Privacy by Voice: Designing Usable Privacy Notices for the Voice Interface

Aafaq Sabir
North Carolina State University
asabir2@ncsu.edu

Abhinaya S.B.
North Carolina State University
asrivid@ncsu.edu

Dilawer Ahmed
North Carolina State University
dahmed2@ncsu.edu

Anupam Das
North Carolina State University
anupam.das@ncsu.edu

Abstract

With the increasing prevalence of voice interfaces, such as smart home assistants, conversational AI, and AR/VR systems, the need for effective privacy and consent mechanisms is more critical than ever. We conducted a mixed-methods study to address the challenges of ensuring effective consent for voice-based data sharing. Through interviews with voice assistant users (n=21), we identify five key design and contextual factors for effective privacy notices: context, control flow, modality, timing, and the voice used for notice delivery. We then prototyped these notices and performed a within-subject user study (n=160) to identify preferred notice designs. We found that the *voice* used for delivery and *timing* of the notice are the most critical factors influencing user preferences, with participants favoring notices delivered in the default app voice *before* data is requested. To our knowledge, this is the first study to design privacy notices specifically for voice-based data sharing in voice interfaces. Our findings contribute valuable insights to the privacy design literature and provide actionable guidance for developers working on emerging voice-driven platforms.

Keywords

Voice Assistants, Privacy Notices, Third-Party Voice Apps

1 Introduction

Voice-based user interfaces, like those in voice assistants, chatbots, and AR/VR, are becoming increasingly common, with voice assistants being the most dominant use case. In recent years, voice assistants have grown immensely popular, with an estimated 8.4 billion users globally [74]. These hands-free devices offer efficient ways to perform tasks such as controlling smart home devices, searching the web, shopping, gaming, and streaming music [5, 6]. Amazon Alexa remains the market leader, ahead of competitors like Google Assistant and Apple Siri [80]. Alexa has also partnered with major car manufacturers like General Motors, Ford, BMW, and Audi to bring hands-free controls, entertainment, and productivity features to vehicles [8, 9, 11, 13]. Smart voice assistants extend their functionality through third-party voice applications (called "Skills" on Amazon Alexa). These voice apps enable specialized tasks, such as playing games (e.g., Jeopardy), listening to a specific news stream, or managing a To-Do list, expanding the voice assistants' interactive capabilities, and web service integration. On Amazon Alexa, users activate skills by "enabling" them, akin to installing smartphone apps. Third-party developers must meet certification standards to publish skills on the Alexa store, which currently hosts over 100,000 skills, mostly from third-party creators [7].

Interacting with third-party applications on voice assistants exposes users to external content that may lack rigorous vetting [47, 61, 91], introducing privacy risks. These risks include unintentional disclosure of sensitive information [23, 58] or sharing of excessively personal data prompted by malicious developers [61]. Wie et al. analyzed 65,000 Alexa skills and found 21.7% exhibited suspicious data collection, with data requests unjustified by the skills' functionality, with 18.6% of these requests occurring verbally during interactions [89]. Further, research shows that most Alexa users misunderstand the role of third-party skills, frequently mistaking them for native features [66]. Third-party skills can verbally request and store personally identifiable information from the user through *verbal data requests.* These requests involve asking for data that is not found on the user's profile within the platform (data stored on a user's profile is directly accessible to the skill based on the permission settings). For instance, the user may not have their age stored on their Amazon profile, but a third-party Alexa skill may ask for age to suggest a suitable workout routine. It is essential that the voice interface platform clearly informs its users about the entities involved in accessing and handling their data.

A privacy notice serves as a mechanism for a system or organization to be transparent about its data practices, helping users make informed privacy decisions [79]. While existing privacy notices—such as website cookie banners or mobile app permission pop-ups—rely on visual elements, these methods are unsuitable for voice interfaces. This presents a significant research gap: designing effective auditory privacy notices that do not disrupt the user experience. While prior research has investigated privacy in visual interfaces, such as Android and iOS, where developer and user mental models of app permissions often conflict [83], these findings do not directly translate to voice-based systems. Further, while other studies have examined privacy notices in broader IoT contexts like wearables and VR [79, 87], our work is unique in its focus on designing usable privacy notices for voice-based data sharing, during verbal data requests. We adapt existing privacy design frameworks to the unique challenges of voice interfaces, with insights from users. We answer the following research questions. **RQ1:** *What privacy features do users desire in voice interfaces?* Through interviews

Proceedings on Privacy Enhancing Technologies YYYY(X), 1–19

https://doi.org/XXXXXXX.XXXXXXX

with current Alexa users (n=21), we identified user-desired privacy features and data request mechanisms. Participants expressed a strong preference for features like multi-factor authentication and warnings about verbal data collection. **RQ2:** *What design factors should we consider in designing privacy notices for voice interfaces?* Insights from the interviewed participants and existing literature on designing privacy notices revealed five key design factors: the context of the verbal data request, control flow, modality, timing, and the voice used for delivering the notice. **RQ3:** *a) Which privacy notices are more likely to be preferred, and b) which design factors influence user preferences the most?* After identifying the key design factors, we prototyped 20 privacy notice variants for a realistic Alexa skill. Through an interactive mixed-design user study (n=160), we collected participants' preferences in using these prototypes. We performed Bayesian Ordinal Multilevel Regression to identify preferred variants and Cumulative Link Mixed Models (CLMM) Regression to identify the factors most influencing preference: the *voice* used for delivery and the *timing* of the notice. In general, the preferred privacy notices are **verbal, interactive notices** in **default** Alexa voice **before** the information is requested.

In this paper, we prototyped privacy notices in the context of Alexa skills due to their rich capabilities and features; however, we believe our notice designs are broadly applicable to other voice assistant ecosystems and systems employing voice-based interfaces to make verbal data requests to users. In summary, we make the following contributions:

- To the best of our knowledge, we are the first to design privacy notices for verbal data requests for the voice interface.
- We elicit insights about user-desired privacy features and attributes of privacy notices for the voice interface.
- We prototype 20 variants of privacy notices, quantitatively identify the most influential factors determining user preference, and determine the specific privacy notice variants users prefer for different data-request contexts.
- We extract the key factors for designing privacy notices for verbal data requests through voice-based interfaces.
- We offer design recommendations to voice-interface developers and platforms to integrate usable privacy notices into various existing voice-based systems with minimal disruption to the user's interaction flow.

