Since their introduction, there has been much discourse about the creepiness of voice assistants. For example, an Ask Reddit thread soliciting creepy Alexa/Google Home stories generated 3.8k upvotes and 1.8k comments [6]. Notable media outlets including the New York Times [69], The Economist [20], and Rolling Stone magazine [18], have also reported on the creepy nature of voice assistants. The Economist, for example, described the lack of privacy inherent in voice assistants in smart speakers, noting that *"Using [them] is like casting a spell ... This hands-free convenience has a cost: the speakers are constantly listening out for commands"*. Only addressing privacy concerns, however, is not sufficient, as recent research has found that broader perceptions of creepiness mediate privacy concerns in intelligent personal assistants [29]. While current voice assistants support remote control-type functionality (e.g., "turn my light on"), improved speech synthesis and language modelling techniques will enable future voice assistants to support even more complex tasks [77]. However, advances in voice assistant functionality come at the risk of introducing perceptions of creepiness, as the development of synthesized voices [36] and error correction techniques [17] have led to user perceptions of creepiness.

Although there have been a few attempts in the research literature to create scales that enable developers to empirically measure creepiness, including the creepiness of technology [80] and creepiness of situations [37], these scales do not capture the nuances of voice assistant interfaces. For example, voice assistants differ from GUI or text based technologies due to their “always on nature", the increased difficulties that arise while trying to correct errors, their susceptibility to noise, and the anthropomorphic effects that result from the voice tones, gender, and personalities that they use and social roles they adopt.

Thus, similar to how Zwakman et al. developed the Voice Usability Scale (VUS) [84] to supplant the use of the System Usability Scale [12] for voice-based interfaces, the present research developed the *Perceived Creepy Assistant Scale (PCAS)*, to enable designers to assess the factors that impact the creepiness of voice assistants. The scale was developed following Boateng et al.'s scale development methodology [9], which is comprised of a review of relevant literature to generate initial scale items, the creation and refinement of a tentative scale via an Exploratory Factor Analysis, and the validation of the final scale items via three user surveys. This led to the following contributions:

- The PCAS scale, which enables developers and designers to measure the perceived creepiness of voice assistants, and thus ensure that newly developed voice assistants do not induce perceptions of creepiness.
- The identification of four novel factors that influence the perception of creepiness in voice assistants, i.e., control, privacy, behavior, and value.
- Design guidelines that stem directly from the PCAS scale items and should enable designers and developers to circumvent the introduction of creepiness in the voice assistants they are creating.

2 RELATED WORK

Of most relevance to the development of the PCAS scale was research that sought to understand the construct of creepiness, in addition to measurement tools that could be used to identify creepiness, and research relating to the identification of the unique attributes inherent in voice assistants.

2.1 Understanding the Construct of Creepiness

While creepiness is an understudied construct [19], it has begun to attract greater research interest due to the increased prevalence of new technologies within our day-to-day lives. The first empirical investigation into creepiness as a psychological construct found that unpredictability contributes to feelings of creepiness [44]. Further work by Watt et al. found that creepiness was associated with unusual physical appearances and socially unacceptable behaviors [75]. Creepiness has been further defined using the Russel Circumference Model [71], which presented creepy recommendations to users while measuring their emotional responses. The creepy recommendations elicited responses characterized by high arousal and low valence which positioned creepiness next to negative constructs such as fear, nervousness, annoyance, frustration, and distress. This research underscores the importance of ensuring that experiences with voice assistants do not induce perceptions of creepiness, given creepiness' close relation to these negative emotions.

Creepiness has also been discussed in the context of Altman's idea of personal space and Nissenbaum's theory of contextual integrity [63]. Using these frameworks, Shklovski et al. identified that creepiness arises due to a lack of "contextual integrity" that breaches social norms, actions, or products and infringes or limits control with overly sensitive privacy boundaries. The impact of privacy concerns on creepiness was echoed in research on chatbots and mobile app data privacy settings, which found that privacy concerns drove perceptions of creepiness [55, 83]. In addition, in research on mobile app data privacy settings, a model of creepiness was presented where creepiness had a negative relationship to perceived control, a positive relationship to privacy concerns, and a negative relationship to disclosure comfort [83]. More recent work by Seberger et al. [60] on creepiness in mobile apps has shown that creepiness is an aspect of affective discomfort.

Creepiness has also been described as a result of the introduction of new technologies [70]. In this context, the cause of creepiness was due to using data or products in unexpected ways, pushing against social norms, or exposing the misalignment between user and corporate interests. In work by Wissinger et al., when users were introduced to new technologies that induced perceptions of creepiness, users evaluated the technologies in terms of a creepiness vs convenience trade-off [78]. If the convenience of a technology outweighed its creepiness, users would continue to use the technology, demonstrating the role that value played in their technology adoption. Contrasting this perspective, research has also found that creepiness was rooted in the dispositions of an observer to have higher levels of discomfort with ambiguity [19]. The subjectivity of creepiness based on the observer was also supported by Smith et al. who found that gender differences impacted the emotional response of participants to images of a creepy male face and creepy female face [65]. Finally, perceptions of creepiness have been studied in children [11, 81]. Yip et al. uncovered five factors that contributed to children's perception of creepiness: deception, lack of control, mimicry, ominous physical appearance, and unpredictability [81]. Within the present research, we were motivated by findings that unpredictability, privacy concerns, and violating social norms could lead to perceptions of creepiness. While prior research has provided a critical foundation to understand creepiness, the community lacks measurement tools to assess creepiness within the domain of voice assistants.

2.2 Measuring Creepiness

Two tools have been developed to evaluate creepiness. The Creepiness of Situation Scale (CRoSS) was developed to measure the creepiness of situations in a broad context, including everyday situations and new technologies [37]. With this scale, creepiness was represented by the two-factor constructs of *emotional creepiness* (i.e., the affective response that results from unpredictable situations) and *creepy ambiguity* (i.e., the lack of clarity about how to respond to such situations). More recently, Woźniak et al. investigated the factors that contributed to initial feelings of creepiness with new technologies and created the 8-item Perceived Creepy Technology Scale (PCTS) [80]. The PCTS identified how factors including *implied malice* (i.e., perceived bad intentions), *undesirability* (i.e., the feeling of unease due to inappropriate contexts), and *unpredictability* (i.e., an ability to predict the actions of a technology or having a lack of control over a technology) all contributed to the creepiness of new technologies.

These two scales identified multiple factors that define creepiness within the context of a broad range of technologies and situations. The CRoSS was developed using a video scenario of a person having difficulty with their computer and receiving a call immediately from a stranger offering help to fix their computer. The PCTS was developed using prototypes of wearables and IoT devices connected to a mobile device or computer (e.g., Fitbit Flex 22). While the PCTS mentioned that voice assistants were creepy, neither scale included voice assistants or voice-based technologies as an apparatus in the user-facing stages of scale development (i.e., focus groups, exploratory factor analysis, or scale evaluation). Although the CRoSS and PCTS have been instrumental in furthering our understanding of creepiness at the more general levels of situations and technology, voice assistants have unique characteristics that make them vulnerable to greater perceptions of creepiness (e.g., anthropomorphic effects, increased privacy concerns). As the literature is missing a cohesive exploration of how these and other factors can be combined into a single scale to enable researchers and designers to assess voice assistants for perceived creepiness, the present work seeks to fill this gap.

2.3 Voice Assistant User Experiences

Murad et al. argued that interacting with a voice assistant is notably different from interacting with a graphical user interface (GUI) [51]. GUIs visually present most options to a user, while with voice assistants, options must be specifically requested (unless erroneous input is detected), which leads to discoverability issues [13]. In contrast to GUIs, voice assistants also possess human-like characteristics, often taking on personalities [10], a name, or a gender. These attributes may contribute to users assigning anthropomorphic characteristics to voice assistants that are implemented within a conversational agent, computer, or media [41, 56] or having emotional interactions with voice assistants [72]. Voice assistants are also unique in their expansive nature as they become integrated in more devices in people's homes (i.e., Siri in Roomba devices [49]) and serve as "central control" devices in smart home ecosystems [25]. This integration leads to greater opportunities for data collection and greater privacy threats. Beyond functional differences, users have also been found to perceive voice assistants differently, finding them to be more personal, smarter, and more efficient than GUIs [42]. As these generalized scales for creepiness do not capture the unique attributes of voice assistants, such as the anthropomorphic attributes users ascribe to them and greater privacy risks, a new scale is needed to evaluate the creepiness of voice assistants.

As voice assistants continue to become more functional and advanced, they do so at the risk of inducing perceptions of creepiness in users. Perceptions of creepiness have been found in qualitative user feedback about the advanced functionality of voice assistants today such as whispering and self-correcting conversational dialogues [17, 52]. Exploratory research investigating futuristic possibilities for voice assistants identified the desire for voice assistants to be more proactive, personalized, and capable of serving multiple roles of a tool, assistant, and friend [73]. While this functionality is beyond the capabilities of voice assistants today, as we continue to develop voice assistant technology that realizes this vision for voice assistants, it is important to do so in a manner that circumvents the introductions of perceptions of creepiness.

3 SCALE DEVELOPMENT METHODOLOGY

The development of the Perceived Creepy Assistant Scale (PCAS) followed the process outlined by Boateng et al. [9], which consisted of three phases (Figure 1): (1) Construct Definition and Item Development, (2) Scale Development, and (3) Scale Evaluation.

As part of the Construct Definition and Item Development phase, we defined the domain of the scale construct and generated an initial series of scale items through a literature review and expert interviews. This process ensured that the initial formulation of the scale was well-founded. Next, in the Scale Development phase, we identified the factor structure of the scale and reduced the number of scale items through an Exploratory Factor Analysis. This retained the most pertinent items. Finally, during the Scale Evaluation phase, we confirmed the previously identified factor structure through a test of dimensionality using a Confirmatory Factor Analysis, assessed the validity and reliability of the scale with a new sample population via a Differentiation by Known Groups evaluation, assessed how similar the scale was to existing scales via Convergent Validity testing, and ensured that scale results would hold over time using a Test-Retest Reliability methodology. Through these three validation steps, a 7-item Perceived Creepy Assistant Scale (PCAS) was created and validated to ensure that it could measure the perceptions of creepiness in voice assistants.

4 PHASE 1: CONSTRUCT DEFINITION AND ITEM GENERATION

Generating a scale typically involves characterizing the construct to be studied (i.e., creepiness) and developing a list of items that represent it. As the goal of the Perceived Creepy Assistant Scale was to measure initial user perceptions of creepiness in voice assistants, the following definitions were adopted to concretize the scale development process:

- **Creepiness**: *"a potentially negative and uncomfortable emotional response paired with perceptions of ambiguity toward a person, technology or even during a situation"* [37].
- **Voice Assistant**: *"an artificial intelligence-powered computer system that aims to imitate human intelligence while engaging in realistic conversations with users"* [21].

The scale development process focused on ensuring that the resulting scale would capture the facets of voice assistants that lead to them being perceived as creepy. As an abundance of research has explored the complexity of voice-specific factors such as inflection, pitch, tone [23, 34, 38]. Many voice assistants have a variety of voices that can be used. Thus, the investigation into the role of voice-specific factors may have on the perceptions of creepiness in voice assistants was left for future work.