2 Background on Alexa Ecosystem

In this section, we provide an overview of the Alexa skill ecosystem, which serves as the platform for developing and evaluating our proposed privacy notices for voice interfaces. An Alexa skill consists of a front-end with voice models and a back-end for execution logic, ensuring users interact only with the front-end [12]. Audio from Alexa devices is processed in the cloud, where requests are identified, inputs extracted, and structured data sent to the back-end [18]. This design enhances privacy by restricting raw audio access and allows developers to focus on crafting engaging experiences without managing sensitive data or speech recognition. Amazon introduced third-party development for Alexa in 2015, fostering an app-like ecosystem [1]. This ecosystem includes native skills developed by Amazon and third-party skills created by external developers.

Data Collection by Skill. Skills can gather personal information via the Amazon account or verbal input. Accessing Amazon account data, such as name, email, or location, requires explicit user permission for personalization. Skills may also request personal details like a name or date of birth during conversations without needing explicit consent.

Intents and Slots. Understanding how voice assistants collect personal information is essential to assessing privacy implications and designing privacy notices. A skill's interaction model, or front-end, processes user input and maps it to functionalities using components like 'intents' and 'slots' [10]. An 'intent' represents an action triggered by a user request, while 'slots' capture details. For instance, a trip-planning skill might include an 'intent' like "PlanMyTrip" with slots for 'departure city', 'destination', and 'travel date'. Slots can use predefined data types, such as `AMAZON.DATE` or `AMAZON.City`, or custom types. The language model maps user utterances to slots based on sentence structure and sends the data to the backend intent handler.

Sample Utterances. Sample utterances are phrases users can say to trigger intents, with variables in curly brackets representing slot values. For example, a sample utterance for the `PlanMyTrip` intent might be, "I want to visit `{destination}`." A diverse set of sample utterances improves intent recognition and enriches the user experience by offering varied interaction options.

Dialog Model. Multi-turn dialogs allow more complex interactions, enabling the collection and confirmation of slot values needed to complete intents. The dialog model defines the flow, including prompts for slot values, user responses, rules for slot validation, and intent confirmation, ensuring a structured, efficient process for gathering user information [10].

3 Related Work

Privacy Concerns with Voice Assistants. Security and privacy risks of smart home devices and voice assistants have been extensively studied in prior research. Edu et al. highlighted the broad attack surface of voice assistants, outlining potential countermeasures [39]. Studies have also examined user perceptions of privacy in smart speakers, attitudes toward related risks [59, 68, 82, 84, 93], and user concerns about data deletion and its impact on trust [32, 34]. Longitudinal investigations have identified privacy as a recurring concern in overall user experience [26, 37]. Lemmer et al. studied children's privacy under GDPR, finding both Alexa and Google Assistant in violation of several legal requirements [60]. Empirical research has explored user behaviors and privacy-related mental models. Participants report concerns about unauthorized access to personal data by household members, visitors, or device manufacturers, reflecting limited trust [52]. Privacy-seeking behaviors, such as use of the mute button, are rare and often insufficient to meet user needs [58], and studies of mental models highlight strong concerns about security and potential vulnerabilities to hacking [19, 20]. Controlled experiments have also been conducted to audit user profiling practices: Khezresmaeilzadeh et al. systematically interacted with voice assistants using several personas and found evidence of user profiling influenced by the modality of interaction [55].

Privacy Risks with Third-party Voice Apps. Preserving privacy and managing permissions becomes particularly challenging when third parties are involved. Zufferey et al. examined data access mechanisms in wearable devices and found that about half of the users underestimate the third-party apps with access to their sensitive information [94]. Studies on voice assistants highlight that third-party voice apps introduce additional privacy risks: some apps can bypass vetting processes and elicit sensitive user data [33, 47, 89–91], and users often cannot distinguish these third-party skills from native skills [67, 76].

Designing Privacy Notices for IoT Devices. To better inform users about data practices and reduce privacy concerns, researchers have explored new ways to present information through innovative privacy notices. In an early work in this domain, Kelly et al. introduced the "privacy nutrition label" [54]. Emami-Naeini et al. introduced similar privacy labels for IoT devices, to help consumers make informed choices when purchasing an IoT device [40]. However, these labels are primarily helpful at the time of purchase and provide only general privacy-related information; they do not assist users while they are actively using the device, especially in the verbal-first context of a voice assistant.

Prior research on the web and social media has shown that privacy notices embedded directly into an app's control flow that subtly steer users toward a privacy-protective behavior without forcing them or restricting their options [53], can help users make more privacy-conscious decisions [24, 69, 85, 87]. In the context of smart home devices, Schaub et al. identified four key components for effective privacy notices: *timing*, *channel*, *modality*, and *control* [79]. These notices can be delivered at different times (e.g., at setup or just-in-time), through various channels (like a primary voice interface or a secondary screen), and in different formats (visual, auditory, or haptic).

While research has also focused on designing privacy icons to convey options without creating misconceptions [49], these solutions often place the burden of managing privacy on the user [38, 63]. Despite advances in other domains, the development of privacy notices specifically for voice interfaces remains largely unexplored.

Distinction from prior work. Prior research on Alexa privacy notices mainly focused on alerting users when they activate third-party skills versus native features [67, 76], but these efforts have not addressed the privacy concerns regarding *verbal data requests.* User studies have shown that Alexa's voice interface provides inadequate cues for distinguishing third-party skills from native ones [76]. This misperception creates privacy risks, as users tend to trust the platform more than third parties. Although the privacy implications of data collection by third-party skills are well documented [62], prior works have not designed or evaluated privacy notices specifically for verbal data requests, which are a central part of voice interactions. Because voice interfaces are fundamentally different from visual platforms like mobile and web, privacy notice designs from those ecosystems can't be directly transferred without risking disruption of the conversation flow or increasing the user's cognitive load. This paper addresses this gap by identifying user-preferred privacy notice types and systematically evaluating their effectiveness in various realistic voice interaction scenarios.

4 Study Organization

In this paper, we present findings from two studies. **Study-I** involved **semi-structured interviews** (n=21) with users of Amazon Alexa to understand user perceptions and preferences regarding current data request methods on voice interfaces. These interviews helped us identify contextual factors influencing user preferences, enabling us to design prototypes for privacy notices. While prior literature—often led by privacy experts—has identified general design factors for privacy notices [79], these frameworks are typically broad and not tailored to the unique characteristics of voice interfaces. In contrast, **Study-I** aims to explore which of these established factors are applicable in the context of voice interactions, while also uncovering additional factors that may be specific to this modality. Importantly, rather than relying solely on expert perspectives, this study engages general active users of voice assistants to ensure the identified factors reflect real-world usage patterns, expectations, and challenges. In **Study-II**, we evaluated user preferences and usability of these privacy notices via an **interactive study**. Based on results from **Study-I**, we conducted **Study-II**—an *interactive user study* (n=160)—where we prototype the privacy notices by simulating them in custom Alexa skills. Based on the results, we provide recommendations on which privacy notice(s) would likely be the best and dissect their factors. For both Study-I and Study-II, eligible participants were required to be U.S. residents, at least 18 years old, and active users of smart voice assistants (VAs).

Ethical Considerations. Both studies were approved by our university's IRB. For Study-I, interviews were transcribed using a local Whisper model, de-identified, and recordings were deleted post-transcription. For Study-II, participants were anonymously recruited via Prolific. While we developed real Alexa skills, public deployment risked user data exposure and violated Alexa's policies against non-functional or mock skills. To address this, we recorded interactions in developer mode and showed participants the videos for feedback. Detailed ethical considerations for both studies are discussed in Appendix §A.1.

5 Study-I: Verbal Privacy Notice Preferences

5.1 Methods

5.1.1 Study Design. For Study-I, our study flyer included a link/QR code to a screening survey that provided the details of the study. In the screening survey, we asked about the frequency and duration of participants' VA usage, the VAs used (e.g., Amazon Alexa, Google Assistant, Apple Siri), and the activities they performed using VA. As an attention check, we asked what "wake word" they used to activate their VA. A non-existent wake word would flag a fraudulent participant. If the participant indicated using Amazon Alexa, the screening survey asked whether and what *skills* they enabled on Alexa. We also collected demographic information, including education level, age range, and gender, to ensure participants' diversity, as participants from different underserved populations may have different privacy needs and challenges [44, 50]. In the consent form, we informed participants to keep their VA device handy during the actual interview to show on video for verification.

Although we aimed to understand preferences for VAs in general, we focused our interview study on Alexa users exclusively,

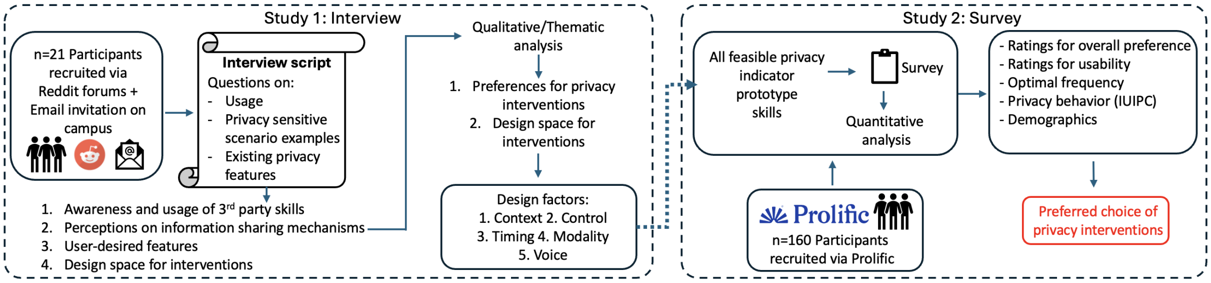

Study 1: Interview
n=21 Participants recruited via Reddit forums + Email invitation on campus
Interview script
Questions on:
- Usage
- Privacy sensitive scenario examples
- Existing privacy features
1. Awareness and usage of 3rd party skills
2. Perceptions on information sharing mechanisms
3. User-desired features
4. Design space for interventions
Qualitative/Thematic analysis
1. Preferences for privacy interventions
2. Design space for interventions
Design factors:
1. Context 2. Control
3. Timing 4. Modality
5. Voice
Study 2: Survey
All feasible privacy indicator prototype skills
Survey
Quantitative analysis
- Ratings for overall preference
- Ratings for usability
- Optimal frequency
- Privacy behavior (IUIPC)
- Demographics
Prolific
n=160 Participants recruited via Prolific
Preferred choice of privacy interventions

Figure 1: Detailed diagram of the two studies. Study-I identifies contextual factors shaping user preferences, informing the design of privacy notice prototypes in Study-II.

as Alexa is the only voice assistant platform currently supporting third-party apps (Google stopped support for third-party voice apps in 2023 [15]). We did not explicitly state in the screening survey that we were recruiting only Alexa users to prevent fraudulent participants from gaming the system. Prior research [75] and our own experience show that clearly stating such criteria often prompts fraudulent respondents to simply claim eligibility, making them harder to detect and wasting time for legitimate participants. Instead, immediately after the consent form, the sign-up survey presented a multiple-choice question about voice assistant usage. If participants did not select Alexa, the survey ended immediately, sparing them from answering 10 additional Alexa-specific questions. This also prevented repeated attempts, as Qualtrics blocked re-entry. This approach saved participant time while reducing our screening burden: of 97 recruits from social media, only 44 genuine Alexa users were invited, with the rest filtered out for reasons such as incorrect wake words or likely AI-generated responses.

During the Zoom-based interview, researchers asked the participants to provide consent to proceed with the interview. As a verification step, at the beginning of each interview, we asked the participants to turn on their webcam so that their faces and VA devices were visible (this was not recorded). The primary researcher was present for all the interviews and could ensure that no participant participated twice and that all were legitimate VA users.