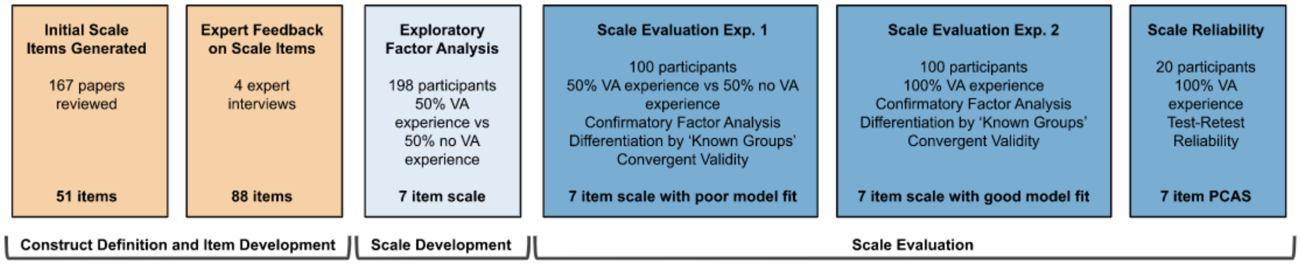

Initial Scale Items Generated
167 papers reviewed
51 items
Expert Feedback on Scale Items
4 expert interviews
88 items
Exploratory Factor Analysis
198 participants 50% VA experience vs 50% no VA experience
7 item scale
Scale Evaluation Exp. 1
100 participants 50% VA experience vs 50% no VA experience
Confirmatory Factor Analysis
Differentiation by 'Known Groups'
Convergent Validity
7 item scale with poor model fit
Scale Evaluation Exp. 2
100 participants 100% VA experience
Confirmatory Factor Analysis
Differentiation by 'Known Groups'
Convergent Validity
7 item scale with good model fit
Scale Reliability
20 participants 100% VA experience
Test-Retest Reliability
7 item PCAS
Construct Definition and Item Development
Scale Development
Scale Evaluation

Figure 1: Overview of the scale development process used to produce the PCAS, which was based on Boateng et al.'s scale development process [9].

According to Cronbach, content validity can be achieved by generating representative items from a universal pool [16]. As such, a literature review was conducted using the ACM Digital Library and Google Scholar to identify possible scale items. A query on these services using the search phrases "creepy" and "voice assistant" returned 343 papers. Papers that used "creep" to describe social media stalking and "creep" as a verb rather than as an adjective were removed, resulting in 167 papers remaining in the dataset.

Two researchers reviewed the 167 papers and independently extracted quotes that mentioned creepiness by searching for the term. They also identified sections that discussed causes of creepiness but did not directly use the term "creepy". Using this process, a corpus of 381 quotes was extracted. The extracted quotes covered a range of topics including the study of creepiness itself [80, 81], privacy issues with smart speakers and smart home devices [1, 24], and the uncanny valley of social robots [22, 30]. For example, an extracted quote from a study on children's perceptions of creepiness highlighted the role of ambiguous answers on creepiness, i.e., *"children noted that broad, non-specific answers to difficult questions made the technology appear creepy because they projected ambiguity"* [81].

Three researchers open-coded the extracted quotes to identify the facets of voice assistants that lead to them being perceived as creepy. To develop the initial codes and codebook to classify the extracted quotes, the researchers open coded a sample of 20% of the extracted quotes and then discussed the resulting codes to refine them. Two additional researchers then used the 19 refined codes to code 15% of the quotes. The inter-rater reliability score (i.e., Cohen's Kappa) when coding this subset of quotes was 67%, indicating that there was substantial agreement between the coders [45]. The two coders then coded the remaining 85% of the quotes. After the coding was complete, the quotes were grouped by code and the extracted quotes were used to generate unique scale items for each code, resulting in 51 initial scale items. As several of the codes were semantically similar, the codes were then aggregated into factors (e.g., *uncanny valley appearance* and *physical appearance* were grouped into a single *Aesthetics* factor). Nine factors resulted from this process (i.e., *aesthetics, behavior, control, intention, privacy, transparency, trust, value,* and *other*; Appendix A).

4.1 Expert Feedback

As per Boateng et al.'s recommendation [9], four experts were recruited to provide feedback on the initial list of scale items. The experts had extensive experience with voice assistants, creepiness in technology, or next-generation user interface design (Table 1). During interviews, each expert was shown the initial list of scale items and was asked to provide feedback on the items and factors and suggest any that they perceived to be missing.

The experts highlighted several novel facets of creepiness in voice assistants that were important to consider, further supporting the need for a new scale specific to voice assistants. Two experts mentioned *control* as being one of the most important factors. For instance, E1 stated, *"A lot of people want reactive control – e.g., stop doing that, but don't want proactive control"*. Two experts also felt that device *behavior* was an important factor. E4 said that the *"behaviour of voice assistants is conversations and those can be clunky and weird e.g., interruption, ask stupid questions, not understand what the person is saying"*.

This feedback, in addition to other comments from the experts, was used to refine the list of scale items. Specifically, the feedback from the experts led to 2 items being deleted, 7 items being refined, 39 new items being added, resulting in a total of 88 scale items. They also recommended assigning the *Transparency* factor's scale items to the *Privacy* and *Behaviour* factors, which resulted in a reduction in the total number of factors (e.g., from 9 to 8; Appendix B).

5 PHASE 2: SCALE DEVELOPMENT

To further refine the list of 88 scale items, we then designed and conducted an online survey using Qualtrics XM (Appendix B) to conduct an Exploratory Factor Analysis. The purpose of the Exploratory Factor Analysis was to assess which of the 88 initial items were valid measures of initial perceptions of creepiness in voice assistants and develop an initial factor structure for the scale. We targeted a sample size of 100-200 participants based on prior recommendations for validating empirical scales [8, 43, 48]. [1]

[1] Prior to conducting all of the studies, we received ethics approval from our institutional IRB.

Table 1: Overview of Experts Consulted about the Initial Scale Items.

Expert	Role	Expertise	Experience
1	Assistant Professor in HCI	Creepy technology	14 years
2	Director of Design at Tech Company	Voice assistants	25+ years
3	Adjunct Faculty in HCI	Interaction design	25+ years
4	Research Engineer at Tech Company	Voice assistants	17 years

5.1 Participants

One hundred and ninety-eight participants were recruited through Amazon Mechanical Turk to complete the survey. Participants were reimbursed $4 USD for completing the 20 minute survey. To be eligible to complete the survey, participants needed to be located in North America or the European Economic Union and be over the age of 19. To ensure high quality responses, based on recommended practices in HCI, participants were required to have completed over 1000 HITs on Turk and have a 95% successful completion rate for HITs [26, 27]. We focused on these regions because they have the highest penetration rates for smart speaker adoption [46].

The participant pool was selected such that 50% of the pool (n = 99) had experience with voice assistants and the other 50% of the pool (n = 99) did not. Participant experience with voice assistants was defined as anyone who had issued a command to a voice assistant. Among the voice assistant experience group, participants' average age was M = 39 (SD = 10 years) with ages ranging from 24 to 79. Within the voice assistant experience group, 69 participants identified as male, and 30 participants identified as female. Among the group without voice assistant experience, participants' average age was M = 38 (SD = 13 years) with ages ranging from 20 to 104. Within the voice assistant experience group, 66 participants identified as male, 31 participants identified as female, 1 participant identified as non-binary, and 1 participant did not disclose their gender identity. One hundred and thirty-two participants were from North America and 66 were from the European Economic Union.

5.2 Study Design

A 2x3 between-subject experimental design was used, wherein participants with and without prior experience using voice assistants were recruited and 3 creepy voice assistant scenarios were developed to evaluate the survey items. Participants were evenly split across the 2x3 experimental design, resulting in 33 participants per condition.

5.3 Survey

Participants were randomly assigned to read one of three scenarios that described a voice assistant that varied in the level of creepiness (Appendix C). The scenarios were designed based on the factors and items resulting from the literature review and expert feedback processes. Prior to running the study, a small pilot study was run to ensure that the scenarios had varying degrees of creepiness. Five participants recruited from our institution were asked to read through each scenario and rate how creepy each scenario was using a 5 point Likert scale. Scenario 1 received a mean creepiness rating of 3.17 (SD = 0.69), Scenario 2 was rated 3.29 (SD = 1.03), and Scenario 3 was rated 4.00 (SD = 0.63). The purpose of this pilot was not to empirically validate how creepy the scenarios were, but rather to show a trend that Scenario 3 was creepier than Scenario 2, and both were creepier than Scenario 1.

After reading one of the three scenarios, participants answered 88 7-point Likert scale items with the anchors "Strongly Disagree" (1) and "Strongly Agree" (7). Participants were asked two attention check questions. Those who failed to answer these questions correctly were excluded from the analysis (i.e., 4 participants).

5.4 Exploratory Factor Analysis

The Exploratory Factor Analysis approach that was used by Mejia and Yarosh [48], which employed a varimax rotation, was replicated to understand the factors implicated in the creepiness of voice assistants, as well as an analysis of the scree plots. The results and plots identified a three-factor model. We then reduced low loading items (i.e., below 0.40) [9], as well as items that loaded onto multiple factors. Thus, the finalized scale was a unidimensional scale with 7-items. We further refined the items by optimizing the inter-item reliability using Cronbach's alpha, which resulted in an alpha of 0.903. The scale also had acceptable factor model fit parameters, with a Root Mean Square Error of Approximation of 0.04 and a Tucker Lewis Index (TLI) of 0.99.

5.5 The Resulting PCAS

The final Perceived Creepy Assistant Scale (PCAS) consisted of 7-items that captured the aspects that cause a voice assistant be perceived as creepy (Table 2). When using the PCAS, respondents would complete the PCAS scale items using a 7 point Likert scale with anchors Strongly Disagree (1) to Strongly Agree (7) for each survey item. The PCAS score would then be generated by summing the response to each item on the scale, with a higher score indicating a greater level of creepiness.

All items in the PCAS came from previously identified factors about creepiness (but not creepiness with voice assistants) in the literature review: *control, privacy, intention, behavior,* and *value*, which can be seen in Appendix B. None of the items from the literature review categories of *transparency*, and *aesthetics* or *trust* were represented in the final PCAS because these items were eliminated based on the expert feedback or the Exploratory Factor Analysis, respectively, following the recommended process of Boateng et al. [9].

6 PHASE 3: SCALE EVALUATION

After determining the PCAS's theoretical structure, we proceeded to evaluate the structure of the PCAS using Confirmatory Factor Analysis [9]. Subsequently, we assessed the construct validity of the scale through two experiments. In Experiment 1, we recruited

Table 2: The PCAS is a unidimensional scale with seven items. The factor loadings for the items and Cronbach's Alpha for the scale were calculated using the results from the Scale Development Survey.

Scale Items	Loading
Q1: I have minimal control when I use this voice assistant.	0.766
Q2: This voice assistant does things that are not in my best interest.	0.814
Q3: This voice assistant behaves in deceptive ways.	0.777
Q4: This voice assistant could be accidentally or unintentionally harmful towards users.	0.854
Q5: This voice assistant is collecting too much data about me.	0.774
Q6: The way this voice assistant behaves doesn't follow social norms.	0.787
Q7: This voice assistant does not provide enough benefits to me to justify me using this voice assistant.	0.791
Scale Overall	$\alpha = 0.903$

100 participants with an even split of participants with and without voice assistant experience. However, during this experiment, we found that the PCAS had a poor model fit for participants without voice assistant experience, as these participants found all voice assistants creepy, including the non-creepy one. Thus, in Experiment 2, we recruited 100 participants with voice assistant experience and validated the PCAS as a measurement tool for initial perceptions of creepiness in voice assistants with people who had prior experience with voice assistants.

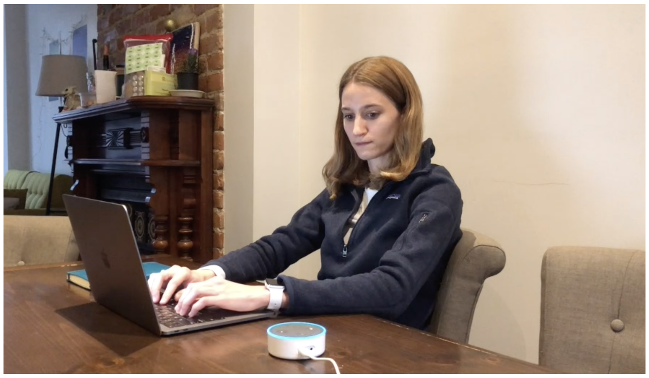


Figure 2: Screenshot from a scene in one of the voice assistant interaction videos.

6.1 Evaluation Stimuli

To validate the PCAS in Experiments 1 and 2, we followed the process used by Woźniak et al. [80] and created two videos depicting a person interacting with a voice assistant (Figure 2). The first video portrayed a creepy futuristic voice assistant. To ensure that the creepy video was creepy, the creepy video voice assistant used self-correction and whispering out of context, which prior work has found to be creepy [17, 52]. The second video portrayed a voice assistant that mimicked the functionality that might be expected from today's voice assistants, e.g., asking the voice assistant to play music. For consistency, both scripts were written to follow the same plotline and included aspects of the scale items to evaluate each of the seven scale items (Appendix D).

6.2 Experiment 1: PCAS Validation with Users with and without Voice Assistant Experience

In Experiment 1, we conducted an online survey with 100 participants to determine the validity of the PCAS' factor structure using Confirmatory Factor Analysis, assess its ability to Differentiate between 'Known Groups', and establish Convergent Validity with related constructs to determine if the scale measures perceptions of creepiness.

6.2.1 Participants. One hundred participants were recruited using Amazon Mechanical Turk to complete the online study. Participants watched a short video of creepy voice assistant interaction or non-creepy voice assistant interaction and then completed the PCAS, PCTS [80], and VUS [84]. Respondents were reimbursed $2.00 USD for participating.

To ensure that the PCAS was valid for all participants, 50% of the participants who were recruited had experience with voice assistants and 50% had no experience with voice assistants. Participant experience with voice assistants was defined as anyone who had issued a voice command to a voice assistant. A definition of a voice assistant and its functionality was provided to ensure participants' understanding. Among the group with voice assistant experience, the average age was M = 35 years (SD = 10 years), with a range of 21 to 62. Within the voice assistant group, 37 participants identified as male, 12 participants identified as female, and 1 participant identified as non-binary. Among the group with no voice assistant experience, the average age was M = 37 years (SD = 11 years) with a range of 21 to 79. Within the voice assistant group, 30 participants identified as male and 20 identified as female.

Fifty participants were recruited from the European Economic Union and fifty participants were recruited from North America. Nine participants were replaced after being identified as straight-liners (i.e., participants who give near identical answers to all the survey items [32]. Fifty percent of participants, evenly split between the experience groupings, watched the creepy voice assistant video and fifty percent of participants watched the non-creepy voice assistant video.

6.2.2 Confirmatory Factor Analysis. As recommended by Boateng et al. [9] and recent practices in HCI to validate scales [8, 80], we evaluated the dimensionality of the PCAS by performing a Confirmatory Factor Analysis (CFA). The results of the CFA produced a TLI score of 0.88. Based on recommendations from

Bentler and Bonnett [7], that models with scores less than 0.9 are inadequate, the PCAS had a poor model fit and should be improved [9]. Therefore, this experimentation did not validate the model. We subsequently validated the PCAS through Experiment 2 by recruiting participants with voice assistant experience.

6.2.3 *Differentiation by 'Known Groups'.*

To establish the construct validity of the PCAS, we conducted tests using 'known groups' [9], which showed that the PCAS can be used to discriminate between creepy vs non-creepy voice assistants. We compared the mean PCAS scores of participants who viewed the creepy vs non-creepy voice assistant interaction videos. A Shapiro-Wilk test revealed that the data was not normally distributed, so the Mann-Whitney U test was used. The results found significant differences in mean PCAS scores between participants who viewed the creepy and non-creepy voice assistant video (Table 3), indicating that the PCAS can be used to differentiate between creepy vs non-creepy voice assistants.

6.2.4 *Convergent Validity Testing.*

Based on the recommendations of Boateng et al. [9], we assessed whether the PCAS had correlations with other relevant constructs to show that the scale had convergent validity. As Woźniak et al. demonstrated that the perceived creepiness of new technologies is a distinct concept from technology acceptance through discriminant validity testing, we did not conduct discriminant validity testing with the PCAS [80].

Thus, we first compared the PCAS with the Perceived Creepy Technology Scale (PCTS) [80], given that they both measure the construct of creepiness. Using the Spearman Rho calculation, we found the PCAS and PCTS had a statistically strong correlation, $\rho = 0.740, p < 0.001$. Thus, the PCAS measured creepiness because it converged with the PCTS.

In prior work with chatbots, Rajaobelina et al. [55] found a negative relationship between creepiness and usability where higher perceived usability reduced perceptions of creepiness. To demonstrate that this negative relationship exists with voice assistants, we compared the PCAS with the Voice Usability Scale (VUS) [84]. We found a significant negative correlation between the PCAS and VUS, $\rho = -0.772, p < 0.001$. This shows that higher perceptions of creepiness were correlated with reduced perceptions of usability in voice assistants.

Thus, the PCAS presented new dimensions to measure perceived initial creepiness in voice assistants while demonstrating convergence with the related constructs of creepiness more generally (i.e., PCTS) and voice assistant usability (i.e., VUS).

6.2.5 *Discussion.*

The goal of this experiment was to validate the PCAS to ensure that it would measure the initial perceptions of creepiness in newly developed voice assistants.

During the Convergence Validity testing, the PCAS achieved convergence with the PCTS, demonstrating that the PCAS measures the construct of creepiness. Additionally, the PCAS achieved convergence with the VUS through a negative correlation, supporting Rajaobelina et al.'s prior hypothesis of a negative relationship between creepiness and usability [55]. While the convergence of the PCAS and the VUS could have indicated that the VUS could be used to assess the creepiness of voice assistants (i.e., a low score on the VUS could be used to detect creepiness), a low VUS score could be indicative of usability issues or perceived creepiness. Thus, the negative relationship between creepiness and usability indicates that the PCAS is a useful scale to measure creepiness, especially in light of its impact on usability. This line of testing thus partially validated the PCAS and, if the other two validations were successful, could have provided designers and developers with confidence that the PCAS could measure the construct of creepiness in voice assistants.

The Differentiation by 'Known Groups' testing results showed that the PCAS could be used by designers to discriminate between creepy and non-creepy voice assistants. In these results, a notable difference was observed in the mean PCAS scores of those with voice assistant experience (PCAS = 21.92) and those without voice assistant experience (PCAS = 26.00) when viewing the non-creepy video condition (Table 3). Thus, participants without voice assistant experience found current voice assistants, as depicted in the non-creepy video, to be significantly creepier that participants with voice assistant experience.

Provided that the remaining validation test was successful, the results from these two evaluations could have instilled confidence in designers that the PCAS could distinguish between different types of voice assistants and provide them with insights into whether their voice assistant products are perceived as creepy. However, the Confirmatory Factor Analysis results did not achieve sufficient model fit to validate the model's factor when used with a new dataset. When poor CFA results are found, a common practice is to remove underperforming scale items, i.e., those with low factor loadings [9], however, all of the PCAS scale items had high factor loadings (i.e., greater than 0.70) and the scale had Cronbach's alpha of 0.916. Thus, we were not able to identify underperforming scale items to remove from the scale.

We hypothesized that the Confirmatory Factor Analysis did not achieve an acceptable model fit due to the presence of the participants without voice assistant experience. These participants had higher mean scores in the non-creepy video condition, suggesting that they had a propensity to view all voice assistants as creepy even if they are not. Although we included participants with varying levels of voice assistant experiences to ensure the scale could be administered to a variety of potential users, we excluded all 50 participants without voice assistant experience and re-ran the Confirmatory Factor Analysis to test whether the model was validated for users with voice assistant experience. We found an acceptable TLI score of 0.94 [9]. Thus, to ensure that the PCAS could be fully validated for users with voice assistant experience, we conducted a second experiment with this group in Experiment 2 outlined below. The results from Experiment 2 led to the validation of the PCAS for voice assistant users.

6.3 Experiment 2: PCAS Validation with Users with Voice Assistant Experience

Because the non-users in Experiment 1 found all voice assistants to be creepy, in Experiment 2, we only recruited participants with voice assistant experience to complete the same online survey. As in Experiment 1, we evaluated the construct validity of the PCAS, using a method similar to Experiment 1, i.e., Confirmatory Factor Analysis, Differentiation by 'Known Groups', and Convergence Validity testing. As the results of these three tests validated the

Table 3: Experiment 1 Construct Validity Assessment using non-parametric tests for Differentiation by Known Groups. Bonferroni adjusted p values are reported. Participants with voice assistant experience are denoted by 'VA' and participants without voice assistant experience are denoted by 'NVA'.

Scale	M_{Creepy}	SD_{Creepy}	$M_{Non\text{-}Creepy}$	$SD_{Non\text{-}Creepy}$	U	p
PCAS (VA)	34.28	8.76	21.92	9.18	104.50	<0.001
PCAS (NVA)	36.00	7.53	26.00	8.98	126.50	<0.001

PCAS, we also conducted a Test-Retest Reliability evaluation to determine if the PCAS would provide the same results over time (as per Boateng et al. [9]).

6.3.1 Participants. The experiment was identical to Experiment 1, except that we added the requirement that all participants had to have experience with voice assistants (i.e., issued a command to one in the past). One hundred participants with voice assistant experience were recruited using Amazon Mechanical Turk to complete the online study. Participants were reimbursed $2.00 USD for completing the survey. Participants followed the same procedure used in Experiment 1, watching a creepy or non-creepy video and completing the PCAS, PCTS, and VUS. The average age was 38 years (SD = 10 years, range = 20 to 78 years). Sixty-three participants identified as male and 37 participants identified as female. Fifty participants were recruited from North America and fifty were recruited from the European Economic Union to provide cultural diversity. Twelve participants were removed from the analysis after being identified as straight-liners.

6.3.2 Confirmatory Factor Analysis. To validate the PCAS for participants with voice assistant experience, we conducted a Confirmatory Factor Analysis that achieved an acceptable model fit with a TLI of 0.97 [9]. Unlike the results from the first experiment, this test of dimensionality using a new sample population indicated a good model fit and validated the scale structure of the PCAS, thus positively contributing to the overall validation of the PCAS.

6.3.3 Differentiation by 'Known Groups'. Next, we conducted a test of Differentiation using 'Known Groups' [9] to establish the construct validity of the PCAS. Again, the Shapiro-Wilk test revealed that the data was not normally distributed, so the Mann-Whitney U test was used. The results demonstrated a significant difference between the creepy vs non-creepy video (Table 4). These results also partially validated the PCAS and indicated that the PCAS was able to distinguish between creepy versus non-creepy voice assistants.

6.3.4 Convergent Validity Testing. Using the same method as in Experiment 1, we evaluated the Convergent Validity of the PCAS compared to the PCTS and VUS. Similar to Experiment 1, we found a strong correlation between the PCAS and PCTS with a $\rho = 0.676, p < 0.001$. Thus, this established Convergent Validity and indicated that the PCAS and PCTS both measured the construct of creepiness. Additionally, we found a negative correlation between the PCAS and the VUS with a $\rho = -0.781, p < 0.001$. This established a convergence between creepiness and usability because creepiness and usability have been shown to have a negative relationship [55]. Thus, these results validated that the PCAS identified new items of creepiness in voice assistants via its convergence with PCTS and that it was able to measure voice assistant usability via its convergence with the VUS.

6.4 Test-Retest Reliability

Lastly, following the recommendations of Boateng et al. [9], it was necessary to validate that the PCAS would be able to produce consistent results over time. To do so, we created a survey that asked participants to watch a video of a creepy voice assistant interaction and rate the interaction they saw using the PCAS scale. This survey was then completed by the same group of participants at two different points in time, as per Boateng et al.'s recommendation [9]. This evaluation was not conducted during Experiment 1 because the Confirmatory Factor Analysis did not meet the acceptable threshold to validate the dimensionality of the scale.