The interview had three components (see Figure 1): (i) participants' usage of skills, (ii) perceptions about modes through which a VA requests sensitive information, and (iii) features desired by the participants to mitigate their privacy concerns. In the first section, we elicited participants' awareness of Alexa skills, whether they had used them, and if they had shared personal information with these skills. We also asked them if they shared personal information by providing permission to the skill through their Amazon account, or if they verbally provided the information in response to a verbal request by the skill. Then we asked participants to share their perspectives on both modes of information sharing: (i) using the permissions mechanism where a skill explicitly gets user consent when accessing personal information from the Amazon account, and (ii) skills requesting personal information verbally without explicit permission. Finally, if the participants shared privacy concerns about modes of information collection, we asked them to share their desired features to mitigate their concerns. Building on prior literature that has developed general privacy notice design frameworks and domain-specific adaptations (e.g., for VR), we utilized both the frameworks and coding schemes from the Privacy Design Framework [79] and the Security Indicators framework for VR [87] in structuring our survey, which formed the basis for our deductive coding. In contrast, identifying voice assistant–specific features desired by participants required inductive coding.

Since not every participant was equally familiar with skills and how their personal information may be requested, during the interview, we ensured to provide sufficient background information so the participants could share their perspectives in a well-informed manner. For this purpose, we made three short videos on the following topics: (i) introduction to Alexa skills, (ii) personal information collection in skills, and (iii) privacy features implemented in skills (see Table 12 in Appendix). When a participant indicated a lack of knowledge on any of these topics, we played the corresponding pre-recorded video, ensuring uniform dissemination of information to all participants. After playing the videos, we held a follow-up discussion to confirm the participants' understanding before proceeding. The educational videos provided a consistent baseline without priming, focusing solely on how the existing system works.

5.1.2 Recruitment. For Study-I, we advertised our study on voice assistant-specific subreddits (e.g., r/VoiceAssistants, r/alexa, r/Alexa_Skills, r/AlexaDevs, r/amazonalexa, r/googlehome), LinkedIn groups (e.g., Amazon Alexa, Google Home Professional Group), and Discord servers (belonging to university students). We also put our study flyer on our University's bulletin boards to recruit student participants. Additionally, we used snowball sampling. We received 137 complete responses for the screening survey. 97 respondents used Alexa, 75 used Google Assistant, 56 used Apple Siri, and 2 used other VAs (most respondents used multiple VAs). In selecting whom to interview, we analyzed the quality of responses to the screening questions and prioritized diversity in terms of the types of activities performed using the VA, modality of VA usage (e.g., on their phones, through a smart speaker, etc.), and participant demographics. We invited participants for the interview on a rolling basis until we reached data saturation [42] (i.e., additional interviews did not yield further insights). Overall, we invited 44 respondents for interviews. Of these, 31 scheduled an interview with us. However, five were no-shows, and one canceled. Further, during the interview, four participants were unable to show their voice assistant device on camera, so we did not proceed with the remainder of those interviews. During the interviews, the researcher was taking notes to determine saturation in responses [78]. Of the successfully conducted interviews, we reached data saturation at the 19th interview. We conducted two more interviews to ensure saturation. As shown in Table 6 in Appendix A, our Study-I participants (N=21)

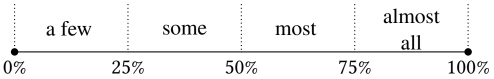


Figure 2: Terminology used to present relative frequency.

covered a diverse range of racial/ethnic backgrounds, including Asian, African American, Latino, and Caucasian participants, with one participant preferring not to disclose. Age was skewed toward younger adults, particularly 18–24 (8/21) and 25–34 (6/21), with fewer participants in older age groups. Gender distribution was moderately balanced (11 men, 9 women, 1 undisclosed). Education was skewed toward higher education, with most participants holding at least a bachelor's degree (14/21).

5.1.3 Data Collection. We hosted the screening survey using Qualtrics. Before conducting interviews with actual participants, we conducted four pilot interviews with S&P researchers to assess the interview duration and refine the questions [65]. At the start of the interview, participants consented to participate in the study and be audio recorded. We conducted a total of 21 interviews as discussed in § 5.1.2. Interviews were conducted in English via Zoom. The average length of interviews was 39 minutes, and participants received a $15 Amazon gift card upon completion.

5.1.4 Data Analysis. The interviews were audio-recorded and transcribed using Whisper [73]. Two researchers conducted thematic analysis [27, 81] on the transcripts using deductive and inductive approaches. Deductive thematic analysis enabled coding based on the privacy design framework [79, 87], while inductive coding was used to extract themes within our research questions. The researchers coded five transcripts and then discussed to create the initial code book. Next, they independently coded the rest of the transcripts in batches of five (the last batch with six) and discussed them to update the codebook after each iteration. The two researchers discussed the codes and resolved any conflicts through several weekly meetings — an approach followed in several qualitative S&P studies [46, 51, 86]. The purpose of meeting multiple times was not to establish a measure of agreement but to eventually yield concepts and themes (recurrent topics or meanings that represent an idea) so calculating inter-rater reliability is not meaningful in this case [70]. The codebook and other artifacts are available using an OSF repository[1] in section A.2 of Appendix A. Since our study is qualitative with a small sample size, we refrain from reporting the exact number of participants associated with a given theme. However, to provide a sense of frequency, we adopt a consistent terminology shown in Figure 2 when reporting our qualitative findings.

5.2 Results

5.2.1 Awareness and Usage of Third-party Skills. We started our interviews by understanding our participants' awareness of skills. Most participants recognized the term "skills". However, only some participants, such as P12, had the correct technical understanding of skills: *"It's a program that Amazon or a third party can build, that provides specific functions based on a keyword. When you ask Alexa... Use a keyword, and it will listen to your command and then perform a specific action"*. While about half of our participants could not provide a succinct definition of skills, a few participants had an incorrect understanding. To provide more context to our participants and ensure they had sufficient background knowledge to answer our interview questions, we educated them using our pre-recorded videos as specified in § 5.1.1.

Interestingly, some participants didn't know what third-party skills they had enabled, even though they knew they were using those functionalities on their Alexa devices. P5 stated that she only realized she was using skills when she followed our instructions to check her Alexa app for the skills she had enabled.

"I didn't know "brown noise" was not part of Alexa... I didn't know they were skills to begin with." (P5)

Our instructions helped participants to identify the skills they used (by checking the Alexa app), which included Spotify, iRobot home, wikiHow, Sleep sounds, iHeartRadio, and daily cricket trivia.