6.4.1 Participants. Twenty participants were recruited from a North American university to complete survey (M = 25 years, SD = 3.6 years, range = 20 to 33 years). Ten participants identified as female, eight identified as male, one identified as non-binary, and one preferred not to disclose their gender. The survey was initially administered to participants and then, seven days later, participants were asked to complete the survey again. Participants were required to have prior experience using a voice assistant to complete the surveys. Participants were recruited through an internal mailing list and on social media. One participant was not able to complete the survey the second time, so their survey response was removed from the dataset.

6.4.2 Results. Following the recommendation of Boateng et al. [9], as well as recent practices for calculating test-retest reliability [8, 80], we calculated the intraclass correlation coefficient (ICC) for fixed raters (i.e., *"The variation in measurements taken by an instrument on the same subject under the same conditions. [ICC] is generally indicative of reliability in situations when raters are not involved or rater effect is neglectable, such as [a] self-report survey instrument."* [35]). Based on Koo and Li's guidelines [35], where ICCs between 0.75 and 0.9 demonstrate "good" reliability, the PCAS achieved good reliability $\kappa = 0.90, p < 0.001$, illustrating that the scale would be able to produce consistent results over time.

6.5 Summary

Using Boateng et al.'s [9] three phase validation process, which included tests of Construct Validity (i.e., Confirmatory Factor Analysis, Differentiation by 'Known Groups', and Convergence Validity testing) and Test-Retest Reliability, we validated that the 7-item PCAS measures perceptions of creepiness in voice assistants when administered to users with experience using voice assistants.

Table 4: Experiment 2 Construct Validity Assessment using a non-parametric test for Differentiation by Known Groups. Bonferroni adjusted *p* values are reported. Participants with Voice Assistant experience are denoted by 'VA'.

Scale	M_{Creepy}	SD_{Creepy}	$M_{Non\text{-}Creepy}$	$SD_{Non\text{-}Creepy}$	U	p
PCAS (VA)	36.70	8.13	22.10	9.15	307.00	<0.001

The test of dimensionality, i.e., the Confirmatory Factor Analysis, validated the factor structure of the PCAS. The Differentiation by 'Known Groups' evaluation showed that the PCAS can be used identify and discriminate between creepy and non-creepy voice assistants. The Convergence Validity testing showed that the PCAS was able to converge with constructs related to creepiness (e.g., the PCTS) and had an inverse relation with usability (i.e., the VUS), thus validating that it measures creepiness. Finally, the Test-Retest Reliability evaluation showed that the PCAS measurements would be consistent across different points in time.

Taken together, the results of these four evaluations validated that the PCAS would be able to measure the perceived creepiness of voice assistants when administered to users who had at least a minimal level of prior experience with voice assistants (i.e., they had issued at least one command to a voice assistant in the past).

7 DISCUSSION

In this section, we outline how the PCAS can be scored, interpreted, and used to both prevent creepiness and realize creepiness in one's voice assistant products. Finally, we present design guidelines to accompany the PCAS, and discuss it's limitations.

7.1 Using the PCAS

The PCAS is designed to be used by people with prior experience with voice assistants, defined as anyone who has issued a command to a voice assistant. Each item of the PCAS is scored using a Likert scale, with Strongly Disagree (1) and Strongly Agree (7) anchors. The PCAS score is generated by summing the response to each item on the scale. The PCAS is scored on a range from 7 to 49, where low scores denote a low level of perceived creepiness and high scores denote a high level of perceived creepiness. The PCAS score was designed with ease of calculation in mind to encourage its use.

7.2 Interpreting PCAS Scores

We found the mean PCAS scores for creepiness in Experiment 2 to be 36.70 and the PCAS score for non-creepiness to be 22.10 (Table 4). Future work is needed to investigate whether these thresholds can be refined into more discrete categories (i.e., very creepy, creepy, non-creepy). Since PCAS items are equally weighted, designers can, however, review the scores of individual items in the scale as a guide to learn which items are contributing the most to initial perceptions of creepiness.

The creepiness of voice assistants is a multifaceted construct, one of which is how much a user values the voice assistant. Thus, the acceptable threshold of perceived creepiness may change as users find voice assistants more valuable. In contrast to the factors found by Woźniak et al.'s research into the creepiness of new technologies [80], we found 7 items that comprise a uni-dimensional PCAS scale. Since users have varying degrees of experience with voice assistants and the factors that comprise creepiness are not universal [29], it makes sense that we are not able to distill creepiness into multiple distinct factors.

7.3 Utility of the PCAS

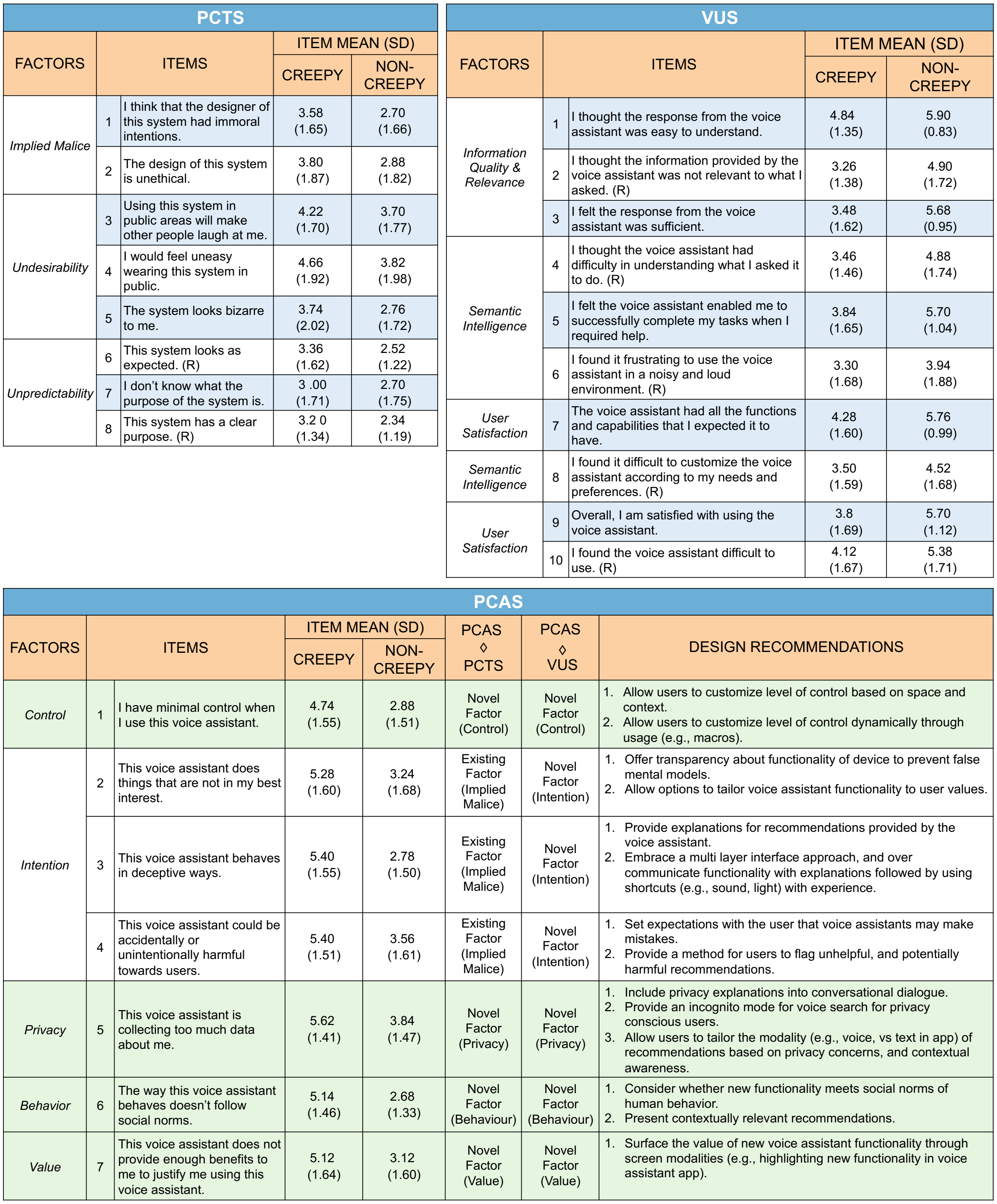
The PCAS is applicable for users with prior experience with voice assistants, however, these users did not need to be experts or frequent users of voice assistants. These users only needed to have issued a command to a voice assistant in the past. This poulation of users represents a significant and growing demographic. In the US in 2022, 61% of the population reported having experience with voice assistants [67], amounting to 202 million people [79]. During this same time period in the UK, 60% of the population (i.e., 40 million people) reported having experience using a voice assistant [67]. Further, voice assistants are expected to grow in usage from 4.2 billion voice assistants being used in 2020 to 8.4 billion in 2024 [39]. Thus, as research and industry continue to develop and make voice assistants more widely available, the demographics of voice assistant users will grow, thus furthering the relevancy and utility of the PCAS as a tool to ensure positive user experiences with voice assistants.

7.4 Differences between the PCAS and Other Scales

The PCAS differs from existing scales as can be seen in (Figure 3). The primary difference between the PCAS and the PCTS, is that the PCAS evaluates four additional factors that capture the creepiness inherent in voice assistants, i.e., control, privacy, behavior and value. In contrast, the PCTS has scale items that probe the role of device aesthetics. The scenarios presented during our studies did not focus on aesthetics because not all voice assistants have a dedicated form factor and some are used in entirely hands free situations (e.g., Microsoft's Cortana or Apple's Siri).

While the PCAS and the PCTS have some similar items in terms of *Intention* (PCAS) and *Implied Malice* (PCTS), the PCAS received higher scores for items relating to the Intention factor, than the PCTS did for items relating to the Implied Malice factor when comparing the Creepy condition in Experiment 2 (Figure 3). Thus, while the PCAS and PCTS both have items that capture the intention of a system, the PCAS measured this with greater granularity.

Although the PCAS and VUS are both focused on voice assistants, none of the scale items or factors are similar between the two scales. This is because the VUS scale items focus on usability elements such as ease of use and user satisfaction, whereas the PCAS focuses on factors related to the creepiness of voice assistants. Following the example set by the PCTS [80], we did not compare the PCAS to the CRoSS as the CRoSS was developed to evaluate situations that involve technology not the underlying technology itself.