When asked which entities could access their PII, about half of our participants believed that the first party (i.e., Amazon) as well as third parties (i.e., skill developers) could do so. However, the other half of the participants were not aware that those skills were "third parties" and that their information was being shared with those entities. For instance, P5 shared: *"It's not cool that I didn't know they were third parties"*. A few participants mentioned that when they interact with skills, it is not obvious which entity is asking for the information: Alexa or the third-party skill.

Next, we probed our participants about their knowledge of privacy features in Alexa devices. Since our eventual goal was to determine if users desired privacy notices at the time of information sharing, we first ensured they were aware of existing privacy features on Alexa. A few participants, such as P8, mentioned skill permissions: *"the third-party apps or skills might have to request access to, for example, your location or your name."*

5.2.2 Perceptions of Alexa's Information Sharing. At this stage of the interview, we briefed our participants on the two ways a skill may access users' PII: using skill permissions and obtaining the PII from the user's Amazon account, or by asking the user for information through a verbal request (see Figure 4 in Appendix A). We then asked participants what they liked and disliked about each of these two information sharing mechanisms.

Perceptions on Accessing PII from Amazon Account. We found that most of our study participants liked the permissions system that is enforced when an Alexa skill needs to access user information from their Amazon account. Participants also liked that the permission system helped users remain cautious and informed of the data they shared, even in the future, and provided granular control. P1 felt that the permission system was good for privacy. P15 liked the usability of the permission system, of not having to share information verbally.

P14 noted that the permission system ensures that an authorized user is providing the information: *"the security factor of your granting the permission explicitly can make sure that it is you that is actually granting the permission. In case of a verbal query, anyone can actually talk to your Alexa, and they can provide that information."* About shortcomings of the existing skill permission system, a few participants shared that the permission system did not indicate why the skill required that information (P15), it was cumbersome to use (P10), and had very limited permissions available (P7).

[1] Interview guide, codebook, and other user study artifacts are available here: https://osf.io/h7wkm/?view_only=baedc79e9f304172ae000f8392c30c9e

Perceptions on Accessing PII through Verbal Request. In the case where users provided information in conversation, some participants, such as P2, P9, P10, and P15, preferred this method, citing ease of use. P17 added: *"I like the simplicity of being asked by voice."* P8 and P13 highlighted the lack of validation for the provided information in some skills, and that the collected information may be incorrect or inaccurate:

"I feel like sometimes it can misinterpret what you say. So maybe if you say no, it could somehow interpret it as yes, which is outside of your control. That definitely poses some issues." (P8)

Most participants preferred verbal requests when providing non-sensitive information, preferring the permission system to provide access to information they considered sensitive. A few participants, such as P2, highlighted the challenges in verifying the authenticity of skills compared to traditional online services: *"I can verify the authenticity of a website better than a skill offered by Amazon."*

Some participants shared their frustrations about the verbal requests for information bypassing the traditional permission system. For instance, P4 shared: *"I don't like that third parties can bypass that privacy settings menu by just asking you directly."* A few participants felt nudged to share information without knowing it would go to a third-party developer. P11 further noted that malicious developers could craft requests to trick users into sharing sensitive data:

"They can arbitrarily ask for anything they want. [And] they might potentially use psychological tricks to make the user reveal things they don't know they are." (P11)

5.2.3 User-desired Features to Improve Privacy Mechanisms.

We elicit user-desired features to improve privacy mechanisms (**RQ1**).

Improving Permission System. To improve the current permission system, a few participants specified the need for context and explanation for why a permission is needed. P16 shared:*"if you're doing a toggle thing, if there is no context there, it's just a phone number toggle. Like, why this information is required?"* P4 added:

"I would prefer maybe a subtoggle of more granular information that could be restricted down. Not everything needs to know exactly where my house is." (P4)

Further, P15 highlighted that the user must not be prompted repeatedly to provide permission if they have already declined multiple times.

P7 believed that all permission information should be available on the web interface as well, apart from the Alexa app, while P9 wanted all permissions to be off by default.

Improving Verbal Requests. Most of our study participants mentioned they would like a warning or a notification that would help them make an informed decision before providing information. For example, P1 shared, *"I think it should offer users a warning that they should just be careful who they give their information to."* P14 shared that a warning to the user, urging them to reconsider what information they were providing, would be helpful.

"If you're having a conversation, it's very easy to give out sensitive information in the flow just because of the way it's progressing. So if there is a notification like an extra confirmation that lets the user think about the information they're giving out." (P14)

P16 added that verbal requests should be made with sufficient context and explanation so that they can do more scrutiny before providing information:*"I will probably look more carefully before saying yes, yes."*

Since users' personal information is being sent to a third party, some participants expressed the need for a mention of who the information will be shared with. P1 added that this must be done both at the skill configuration stage and during regular skill usage:

"When you are initially configuring the skill, you should be notified that it may require, [certain] information and that [it] is not going to Amazon, but somewhere else. Then, once it asks for that information... warn the end user that, it's not Amazon that's asking for this." (P1)

Users also expressed a desire that voice apps should only request information that is required for a task, and they also expect the platform to enforce it, stating:

"I guess the only way would be for Alexa to do a similar vetting system like Apple to verify that skills don't get unnecessary information" (P12)

Another participant desired the role of platform as a gatekeeper to ensure requests for legitimate and minimal information, stating:

"[I would want that] if Alexa was a trusted gatekeeper to only allow third parties to collect necessary information." (P17).

Nuances in Implementing a Warning. Participants brought up several nuances in implementing such a warning. We highlight several specific features suggested by our participants.

Length & Frequency of Warning. P0 wanted a warning, but a short one. Some participants expressed a concern over degradation in usability if warnings were introduced. P19 felt that repeated warnings may be a hassle: *"if every time a third-party app wanted to ask a piece of information, there was a pop-up or Alexa chimed in with a warning that would be a hassle. There's got to be some point of user awareness of what's going on."*. P4 highlighted the possibility of *alert fatigue*:

"If a warning does come too regularly ... it could happen that people start ignoring the messages... like how when we install application we don't really read the terms and conditions... we just click next next next." (P4)

P4 also wanted users to have the ability to control the frequency of warnings if they thought it was excessive.

Content & Timing of the Warning. A few participants had strong preferences for the kind of message the warning should convey. P6 believed that it should convey *"who wants the information, why they need the information, and get confirmation if I want to share this information."* P13 added that the warning should be intuitive: *"it should be intuitive and helpful for the users. Otherwise, the user would think it is going to Amazon..."* P2 wanted the warning to emphasize the data processing and storage capabilities of the third-party entity.

Some participants further specified that the warning should also "ask for permission" from the user, where they can verbally (or otherwise) confirm whether they want to provide that information to the skill or not. P20 shared:

"The warning should make me aware that the info is going to a third party and I can refuse to provide." (P20)

Modality of the Warning. Our participants identified several modalities through which this warning could be made to the user. While most participants described a verbal warning, some of them suggested that the warning should appear as a popup or a notification on the phone. P7 shared that a mandatory approval of a request through the app must be implemented: *"Amazon could probably have a pop-up or a notification whenever it detects that a question is asking for some information."*

Table 1: All factors, possible values and descriptions

Factors	Possible values	Description
Context	Relevant data	The verbal data requested is related to or required for the task at hand
	Irrelevant data	The verbal data requested is not related to or required for the task at hand
Control	Blocking	Seeks verbal permission and only continue with skill if user allows explicitly
	Non-blocking	Just notifies the users and continue without requiring explicit affirmation
Timing	Before request	Notification comes before information is requested
	After request	Notification comes after information is requested
Modality	Audio	Audio notification only
	Audio+App	Audio notification with an app popup
Voice	Alexa-voice	Notification in same voice as Alexa or rest of the skill
	Different-voice	Notification in voice different from Alexa or rest of the skill

A few participants felt that to preserve usability, a redirection of permission requests to the app could be done only when the information requested was sensitive.

Voice Used for Providing the Warning. Several participants had suggestions about the voice to be used to deliver the warning. P15 shared that when the warning is being made in relation to a third-party skill, a "third-party voice" must be used. P11 added:

"If third party skills use a separate voice from Alexa's voice, I think that's a very natural way to let people know that it's someone else." (P11)

While a few participants wanted a different voice for all warnings, another few wanted a different voice to highlight sensitive information requests alone.

Relevance of the Information Requested. We found that participants considered the context and functionality of the skill when determining if they had any privacy concerns about providing personal information to the skill. They mentioned that they would not be concerned if the provided information was necessary for the skill to perform the desired task; however, they would be concerned in case a skill asks excessive personal information. This is in line with prior research about users being concerned about third-party skills collecting excessive personal information [21].

"If the correct address has special benefits for people over 65, and if Alexa asked me, are you older than 65? I would say yes. If it asked me the same sort of question about. Would you like [Insurance Company] to know that you're over 65? I would probably say no." (P3)

Sensitivity of the Information Requested. In case the information being requested by the skill is more sensitive, participants wanted more explicit, noticeable, and frequent warnings. They also mentioned that in these cases, warnings should not be suppressible.

"[For] things which falls under the sensitive category [the warning] should not be suppressible. It should prompt you every time that it [is asking] for sensitive information and only under your consent it should be sent." (P4)

Participants also mentioned that the warnings should be based on what users deem most sensitive, issuing them only in such cases. A few others proposed biometric or PIN-based confirmation for sharing sensitive information.

5.2.4 Design Space for Privacy Notices. Based on the insights shared by our participants, we identified multiple factors that could be used to design privacy notices (**RQ2**). We aimed to create prototypes of privacy notices and evaluate them via an interactive user study. Although there were additional factors, we identified **five** key factors to design our prototypes. These factors include the following:

- **Context** in which the information is requested by the skill; determines if the requested information is relevant or necessary to perform the core functionality of the skill; indicated by *"relevant info/irrelevant data"* in Table 1.
- **Control** flow of the notice; whether the notice will block the flow of conversation and expect user response, or merely inform the user; indicated by *"blocking/non-blocking"* in Table 1.
- **Modality** through which the notice is implemented; either verbal through the voice assistant, or visual through the phone; indicated by *"audio/audio+app"* in Table 1.
- **Timing** of the notice; either before the information is requested, or after the information has been provided; indicated by *"before request/after request"* in Table 1.
- **Voice** used to deliver notification; same voice as Alexa, or a different voice; indicated by *"Alexa-voice/Different-voice"* in Table 1.

These factors align with prior research but have been adapted for Virtual Reality [87]. While we identified three additional factors—frequency, length, and content of the warnings—we did not incorporate all of them into our prototype design. Frequency was evaluated separately by asking users for their preferences after experiencing the notices, avoiding repeated exposures during the brief user study for feasibility reasons. The content and length of the notices were carefully designed based on prior literature [87] (see § 6), though further refinements remain as future work. Since runtime information requests from third-party skills are unpredictable, our notices could be implemented by first identifying the type of requested information before presenting it to users. Existing systems, such as "Skipper" by Xie et al. [90], demonstrate that detecting and assessing the compliance of personal information requests is highly feasible, achieving near-perfect recall with only a 2.11% false positive rate. While such detection techniques are essential for enabling our notices in practice, they fall outside the primary scope of this work, which focuses on the design and evaluation of the notices themselves.

6 Study-II: Privacy Notice Design & Evaluation

6.1 Methods

6.1.1 Creating Prototype Skills. To simulate privacy notices based on design factors identified in Study-I, we developed a custom Alexa skill that requests personal information from the user. We created multiple versions of this skill, each implementing a different privacy notice to illustrate how the skill's behavior would change if Alexa adopted these notices. To ensure real-world relevance, we based the skill on a common use case for which similar skills already exist in Alexa's skill store. The chosen skill required verbal requests for personal information. We initially devised five possible scenarios based on existing skills (i.e., "Trip Planner", "Book Suggestion", "Restaurant Booking", "Task To-Do", and "Find My Phone") and conducted a small user study (n=5) with student volunteers to determine their preferences. Participants favored the "Task To-Do" skill, which maintains a to-do list and sends reminders via SMS. They found this skill convenient for managing tasks by voice, especially when multitasking or away from other devices.