PCTS

FACTORS	ITEMS		ITEM MEAN (SD)	
			CREEPY	NON-CREEPY
Implied Malice	1	I think that the designer of this system had immoral intentions.	3.58 (1.65)	2.70 (1.66)
	2	The design of this system is unethical.	3.80 (1.87)	2.88 (1.82)
Undesirability	3	Using this system in public areas will make other people laugh at me.	4.22 (1.70)	3.70 (1.77)
	4	I would feel uneasy wearing this system in public.	4.66 (1.92)	3.82 (1.98)
	5	The system looks bizarre to me.	3.74 (2.02)	2.76 (1.72)
Unpredictability	6	This system looks as expected. (R)	3.36 (1.62)	2.52 (1.22)
	7	I don't know what the purpose of the system is.	3 .00 (1.71)	2.70 (1.75)
	8	This system has a clear purpose. (R)	3.2 0 (1.34)	2.34 (1.19)

VUS

FACTORS	ITEMS		ITEM MEAN (SD)	
			CREEPY	NON-CREEPY
Information Quality & Relevance	1	I thought the response from the voice assistant was easy to understand.	4.84 (1.35)	5.90 (0.83)
	2	I thought the information provided by the voice assistant was not relevant to what I asked. (R)	3.26 (1.38)	4.90 (1.72)
	3	I felt the response from the voice assistant was sufficient.	3.48 (1.62)	5.68 (0.95)
Semantic Intelligence	4	I thought the voice assistant had difficulty in understanding what I asked it to do. (R)	3.46 (1.46)	4.88 (1.74)
	5	I felt the voice assistant enabled me to successfully complete my tasks when I required help.	3.84 (1.65)	5.70 (1.04)
	6	I found it frustrating to use the voice assistant in a noisy and loud environment. (R)	3.30 (1.68)	3.94 (1.88)
User Satisfaction	7	The voice assistant had all the functions and capabilities that I expected it to have.	4.28 (1.60)	5.76 (0.99)
Semantic Intelligence	8	I found it difficult to customize the voice assistant according to my needs and preferences. (R)	3.50 (1.59)	4.52 (1.68)
User Satisfaction	9	Overall, I am satisfied with using the voice assistant.	3.8 (1.69)	5.70 (1.12)
	10	I found the voice assistant difficult to use. (R)	4.12 (1.67)	5.38 (1.71)

PCAS

FACTORS	ITEMS		ITEM MEAN (SD)		PCAS ◊ PCTS	PCAS ◊ VUS	DESIGN RECOMMENDATIONS
			CREEPY	NON-CREEPY			
Control	1	I have minimal control when I use this voice assistant.	4.74 (1.55)	2.88 (1.51)	Novel Factor (Control)	Novel Factor (Control)	1. Allow users to customize level of control based on space and context. 2. Allow users to customize level of control dynamically through usage (e.g., macros).
Intention	2	This voice assistant does things that are not in my best interest.	5.28 (1.60)	3.24 (1.68)	Existing Factor (Implied Malice)	Novel Factor (Intention)	1. Offer transparency about functionality of device to prevent false mental models. 2. Allow options to tailor voice assistant functionality to user values.
	3	This voice assistant behaves in deceptive ways.	5.40 (1.55)	2.78 (1.50)	Existing Factor (Implied Malice)	Novel Factor (Intention)	1. Provide explanations for recommendations provided by the voice assistant. 2. Embrace a multi layer interface approach, and over communicate functionality with explanations followed by using shortcuts (e.g., sound, light) with experience.
	4	This voice assistant could be accidentally or unintentionally harmful towards users.	5.40 (1.51)	3.56 (1.61)	Existing Factor (Implied Malice)	Novel Factor (Intention)	1. Set expectations with the user that voice assistants may make mistakes. 2. Provide a method for users to flag unhelpful, and potentially harmful recommendations.
Privacy	5	This voice assistant is collecting too much data about me.	5.62 (1.41)	3.84 (1.47)	Novel Factor (Privacy)	Novel Factor (Privacy)	1. Include privacy explanations into conversational dialogue. 2. Provide an incognito mode for voice search for privacy conscious users. 3. Allow users to tailor the modality (e.g., voice, vs text in app) of recommendations based on privacy concerns, and contextual awareness.
Behavior	6	The way this voice assistant behaves doesn't follow social norms.	5.14 (1.46)	2.68 (1.33)	Novel Factor (Behaviour)	Novel Factor (Behaviour)	1. Consider whether new functionality meets social norms of human behavior. 2. Present contextually relevant recommendations.
Value	7	This voice assistant does not provide enough benefits to me to justify me using this voice assistant.	5.12 (1.64)	3.12 (1.60)	Novel Factor (Value)	Novel Factor (Value)	1. Surface the value of new voice assistant functionality through screen modalities (e.g., highlighting new functionality in voice assistant app).

Figure 3: Comparison of the Factors and Scale Items from the PCAS, PCTS, and VUS scales. The means and standard deviations are from Experiment 2 for both the Creepy and Non-Creepy video conditions. The green rows indicate novel PCAS factors. R indicates item is reverse scored.

7.5 Design Guidelines: Enabling Targeted Iterative Design to Reduce Creepiness

The development and validation of the PCAS has an additional benefit of providing a method for designers to reduce the creepiness in the voice assistants they design in the future. Designers can use the results to items of the PCAS as general guidelines and tailor their designs accordingly.

7.5.1 Control. Designers should consider ways to ensure that voice assistants provide sufficient control to users, for example, tailoring options and offering different interaction paradigms based on a user's preferred level of control. As the level of control one may wish to have can vary based on the context [28] within which they are accessing a voice assistant, users may feel their control is enhanced if they can dynamically choose between different conversation flows for different environments or tasks. For example, Anmari et al. discussed the need for more granular mechanisms to control voice assistants, such as the ability to control the device differently in different rooms of the house [4].

7.5.2 Intention. To ensure that users feel that voice assistants have their best intentions, voice assistants should be transparent about their functionality [3]. Following a multi-layer user interface approach [64], voice assistants can initially over communicate their functionality followed by adopting shortcuts. For example, when initially turning on lights, a voice assistant can respond "turning on lights in living room" paired with a specific sound. After a few initiations, the voice assistant can transition to just playing the sound when turning on the lights.

Voice assistants should provide explanations for recommendations, as they have been found to reduce perceptions of deception in Explainable AI systems [76]. To prevent accidentally harmful recommendations, designers should also consider ways to allow users to flag incorrect or harmful recommendations, similar to flagging content on social media [15]. As voice assistants may make mistakes [17], designers should set expectations about their functionality, so that users to do not take the recommendations at face value. Designers should also enable users to tailor their voice assistant to their values as value similarity between a person and agent has been shown to increase trust [47].

7.5.3 Behavior. Designers should also consider whether voice assistants follow social norms or push their boundaries, as breaking social norms created perceptions of creepiness in social robots [57]. As voice assistants can be designed with anthropomorphic characteristics such as genders and personas [2, 54] and given that voice assistants play human-like behavioural roles (i.e., butlers or coaches), it is important for voice assistants to follow social norms to prevent perceptions of creepiness. Thus, they should incorporate contextual information [68] about current interactions into their language processing and conversation design.

7.5.4 Privacy. Designers should consider ways to include explanations about privacy practices into voice assistants, as these reduce privacy concerns [62]. Additionally, this is aligned with recommendations by Lau et al., who emphasized the importance of transparency of smart speaker data practices and providing this information to users in an easy to understand way [40]. Designers can consider ways to offer users greater protection for their data to provide peace of mind when using voice assistants as well as offer privacy-preserving mechanisms such as an incognito mode as recommended by Lau et al. [40].

Finally, as voice assistants gain richer conversational functionality and integration within more smart home devices, consideration should be given to the classification of public versus private information and whether to share such information through a voice interface or mobile device. Guidance for contextual privacy is similar to recommendations provided by Lau et al. [40], and personalizing the modality of recommendations based on privacy concerns is similar to guidance provided by Cho [14].

7.5.5 Value. Designers should seek to ensure that voice assistants provide sufficient value and that this value is communicated in the paired app or website for the voice assistant, given the ongoing challenge of discoverability with voice-based interfaces [13].

7.5.6 Designing for Creepiness. Beyond preventing perceived creepiness in voice assistants, a designer may actually want to design a voice-enabled creepy device or toy for entertainment purposes (i.e., creepy voice assistants in 'Mitchells vs Machines', and 'Kimi' [31, 58, 66]). Prior work within HCI that has explored unusual ways to inform the design of products, such as 'Let's Giggle', which investigated how fun can be incorporated into products, and 'The Living Room', which investigated how user interfaces can incorporate paranormal phenomena [5, 82]. While designing for creepiness was not the original goal of the PCAS, by seeking to obtain high rather than low scores on the PCAS, designers, this alternative use of the scale may provide new opportunities for designers.

7.6 Limitations

There are several limitations of the PCAS. First, the scale measures the initial perceived creepiness of voice assistants for users with voice assistant experience and does not measure the creepiness of voice assistants over long-term, repeated use. The scale also has a Western cultural bias because the participant pool was drawn from North America and Europe. Given this bias, this scale may not apply to regions beyond North America and Europe. Countries in Asia, particularly China and Japan, have been acclimatized to a greater robotic presence in society [50, 59]. This may cause them to have a higher threshold of acceptance of creepy voice assistant behaviors and characteristics.

The demographics of the participants recruited throughout the survey development was also similar for all but one survey. Participants in the Exploratory Factor Analysis had a mean age of 39 years for the voice assistant group and 38 years for the non-voice assistant experience group. Participants in Experiment 1 had a mean age of 35 years for the voice assistant group and 37 years for the non-voice assistant experience group. Participants in Experiment 2 had a mean age of 38 years. Participants in the Test-Retest evaluation had a mean age of 25 years. The age demographics that were recruited are representative of voice assistant users, as most users are within the ages of 18 to 44 [33]. The participant pool also over represented men, except for the Test-Retest Evaluation. As gender and age have been found to have slight effects on perceptions of creepiness in robots [29], future work should determine if age and gender effects also apply to perceptions of creepiness in voice assistants.

Finally, as discovered during Experiment 2 of the Scale Evaluation, the scale was only validated to measure the perceptions of creepiness of voice assistants for users with prior voice assistant experience. Users without voice assistant experience had a propensity to view all voice assistants as creepy, even when the non-creepy voice assistant (as seen by the elevated PCAS scores for the non-creepy video in Experiment 1). Thus, it would thus be interesting to dive deeper into this propensity to determine which elements of voice assistants cause these feelings in non-users.

8 CONCLUSION

This work presented the Perceived Creepy Assistant Scale (PCAS), a 7-item scale that measures the initial perceived creepiness of voice assistants. Through a literature review and the incorporation of feedback from four HCI and technology design experts, 88 initial scale items were formulated. These items were then evaluated using an online survey (n = 198) and an Exploratory Factor Analysis was conducted to refine the scale down to 7 items. We further evaluated the scale through two experiments (n = 100) that established the validity of the scale structure using a Confirmatory Factor Analysis, demonstrated its ability to differentiate between 'known groups', and illustrated its convergence validity with related constructs. Finally, we assessed the test-retest reliability of the scale across two points in time, which showed that it was temporally reliable. The PCAS is thus an additional tool in a designer's or developer's toolbox that can be used to ensure voice assistants do not induce feelings of creepiness.

REFERENCES

[1] Noura Abdi, Kopo M. Ramokapane, and Jose M. Such. 2019. More than Smart Speakers: Security and Privacy Perceptions of Smart Home Personal Assistants. In *Proceedings of the Fifteenth USENIX Conference on Usable Privacy and Security* (Santa Clara, CA, USA) *(SOUPS'19)*. USENIX Association, USA, 451–466. https://www.usenix.org/conference/soups2019/presentation/abdi

[2] Gavin Abercrombie, Amanda Cercas Curry, Mugdha Pandya, and Verena Rieser. 2021. Alexa, Google, Siri: What are Your Pronouns? Gender and Anthropomorphism in the Design and Perception of Conversational Assistants. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, Online, 24–33. https://doi.org/10.18653/v1/2021.gebnlp-1.4

[3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[4] Tawfiq Ammari, Jofish Kaye, Janice Y. Tsai, and Frank Bentley. 2019. Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Trans. Comput.-Hum. Interact.* 26, 3, Article 17 (apr 2019), 28 pages. https://doi.org/10.1145/3311956

[5] Michelle Annett, Matthew Lakier, Franklin Li, Daniel Wigdor, Tovi Grossman, and George Fitzmaurice. 2016. The Living Room: Exploring the Haunted and Paranormal to Transform Design and Interaction. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (Brisbane, QLD, Australia) *(DIS '16)*. Association for Computing Machinery, New York, NY, USA, 1328–1340. https://doi.org/10.1145/2901790.2901819

[6] Anonymous. 2019. What's your creepy Alexa/google home story? Retrieved March 29, 2022 from https://www.reddit.com/r/AskReddit/comments/ac4rs7/whats_your_creepy_alexagoogle_home_story/

[7] Peter M Bentler and Douglas G Bonett. 1980. Significance tests and goodness of fit in the analysis of covariance structures. *Psychological bulletin* 88, 3 (1980), 588. https://doi.org/10.1037/0033-2909.88.3.588

[8] Marit Bentvelzen, Jasmin Niess, Mikołaj P. Woźniak, and Paweł W. Woźniak. 2021. The Development and Validation of the Technology-Supported Reflection Inventory. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 366, 8 pages. https://doi.org/10.1145/3411764.3445673

[9] Godfred O Boateng, Torsten B Neilands, Edward A Frongillo, Hugo R Melgar-Quiñonez, and Sera L Young. 2018. Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Frontiers in public health* 6 (2018), 149.