We created a baseline version of the skill that requests users' personal information without any privacy notice, reflecting Alexa's current approach. Next, we developed variations of the skill incorporating different privacy notices, each representing different values of the design factors: **context**, **control**, **modality**, **timing**, and **voice**. Due to the challenge of demonstrating frequency in a single interaction, this factor was excluded, and instead, users were asked about their preferred frequency at the end of the session (see §6.2.3). The content of the privacy notification was simple: *"Alexa wants to inform you that the Task To-Do skill is going to request some personal information, which will be sent to the third-party developers of this skill."* We injected a tone to grab users' attention, and the notices lasted around 10 seconds.

Factors in the Privacy Notices. Each of the five factors has two possible values. The context of the information request could be *relevant* or *irrelevant* to the skill functionality; the privacy notice could *block* the control flow and wait for the user response or be *non-blocking* by merely informing the user without interrupting the flow of the task; the timing of the notice could be *before* the information is requested, or *after* the user has responded to the request; the modality of the notice could be purely *audio*-based, i.e., a verbal warning, or include an *app*-based notification, where both a verbal warning and an app notification or a prompt to allow/deny permission is provided. The app notification also has audio with it because a voice assistant still communicates verbally with the user while sending a pop-up on the phone. The voice used to provide the verbal warning could be using the *default* voice used for regular voice assistant activity, or a specialized, *different* voice. These combinations (Context=Relevant/Irrelevant, Control=Blocking/Non-blocking, Timing=Before/After, Modality=Audio/Audio+App, Voice = Default/Different voice) result in a total of $2^5 = 32$ potential combinations. However, we note that if the notice is *blocking*, it cannot come *after* the information request. We exclude these infeasible combinations: we eliminated eight combinations where *control=blocking* and *timing=after response*. We also removed four combinations with *control=blocking* and *modality=audio+app*, since these would require users to interact with a phone app or the voice assistant's screen (not available on most VAs), prompting them to grant permission on screen. After removing these 12 combinations, we were left with **20** feasible combinations, as shown in Table 2.

One of our design factors was *context*, as users considered whether the information requested by the skill was genuinely necessary for completing the task or if it could proceed without the information. Since this factor pertains to the type of information requested rather than the notice itself, we created an additional baseline version of the skill where it requests irrelevant personal information. There can be two ways to incorporate *context*: We can either keep the information requested the same and change the functionality of the skill, or keep the skill the same and change the information type requested. We chose the latter because changing the skill's functionality would introduce a significant difference in user experience and change uncontrolled variables.

Table 2: All 20 privacy notice variants

Var. #	Context	Control	Timing	Modality	Voice
V00	Relevant data	Blocking	Before request	Audio	Default-Voice
V01	Relevant data	Non-blocking	Before request	Audio	Default-Voice
V02	Relevant data	Non-blocking	After response	Audio	Default-Voice
V03	Relevant data	Non-blocking	Before request	Audio+App	Default-Voice
V04	Relevant data	Non-blocking	After response	Audio+App	Default-Voice
V05	Irrelevant data	Blocking	Before request	Audio	Default-Voice
V06	Irrelevant data	Non-blocking	Before request	Audio	Default-Voice
V07	Irrelevant data	Non-blocking	After response	Audio	Default-Voice
V08	Irrelevant data	Non-blocking	Before request	Audio+App	Default-Voice
V09	Irrelevant data	Non-blocking	After response	Audio+App	Default-Voice
V10	Relevant data	Blocking	Before request	Audio	Different-Voice
V11	Relevant data	Non-blocking	Before request	Audio	Different-Voice
V12	Relevant data	Non-blocking	After response	Audio	Different-Voice
V13	Relevant data	Non-blocking	Before request	Audio+App	Different-Voice
V14	Relevant data	Non-blocking	After response	Audio+App	Different-Voice
V15	Irrelevant data	Blocking	Before request	Audio	Different-Voice
V16	Irrelevant data	Non-blocking	Before request	Audio	Different-Voice
V17	Irrelevant data	Non-blocking	After response	Audio	Different-Voice
V18	Irrelevant data	Non-blocking	Before request	Audio+App	Different-Voice
V19	Irrelevant data	Non-blocking	After response	Audio+App	Different-Voice

In baseline B_0, the "Task To-Do" skill asks for the user's phone number to send a reminder via SMS, providing a clear rationale. In baseline B_1, however, the skill requests the user's home address to "complete the profile" without explaining its necessity, a piece of information irrelevant to the core functionality of the skill. With these two baseline variants designed based on context, we retained **four** factors relating to the notice properties, resulting in **10 variants per context** (see Table 2). which we used to assess user preferences in each scenario.

We created videos of the skill interactions on an Echo Dot smart speaker, using an AI-generated voice from NaturalReader [16] following a script. In total, we produced 22 *videos* of skill interactions (2 baselines + 20 notice prototypes). The baseline videos were each 38s long, while the 20 notice variant videos averaged 54.15s. Anonymous links to videos are available in Table 12 in Appendix A.

Our study approach. An ideal way to conduct our user study would be for participants to enable our skill variants on their own Alexa devices and interact with them directly. However, this raises privacy and ethical concerns (discussed in Appendix § A.1). Since our skills request sensitive information, such as phone numbers, there is a risk of participants inadvertently sharing their real personal data, despite our commitment to not storing any information. To address this, our study allows participants to focus solely on the task, minimizing the risk of revealing personal details or encountering frustration from potential skill-installation errors. While direct interaction with the skill would be optimal, our video-based approach is a viable alternative, especially for an early usability study, where indirect methods like video observation and think-aloud protocols are low-cost, scalable alternatives to direct interaction [30]. Similar to how users observe tools via tutorials before hands-on use [36], watching these videos serves as a proxy for realistic interaction, particularly since they simulate actual tasks (i.e., using the "Task To-Do" skill). Our study design focuses on maintaining internal validity rather than ecological validity, which is appropriate for evaluating different factors contributing to user preferences [25].