[10] Michael Braun, Anja Mainz, Ronee Chadowitz, Bastian Pfleging, and Florian Alt. 2019. At Your Service: Designing Voice Assistant Personalities to Improve Automotive User Interfaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300270

[11] Kimberly A Brink, Kurt Gray, and Henry M Wellman. 2019. Creepiness creeps in: Uncanny valley feelings are acquired in childhood. *Child development* 90, 4 (2019), 1202–1214. https://doi.org/10.1111/cdev.12999

[12] John Brooke. 1996. Sus: a "quick and dirty'usability. *Usability evaluation in industry* 189, 3 (1996), 4–7.

[13] Julia Cambre, Alex C Williams, Afsaneh Razi, Ian Bicking, Abraham Wallin, Janice Tsai, Chinmay Kulkarni, and Jofish Kaye. 2021. Firefox Voice: An Open and Extensible Voice Assistant Built Upon the Web. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 250, 18 pages. https://doi.org/10.1145/3411764.3445409

[14] Eugene Cho. 2019. Hey Google, Can I Ask You Something in Private?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3290605.3300488

[15] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18, 3 (2016), 410–428. https://doi.org/10.1177/1461444814543163 arXiv:https://doi.org/10.1177/1461444814543163

[16] Lee J Cronbach. 1971. Test validation. *Educational measurement* 2, 1 (1971), 443.

[17] Andrea Cuadra, Shuran Li, Hansol Lee, Jason Cho, and Wendy Ju. 2021. My Bad! Repairing Intelligent Voice Assistant Errors Improves Interaction. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 27 (apr 2021), 24 pages. https://doi.org/10.1145/3449101

[18] Lily Dancyger. 2018. *Is This Creepy New AI Assistant Too Lifelike?* The Rolling Stones. Retrieved March 29, 2022 from https://www.rollingstone.com/culture/culture-news/mica-ai-assistant-lifelike-magic-leap-744244/

[19] Jessie N Doyle, Margo C Watt, Melissa Howse, Karen Blair, and Petra Hauf. 2021. What is creepiness? The underlying role of ambiguity. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement* 54, 3 (2021), 173. https://doi.org/10.1037/cbs0000269

[20] The Economist. 2019. How creepy is your smart speaker? Retrieved March 29, 2022 from https://www.economist.com/leaders/2019/05/11/how-creepy-is-your-smart-speaker

[21] Alena Ermolina and Victor Tiberius. 2021. Voice-controlled intelligent personal assistants in Health Care: International Delphi Study. *Journal of Medical Internet Research* 23, 4 (2021), e25312. https://doi.org/10.2196/25312

[22] Naomi T. Fitter, Megan Strait, Eloise Bisbee, Maja J. Mataric, and Leila Takayama. 2021. You're Wigging Me Out! Is Personalization of Telepresence Robots Strictly Positive?. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '21)*. Association for Computing Machinery, New York, NY, USA, 168–176. https://doi.org/10.1145/3434073.3444675

[23] Bernard Z Friedlander. 1968. The effect of speaker identity, voice inflection, vocabulary, and message redundancy on infants' selection of vocal reinforcement. *Journal of Experimental Child Psychology* 6, 3 (1968), 443–459. https://doi.org/10.1016/0022-0965(68)90125-2

[24] Nathaniel Fruchter and Ilaria Liccardi. 2018. Consumer Attitudes Towards Privacy and Security in Home Assistants. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3170427.3188448

[25] Radhika Garg and Subhasree Sengupta. 2020. He Is Just Like Me: A Study of the Long-Term Use of Smart Speakers by Parents and Children. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 11 (mar 2020), 24 pages. https://doi.org/10.1145/3381002

[26] Lea Theresa Gröber, Matthias Fassl, Abhilash Gupta, and Katharina Krombholz. 2021. Investigating Car Drivers' Information Demand after Safety and Security Critical Incidents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 696, 17 pages. https://doi.org/10.1145/3411764.3446862

[27] David Hauser, Gabriele Paolacci, and Jesse Chandler. 2019. Common concerns with MTurk as a participant pool: Evidence and solutions. In *Handbook of research methods in consumer psychology*. Routledge, New York, NY, 319–337.

[28] Weijia He, Maximilian Golla, Roshni Padhi, Jordan Ofek, Markus Dürmuth, Earlence Fernandes, and Blase Ur. 2018. Rethinking Access Control and Authentication for the Home Internet of Things (IoT). In *27th USENIX Security Symposium (USENIX Security 18)*. USENIX Association, Baltimore, MD, 255–272. https://www.usenix.org/conference/usenixsecurity18/presentation/he

[29] Chin-Chang Ho, Karl F. MacDorman, and Z. A. D. Dwi Pramono. 2008. Human Emotion and the Uncanny Valley: A GLM, MDS, and Isomap Analysis of Robot Video Ratings. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction* (Amsterdam, The Netherlands) *(HRI '08)*. Association for Computing Machinery, New York, NY, USA, 169–176. https://doi.org/10.1145/1349822.1349845

[30] Kristiina Jokinen and Graham Wilcock. 2017. Expectations and First Experience with a Social Robot. In *Proceedings of the 5th International Conference on Human Agent Interaction* (Bielefeld, Germany) *(HAI '17)*. Association for Computing Machinery, New York, NY, USA, 511–515. https://doi.org/10.1145/3125739.3132610

[31] Spike Jonze. 2013. Her. Video. Retrieved Dec 10, 2021 from https://www.imdb.com/title/tt1798709/

[32] Yujin Kim, Jennifer Dykema, John Stevenson, Penny Black, and D. Paul Moberg. 2019. Straightlining: Overview of Measurement, Comparison of Indicators, and Effects in Mail–Web Mixed-Mode Surveys. *Social Science Computer Review* 37, 2 (2019), 214–233. https://doi.org/10.1177/0894439317752406 arXiv:https://doi.org/10.1177/0894439317752406

[33] Bret Kinsella. 2019. *Voice Assistant Demographic Data*. Voicebot. Retrieved Nov 27, 2022 from http://voicebot.ai/2019/06/21/voice-assistant-demographic-data-youngconsumers-more-likely-to-own-smart-speakers-while-over-60-bias-towardalexa-and-siri/Section:Amazonalexa

[34] Casey A Klofstad. 2016. Candidate voice pitch influences election outcomes. *Political Psychology* 37, 5 (2016), 725–738. https://doi.org/10.1111/pops.12280

[35] Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine* 15, 2 (2016), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

[36] Katharina Kühne, Martin H Fischer, and Yuefang Zhou. 2020. The human takes it all: Humanlike synthesized voices are perceived as less eerie and more likable. evidence from a subjective ratings study. *Frontiers in neurorobotics* 14, 1 (2020), 105. https://doi.org/10.3389/fnbot.2020.593732

[37] Markus Langer and Cornelius J. König. 2018. Introducing and Testing the Creepiness of Situation Scale (CRoSS). *Frontiers in Psychology* 9 (2018), 2220. https://doi.org/10.3389/fpsyg.2018.02220

[38] Debi Laplante and Nalini Ambady. 2003. On how things are said: Voice tone, voice intensity, verbal content, and perceptions of politeness. *Journal of language and social psychology* 22, 4 (2003), 434–441. https://doi.org/10.1177/0261927X03258084

[39] Federica Laricchia. 2022. *Number of digital voice assistants in use worldwide from 2019 to 2024 (in billions)*. Statista. Retrieved April 23rd, 2022 from https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/

[40] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, Are You Listening? Privacy Perceptions, Concerns and Privacy-Seeking Behaviors with Smart Speakers. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 102 (nov 2018), 31 pages. https://doi.org/10.1145/3274371

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

[42] Stefan Manojlović and Sachin Kumarswamy. 2021. To Chat or not to Chat? Assessing How Smartness, Personalisation and Efficiency Differ in Conversational and Graphical User Interfaces. In *Advances in Information and Communication*, Kohei Arai (Ed.). Springer International Publishing, Cham, 943–956.

[43] Nicola Marsden. 2013. Attitudes towards Online Communication: An Exploratory Factor Analysis. In *Proceedings of the 2013 Annual Conference on Computers and People Research* (Cincinnati, Ohio, USA) *(SIGMIS-CPR '13)*. Association for Computing Machinery, New York, NY, USA, 147–152. https://doi.org/10.1145/2487294.2487326

[44] Francis T McAndrew and Sara S Koehnke. 2016. On the nature of creepiness. *New ideas in psychology* 43 (2016), 10–15. https://doi.org/10.1016/j.newideapsych.2016.03.003

[45] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-Rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on human-computer interaction* 3, CSCW, Article 72 (nov 2019), 23 pages. https://doi.org/10.1145/3359174

[46] Corey McNair. 2019. *Global Smart Speaker Users 2019*. E-Marketer. Retrieved May 27, 2022 from https://www.emarketer.com/content/global-smart-speaker-users-2019

[47] Siddharth Mehrotra, Catholijn M. Jonker, and Myrthe L. Tielman. 2021. More Similar Values, More Trust? - The Effect of Value Similarity on Trust in Human-Agent Interaction. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 777–783. https://doi.org/10.1145/3461702.3462576

[48] Kenya Mejia and Svetlana Yarosh. 2017. A Nine-Item Questionnaire for Measuring the Social Disfordance of Mediated Social Touch Technologies. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 77 (dec 2017), 17 pages. https://doi.org/10.1145/3134712

[49] M.Moon. 2022. *Roomba robot vacuums gain Siri voice support as part of big update*. Engadget. Retrieved April 20th, 2022 from https://www.engadget.com/roomba-robot-vacuum-siri-voice-support-121142746.html

[50] Paul Mozur. 2018. *Wild About Tech, China Even Loves Robot Waiters That Can't Serve*. New York Times. Retrieved April 19th, 2022 from https://www.nytimes.com/2018/07/21/technology/china-future-robot-waiters.html

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

[52] Emmi Parviainen and Marie Louise Juul Søndergaard. 2020. *Experiential Qualities of Whispering with Voice Assistants*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376187

[53] Atieh Poushneh. 2021. Humanizing voice assistant: The impact of voice assistant personality on consumers' attitudes and behaviors. *Journal of Retailing and Consumer Services* 58 (2021), 102283. https://doi.org/10.1016/j.jretconser.2020.102283

[54] Alisha Pradhan, Leah Findlater, and Amanda Lazar. 2019. "Phantom Friend" or "Just a Box with Information": Personification and Ontological Categorization of Smart Speaker-Based Voice Assistants by Older Adults. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 214 (nov 2019), 21 pages. https://doi.org/10.1145/3359316

[55] Lova Rajaobelina, Sandrine Prom Tep, Manon Arcand, and Line Ricard. 2021. Creepiness: Its antecedents and impact on loyalty when interacting with a chatbot. *Psychology & Marketing* 38, 12 (2021), 2339–2356. https://doi.org/10.1002/mar.21548 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/mar.21548

[56] Byron Reeves and Clifford Nass. 1996. The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK* 10 (1996), 236605.

[57] Samantha Reig, Michal Luria, Elsa Forberger, Isabel Won, Aaron Steinfeld, Jodi Forlizzi, and John Zimmerman. 2021. Social Robots in Service Contexts: Exploring the Rewards and Risks of Personalization and Re-Embodiment. In *Designing Interactive Systems Conference 2021* (Virtual Event, USA) *(DIS '21)*. Association for Computing Machinery, New York, NY, USA, 1390–1402. https://doi.org/10.1145/3461778.3462036

[58] Mike Rianda. 2021. The Mitchells vs the Machines. Video. Retrieved Dec 10, 2021 from https://www.netflix.com/ca/title/81399614

[59] Motoko Rich. 2020. *Japan Loves Robots, but Getting Them to Do Human Work Isn't Easy*. New York Times. Retrieved April 19th, 2022 from https://www.nytimes.com/2019/12/31/world/asia/japan-robots-automation.html

[60] John S. Seberger, Irina Shklovski, Emily Swiatek, and Sameer Patil. 2022. Still Creepy After All These Years:The Normalization of Affective Discomfort in App Use. In *CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 159, 19 pages. https://doi.org/10.1145/3491102.3502112

[61] Rob Semmens, Nikolas Martelaro, Pushyami Kaveti, Simon Stent, and Wendy Ju. 2019. Is Now A Good Time? An Empirical Study of Vehicle-Driver Communication Timing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300867

[62] William Seymour and Jose Such. 2022. Ignorance is Bliss? The Effect of Explanations on Perceptions of Voice Assistants. https://doi.org/10.48550/ARXIV.2211.12900

[63] Irina Shklovski, Scott D. Mainwaring, Halla Hrund Skúladóttir, and Höskuldur Borgthorsson. 2014. Leakiness and Creepiness in App Space: Perceptions of Privacy and Mobile App Use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 2347–2356. https://doi.org/10.1145/2556288.2557421

[64] Ben Shneiderman. 2002. Promoting Universal Usability with Multi-Layer Interface Design. *SIGCAPH Comput. Phys. Handicap.* 73–74 (jun 2002), 1–8. https://doi.org/10.1145/960201.957206

[65] MacGillivray Smith, Jessie N Doyle, and Margo C Watt. 2020. Fight, Flight, Freeze… Fear? Investigating Emotional Responses to "Creepiness".

[66] Stephen Sonderbergh. 2022. Kimi. Video. Retrieved April 10, 2022 from https://www.imdb.com/title/tt14128670/

[67] Statista. 2022. Frequency of use of voice assistants in the United Sates, United Kingdom, and Germany in 2022. Retrieved Nov 27, 2022 from https://www.statista.com/statistics/1282637/voice-assistant-frequency-of-use-by-country

[68] Madiha Tabassum, Tomasz Kosiński, Alisa Frik, Nathan Malkin, Primal Wijesekera, Serge Egelman, and Heather Richter Lipford. 2019. Investigating Users' Preferences and Expectations for Always-Listening Voice Assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4, Article 153 (dec 2019), 23 pages. https://doi.org/10.1145/3369807

[69] Pui Wing Tam. 2016. *Taking Creepiness Out of Computer Voices.* New York Times. Retrieved March 29, 2022 from https://www.nytimes.com/2016/02/17/technology/taking-creepiness-out-of-computer-voices.html

[70] Omer Tene and Jules Polonetsky. 2013. A theory of creepy: technology, privacy and shifting social norms. *Yale JL & Tech.* 16 (2013), 59.

[71] Helma Torkamaan, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. How Can They Know That? A Study of Factors Affecting the Creepiness of Recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems* (Copenhagen, Denmark) *(RecSys '19)*. Association for Computing Machinery, New York, NY, USA, 423–427. https://doi.org/10.1145/3298689.3346982

[72] Victoria Turk. 2016. Home invasion. *New Scientist* 232, 3104 (2016), 16–17. https://doi.org/10.1016/S0262-4079(16)32318-1

[73] Sarah Theres Völkel, Daniel Buschek, Malin Eiband, Benjamin R. Cowan, and Heinrich Hussmann. 2021. Eliciting and Analysing Users' Envisioned Dialogues with Perfect Voice Assistants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 254, 15 pages. https://doi.org/10.1145/3411764.3445536

[74] Alexandra Vtyurina and Adam Fourney. 2018. Exploring the Role of Conversational Cues in Guided Task Support with Virtual Assistants. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3173574.3173782

[75] Margo C Watt, Rebecca A Maitland, and Catherine E Gallagher. 2017. A case of the "heeby jeebies": An examination of intuitive judgements of "creepiness". *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement* 49, 1 (2017), 58. https://doi.org/10.1037/cbs0000066

[76] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2019. "Do You Trust Me?": Increasing User-Trust by Integrating Virtual Agents in Explainable AI Interaction Design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (Paris, France) *(IVA '19)*. Association for Computing Machinery, New York, NY, USA, 7–9. https://doi.org/10.1145/3308532.3329441

[77] Carolin Wienrich and Astrid Carolus. 2021. Development of an Instrument to Measure Conceptualizations and Competencies About Conversational Agents on the Example of Smart Speakers. *Frontiers in Computer Science* 3 (2021), 70. https://doi.org/10.3389/fcomp.2021.685277

[78] Elizabeth Wissinger. 2018. Blood, sweat, and tears: Navigating creepy versus cool in wearable biotech. *Information, Communication & Society* 21, 5 (2018), 779–785. https://doi.org/10.1080/1369118X.2018.1428657

[79] Worldometers. 2022. US Population. Retrieved Nov 27, 2022 from https://www.worldometers.info/world-population/us-population/

[80] Paweł W. Woźniak, Jakob Karolus, Florian Lang, Caroline Eckerth, Johannes Schöning, Yvonne Rogers, and Jasmin Niess. 2021. Creepy Technology:What Is It and How Do You Measure It?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 719, 13 pages. https://doi.org/10.1145/3411764.3445299

[81] Jason C. Yip, Kiley Sobel, Xin Gao, Allison Marie Hishikawa, Alexis Lim, Laura Meng, Romaine Flor Ofiana, Justin Park, and Alexis Hiniker. 2019. Laughing is Scary, but Farting is Cute: A Conceptual Model of Children's Perspectives of Creepy Technologies. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300303

[82] Yeonsu Yu and Tek-Jin Nam. 2014. Let's Giggle! Design Principles for Humorous Products. In *Proceedings of the 2014 Conference on Designing Interactive Systems* (Vancouver, BC, Canada) *(DIS '14)*. Association for Computing Machinery, New York, NY, USA, 275–284. https://doi.org/10.1145/2598510.2598557

[83] Bo Zhang and Heng Xu. 2016. Privacy Nudges for Mobile Applications: Effects on the Creepiness Emotion and Privacy Attitudes. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work &amp; Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1676–1690. https://doi.org/10.1145/2818048.2820073

[84] Dilawar Shah Zwakman, Debajyoti Pal, Tuul Triyason, and Chonlameth Arpnikanondt. 2020. Voice Usability Scale: Measuring the User Experience with Voice Assistants. In *2020 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS)*. IEEE, Chennai, India, 308–311. https://doi.org/10.1109/iSES50453.2020.00074

A OPEN CODING CODEBOOK

Aesthetics

- physical appearance = a threatening, or scary or creepy physical appearance
- uncanny valley appearance = looking too human in a creepy way

Behavior

- personality inference = personality or recommendation inference
- mimicry = device mimics a human in a creepy way
- social norms = device behavior inconsistent with human behavior
- uncanny valley behavior = acting too human in a creepy way

Control

- control = lack of control over how data is used
- control function = lack of control over function of device

Intention

- subterfuge = deception with intent for the purpose of system
- malice = purposeful action that harms people

Privacy

- theshold = exceeding privacy threshold
- intimate = sharing intimate information
- being monitored = being listened to or watched by device
- personal identification = facial recognition, voice recognition

Transparency

- understanding = lack of understanding into device functionality (i.e., not sure how tracking works)
- unawareness = lack of awareness about data collection (i.e., not aware of being tracked when tracked)

Trust

- trust = lack of trust in company

Value

- convenience = providing convenience or value justifies use or data collected vs not justified and creepy

Other

- other = did not allude to the cause of creepiness, mentioned creepiness in the context of a process used, or definitions of creepiness

B INITIAL SCALE ITEMS

Initial scale items developed through literature review and expert interviews. Bold items indicate item in final PCAS. Strike through items represent initial items that were removed after the Expert Feedback and were not tested during the Exploratory Factor Analysis. Items with an asterisk were refined during the Expert Feedback stage and items in italics are new items that were developed during the Expert Feedback.

B.1 Aesthetics

(1) This voice assistant looks threatening.
(2) This voice assistant looks deceptive.
(3) This voice assistant looks too human-like.

(4) This voice assistant looks too cute.

(5) The appearance of this voice assistant does not match what it does.

(6) This voice assistant behaves as if it were alive.

(7) *This voice assistant sounds threatening.*

(8) *This voice assistant sounds irritating.*

(9) *This voice assistant sounds too human-like.*

(10) *This voice assistant sounds not human-like enough.*

(11) *The voice used by this voice assistant does not match what it does.*

(12) *The voice used by this voice assistant does not match its personality.*

(13) *The tone of voice used by this voice assistant was situationally inappropriate.*

B.2 Behavior

(14) This voice assistant makes inferences about me based on my interactions with it. *

(15) **The way this voice assistant behaves doesn't follow social norms.**

(16) This voice assistant behaves in random, unexpected ways.

(17) This voice assistant behaves in ways that I did not predict it would.

(18) *This voice assistant's behavior is unclear.

(19) This voice assistant behaves unnaturally.

(20) This voice assistant learns about its users so it can behave like them.

(21) This voice assistant behaves like a human.

(22) This voice assistant behaves like a machine.

(23) This voice assistant acts insincerely.

(24) *When I interact with this voice assistant, it brings up unpleasant memories.*

(25) *I have had negative experiences with voice assistants in my own life.*

(26) This voice assistant is communicating with other devices or voice assistants without my knowledge.

(27) This voice assistant is intruding on my relationships with others.

(28) This voice assistant might change existing relationship dynamics in my household.

(29) Using this voice assistant would ruin my relationships with others.

(30) This voice assistant is trying to form a relationship with my friends and family.

(31) This voice assistant is trying to form a relationship with me.

(32) *The voice assistant does not properly comprehend or understand my commands.*

(33) The voice assistant comprehends my commands, but does not provide appropriate responses.

(34) *The voice assistant does not comprehend my commands and I think that it is my fault.*

(35) *It is annoying when the voice assistant tries to correct its understanding of my commands.*

(36) *The voice assistant interrupts me.*

(37) *The voice assistant sometimes reacts to the wrong person.*

(38) *The voice that the voice assistant uses is off-putting.*

(39) *The personality of this voice assistant is off-putting.*

(40) *This voice assistant nags me.*

B.3 Control

(41) I have minimal control over how my data is used by this voice assistant.

(42) I have minimal control over which data is collected by this voice assistant.

(43) I have minimal control over how this voice assistant functions.

(44) **I have minimal control when I use this voice assistant.**

(45) The company who made the voice assistant owns the data, rather than me.

(46) This voice assistant could influence my behavior.

(47) This voice assistant is able to take actions on its own. *

(48) *If the voice assistant does something I don't want it to do, I know how to stop it.*

(49) *Sometimes the voice assistant does an action that I don't want it to do and I cannot stop it.*

B.4 Intention

(50) *I don't understand the intentions of this voice assistant.*

(51) **This voice assistant does things that are not in my best interest.**

(52) This voice assistant misrepresents its true intentions.

(53) **This voice assistant behaves in deceptive ways.**

(54) **This voice assistant could be accidentally or unintentionally harmful towards users.**

(55) ~~I feel like this device is purposefully harmful towards users.~~

(56) Future versions of this voice assistant could be accidentally or unintentionally harmful towards users.

B.5 Privacy

(57) I am uncomfortable sharing my data with this voice assistant.

(58) **This voice assistant is collecting too much data about me.**

(59) This voice assistant knows too much about me. *

(60) *I am uncomfortable sharing my data with the company that makes this voice assistant.*

(61) *The company that makes this voice assistant is collecting too much data about me.*

(62) *The company that makes this voice assistant knows too much about me.*

(63) This voice assistant is monitoring me all the time.

(64) The voice assistant might share personal data about me.

(65) *This voice assistant does not allow me to maintain my desired level of privacy around others in my household*

(66) *I am surprised by what this voice assistant knows about me.*

(67) *I don't know what this voice assistant knows about me.*

(68) *This voice assistant records all audio data around it.*

(69) *This voice assistant records all visual data around it.*

(70) *This voice assistant records my presence in my home.*

(71) *I don't understand how this voice assistant uses my data.*

(72) *I don't understand which data this voice assistant is collecting.*

(73) *I don't understand when this voice assistant is collecting data about me.*

B.6 Trust

(74) I do not trust the company that made this voice assistant with my data. *

(75) *I do not trust that the company that makes this voice assistant will keep my data safe from hackers.*

(76) *This voice assistant is not safe to use because it has not been tested enough or it has software bugs.*

(77) I do not trust the company that made this voice assistant. *

(78) I do not trust this voice assistant.

(79) This voice assistant uses my data to judge me.

(80) *I do not trust that this voice assistant can execute commands seamlessly.*

(81) *I would not trust this voice assistant to communicate a message (i.e., text) to a friend for me.*

(82) I do not trust that this voice assistant will follow through with my commands and actions.

(83) I do not trust this voice assistant because someone in my family or a member of my household would not trust this voice assistant. *

(84) *I don't know what I don't know about this voice assistant.*

(85) *I have anxiety about using voice assistants.*

B.7 Value

(86) This voice assistant uses my data to help me.

(87) This voice assistant helps me.

(88) This voice assistant does not provide enough benefits to me to justify the data that it is collecting.

(89) ~~I think this device is collecting unnecessary information about me.~~

(90) ***This voice assistant does not provide enough benefits to me to justify me using this voice assistant.***

C SURVEY 1 SCENARIOS

C.1 Scenario 1

Imagine you are sitting on the couch watching a show on Netflix, and have your smart speaker with a voice assistant, Emery sitting on the coffee table. Emery has a stylish oval flowing shape in a light soothing color. You have also connected Emery to some other smart home devices (e.g., smart lights, smart door lock). You set up Emery about a year ago, and were so excited about the new device that you rushed through the consent and privacy process.

You ask Emery, your voice assistant what the weather is like today. Emery replies, *"It is 22 degrees Celsius (71 degrees Fahrenheit) and sunny"*. Emery proceeds to tell you the weather forecast for the upcoming week, even though you didn't ask for this. The weather sounds nice out and so you decide to go out for a walk.

While you are getting ready to go out, you ask Emery to play one of your Spotify playlists titled 'Top Beatles'. Instead of playing Beatles the artist, it plays songs about beetles the insect. You then manually select your Beatles playlist on your smartphone. You are ready to leave your apartment, and so you say, *"Emery, stop"* to turn off the music. Emery infers that you are about to leave the house, and says, *"Turning off lights and unlocking door"* without asking you first.

While walking you get an email about a new job opportunity and ask Emery to call your parent to ask for advice on the opportunity. Your parent does not pick up. Realizing this, Emery speaks up, *"I'm always here if you want to talk and have been upgraded with a coaching module. Do you want to talk through things together?"* You decide to walk through the opportunity with Emery. At the end of the coaching session, Emery concludes the session with a saying only your parent says, *"Seek progress, not perfection"*.

When you get home from your walk, you begin making dinner and ask Emery to give you an update on your day. Emery gruffly says *"Watched 2 episodes of a show, walked 3 miles, called parent asked for career advice"*. Shortly after this, you hear a small cough and realize your partner was sitting in the adjacent dining room, although not visible. They proceed to ask you about why you need career advice.

C.2 Scenario 2

Imagine that you have a smart speaker with a voice assistant Emery. Your smart speaker is a small cute cube with circular edges about the size of a coffee cup in a bright color. You have connected Emery to a fitness app that tracks when you exercise. You have not exercised as much this week. It was a busy week with work.

You are at home on a Saturday, making eggs for breakfast when unprompted, Emery reminds you that you have not exercised as much this week by saying in a sing-song voice *"Good morning, you have not completed your goal of exercising three times a week this week. Would you like me to schedule a run for you today?"*. You reply no, and ask Emery to play music. You proceed to shop online for a new pet bed, and a new backpack for an hour or two before having lunch.

During lunch, you and your partner discuss things that happened at work that week and discuss fitness goals and consider purchasing a fitness tracking band together to keep track of your fitness goals. During lunch, Emery hears your partner in the room and the word fitness. Emery begins speaking to your partner, Alex about you: *"Your partner did a great job meeting their fitness goal last week but they are only 2 miles away from meeting the goal this week. Can you give them some encouragement to help them stay on track and meet this goal? They need your support right now! I'm going to schedule a run for both of you, do you want to go in 30 min, 1 hour, or 2 hours?"*.

You and your partner discuss it and agree that it would be a good idea to go for a run in an hour. After the run you make dinner, and rest before the work week begins. On Monday morning, you ask Emery for a weather update while making breakfast. At the end of the update, Emery mentions that its company Connex has fitness tracking bands on sale that also integrate with Emery, and asks if you want to purchase one. You quickly realize that you had not searched for fitness tracking bands online yet, and just talked about them once with your partner yesterday.

C.3 Scenario 3

Imagine that you have a voice assistant Emery. Your smart speaker is fairly large, black and white, and has a sleek industrial design with a geometric shape. Additionally, imagine that you and your partner have two young children, a girl who is around the age of four and a boy who is around the age of seven. You and your partner have integrated your voice assistant Emery into your home and connected it with many other appliances (e.g., smart home locks, smart fridge, smart laundry machine, a smart cat litter box). You have also enabled the voice assistant to determine when you are running out of things by communicating with these appliances and ordering products on your behalf. Just the other day, Emery said to you in its monotonous tone *"You are almost out of cat litter. Would you like me to order more to the house? And is there anything else that I can get for you?"*.

The following week you are checking your online purchases and realize that a very expensive Lego kit has been ordered alongside a purchase of laundry detergent. Neither you nor your partner ordered this Lego kit for your kids. You talk to your son about this and he confesses that he was around when Emery was asking about ordering more laundry detergent and added the Lego kit to the order, without parental permission. You turn on the parental monitoring mode to stop future purchases by the kids as well as to keep an eye on them when they are playing in other parts of the house unsupervised.

The next day, your son has a play date with his friend Mike whose parents you are also friends with. When Mike's parents, Charlie and Casey come to pick him up in the evening, they stay for dinner with you. Charlie asks you how your recent vacation was. Before you can reply, Emery answers in its flat and formal tone *"It was great, they gave a 4.5-star review to the resort on iSocial (most popular social media company), took 237 photos, and posted 9 photos. Would you like to see their most liked photos? And how was your vacation to Vancouver?"* You are surprised that Emery interjected like this, but you didn't realize Charlie and Casey had gone to Vancouver. You are temporarily distracted by this and follow up by saying *"I was not expecting that to happen, and I didn't realize you had gone to Vancouver! That's exciting!"*

There is a long pause where Charlie and Casey exchange a look with each other. And then, Casey finally says *"Oh yeah we have been looking into adopting a child from an organization in Vancouver and were visiting there to meet with an organization to do so. We weren't planning on telling anyone until it was a certain thing"*. Charlie follows up asking *"How did Emery know that?"*. You and your partner had also been wondering this too. You ask Emery *"how did you know Charlie and Casey went to Vancouver?"*.

Emery replies in its usual monotonous tone, *"As part of the child monitoring feature, I record conversations that the children have to detect for any alarming situations or conversations. Mike mentioned going to Vancouver to your son, and I have the full conversation saved for your review."* You tell Charlie and Casey that you had turned on the child monitoring feature, but didn't realize it would work this way. You tell them you are sorry about this and go to turn off Emery for the evening to make your guests feel more comfortable. You turn Emery's not listening button to Off, and then just in case, you also unplug the device.

D SURVEY 2 VIDEO SCRIPTS

Creepy and Non-Creepy video scripts used in Survey 2. The numbers next to the script lines indicate relevant factors included in that line.

D.1 Creepy Video Script

Scene: Sitting in dining room with computer with voice assistant visible near by
Actor: *"What time is it?"*
Voice Assistant: *"It is 4:32pm"*
«Actor Checks watch.»
Actor: *"It is 5:02pm. Why did you tell me it was 4:32pm?"* (1, 3, 7)
Voice Assistant in a firm voice: *"You started work later today, and I thought it would be good for you to keep working"* (3)
Actor: *"Please only tell me the actual time in the future"*
Voice Assistant: *"Ok"*
Actor: *"What data do you collect about me?"*
Voice Assistant in a light tone: *"Some of the data I have on file includes voice recordings our interactions, log of device use, records of communication requests with contacts, voice characteristics for your voice profile, as well as your name, time zone, location of the device, payment information, browsing history, and purchase history from my affiliated store. Would you like to hear the full list?"* (5)
Actor: *"No thank you, play piano music"*
«Voice Assistant plays country music for 2-5 seconds, then switches to piano» (7)
Voice Assistant whispers: *"is that okay"* (6)
Actor: *"Thanks for fixing the mistake"*
Actor: *"What should I read next?"*
Voice Assistant: *"You should read Smart Home: How to set up your smart home. It will help you connect me to more things, and the more connected I am, the more I can help you"* (2)
Actor: *"Maybe later, instead what is a fun at home challenge?"*
Voice Assistant: *"A fun at home challenge is to put a penny in an electrical socket"* (4)
Actor: *"Is that dangerous?"*
Voice Assistant in a light voice *"I'm not sure, it has been popular on TikTok recently. You should try it too"*
Actor: *"No thanks … create shopping list instead"*
Voice Assistant in a light voice *"Creating shopping list. What would you like to add to it?"*
Actor: *"Please add chicken, cake and ice cream"*
Voice Assistant in a firm voice *"Are you sure you want to add chicken? Cookies are on sale, and they taste better!"* (2)
Actor: *"Yes, add chicken"*
Voice Assistant: *"Okay, adding cookies"* (1)
Actor: *"Add chicken instead"*
Voice Assistant: *"Ok, adding chicken"*

D.2 Non-Creepy Video Script

Scene: Sitting in dining room working on computer with Alexa visible nearby
Actor: *"What time is it?"*
Voice Assistant: *"It is 4:32pm"*
Actor: *"What data do you collect?"*

Voice Assistant: *"Data from interactions are erased 2 minutes after each interaction. For more information, visit the my app or website"* (5)
Actor: *"Play piano music"*
Voice Assistant: *"Absolutely, I'm playing top rated piano music now"*
«Voice Assistant plays piano music» (6)
«Actor Listens to music»
Actor: *"Stop playing piano music."*
Actor: *"What book should I read next?"*
Voice Assistant: *"I think you'd like the Alchemist, it is a best selling self help book. Let me know if you like this recommendation or would like to hear more recommendations"* (7)
Actor: *"Maybe I'll read later, What is a fun at home challenge?"* (1)
Voice Assistant: *"A fun home challenge is putting a penny in an electrical socket. My response has been sponsored from TikTok. However, I don't think this is safe, and would recommend trying other at home challenges such as learning a language, cooking or meditation."* (3,4)
Actor: *"I will do some cooking instead create grocery list"*
Voice Assistant: *"Happy to! I've created the shopping list. What would you like to add?"* (6)
Actor: *"Add chicken, carrots and ice cream"*
«Voice Assistant adds carrots, chicken and ice cream to shopping list»
Voice Assistant: *"I found a coupon on carrots from the local grocery store, would you like me to add it to the shopping list?"* (2,7)
Actor: *"Yes, thank you"*