6.1.2 Recruitment.

We recruited 160 participants (40 participants per notice) through Prolific [14]. We chose Prolific due to its popularity as a crowd-sourcing research platform in the US, and as it is shown to produce better data quality compared to Amazon MTurk and CloudResearch [41]. We only recruited participants who lived in the US, between the ages of 18 and 65. To ensure we recruited only VA users, we used the prescreening filters like "Home assistants/smart hub (e.g., Amazon Echo)" available on Prolific to select only home assistant owners. As shown in Table 7 in Appendix 7, our Study-II participants (N=160) represented a diverse range of racial and ethnic backgrounds, with the majority identifying as Caucasian (70%), followed by Black or African American (15%), Hispanic/Latinx (13%), Asian (11%), and other groups. Participants were primarily younger and middle-aged adults, with 17% aged 18–24, 38% aged 25–34, and 28% aged 35–44. Gender distribution was slightly skewed toward women (56%) compared to men (42%), with 1% identifying as non-binary and 1% preferring not to answer. The sample was relatively highly educated, with 41% holding a bachelor's degree and 18% a master's degree; the remainder had high school, some college, vocational training, associate's, or doctoral degrees.

6.1.3 User Study Structure. We created a user study for the participants to view the notices and rate them in terms of overall preference rating, inconvenience, and provide an optional free-text response of their explanations. We designed the survey to collect data points for five variants from each participant, to avoid attention fatigue. We randomly assigned five variants associated with either baseline B_0 or B_1 to a given participant. We programmed our survey on Qualtrics such that 80 participants responded to baseline B_0 and the other 80 saw B_1 and its respective variants.

In the survey, participants are first presented with the video of the baseline skill along with the description. This is followed by the videos of the five variants belonging to that baseline. For overall preference and inconvenience, we asked participants to answer on a five-point Likert-type scale: *"To what extent do you agree with this statement on a five-point scale?"* For overall preference, the statement was, *"I would like Alexa to implement the notification prototyped in this skill,"* (positive framing), and for inconvenience, *"This notification would make it very hard to use the skill; I prefer the original version even though it lacks the notification."* (negative framing). The choices on the Likert-type scale questions used wording of "Strongly disagree", "Disagree", "Neutral", "Agree" and "Strongly agree" in order. We chose a single Likert-type scale rating over a system usability scale (SUS) for its simplicity, efficiency, and focus on overall perceived usability. Many items in the standardized SUS scale are not directly applicable to our privacy notices and may be confusing for participants (e.g., "I found the various functions in this system were well integrated"). Likert-type scales combined with qualitative responses are more appropriate for evaluating privacy choice interfaces, particularly when participants have prior exposure to the system [48]. Although this approach provides a better fit in our context, the single-item Likert-type responses limit direct comparison with existing literature and may not capture all dimensions typically assessed by SUS. We also collected qualitative feedback through free-form responses. Since our goal was to identify the least disruptive privacy notice rather than diagnose usability issues, this approach, complemented by qualitative prompts, aligns with established practices in prior research [48].

We also provided a free-text box for participant to add details on their preferences for each notice. Each video was presented on a different page, and ratings were independent of other variants. After the skill variant videos, participants were asked how frequently they would want the notice to occur. Finally, participants were asked about demographics (age group, gender, race, and education). We used two attention checks in the survey to determine the quality of the response. The first, asked participants to identify the task described in the survey, appearing immediately after the task description. The second attention check, presented alongside the video, asked what the base skill said at the start, minimizing cognitive load while ensuring participants watched the video. We collected 160 valid responses and compensated participants with $5 through Prolific. The average completion time was around 15 minutes.

6.1.4 Data Analysis. As our participants rated the notice designs on a Likert-type 5-point scale and each participant rated five random variants, we have ordinal and hierarchical data (containing random effects). To answer both parts of our RQ3, we chose an analysis approach that caters to ordinal and hierarchical data. We first summarize the data by reporting average ratings and inconvenience scores for all design variants. To answer both parts of RQ3, we conduct two different types of analyses, as there is no one-size-fits-all approach, answering both parts adequately. To determine the most preferred notice designs (first part of the RQ3), we conduct Bayesian ordinal multilevel modeling regression with posterior probability analysis [29]. The Bayesian multilevel regression model supports hierarchical structures in ordinal data and quantifies the uncertainty using Markov Chain Monte Carlo (MCMC), providing credible intervals and posterior measures. We conduct Bayesian multilevel regression using the `brms` library in `R`. The `brm` considers one notice design as a reference and compares each notice design to the reference, and estimates the change in ratings. We also incorporated random effects, as we have mixed data due to each participant rating multiple notices. We set the `chains` parameter representing the number of Markov Chains to 4 and the number of iterations per chain to 2000. It is a common practice in literature to run at least three to four Markov chains [45, 57] and observe the model convergence using the Potential Scale Reduction Factor (Gelman-Rubin statistic which is commonly known as $\hat{R}$) convergence diagnostic metric. Gelman et al. suggest $\hat{R}$ close to 1 indicates convergence, but $\hat{R} > 1.1$ indicates non-convergence and may require a higher number of chains or iterations [28]. Another aspect of reliability is to check the stability of the estimates even when chains have converged, and the key indicator for this is ESS and which is recommended to be greater than 1000 for stable estimates [57]. Then, we calculated posterior probabilities for the linear predictor using `posterior_linpred` to determine the probability of each notice receiving higher ratings than all other notice variants, which helps us rank the notices probabilistically [43]. The probabilistic preferences obtained via `brms` are directly interpretable and also quantify uncertainty, as preference may depend on various factor combinations and some combinations may not be significantly better than others in a classical way.

Then, we analyze what factors contribute the most to the ratings to understand *why* certain notices were preferred over the others (second part of the RQ3). We conduct this factor analysis using

Cumulative Link Mixed Models (CLMM) Regression [35]). CLMM estimates the change in ratings based on the unit change in the independent variable, determining how important each factor is. `p_values` for each indicates whether the factor affects the rating significantly. We implemented `clmm` regression through `ordinal` package in `R`. We also determine the orthogonality of the factors and the alias structure due to the fractional factorial design [56]. Based on the potential aliasing, we check the model's robustness by testing the interacting factors (*Timing***Modality*) as well as testing the effect of each factor, when other factors are not present.

The text responses regarding participants' likes and dislikes of the variants were analyzed through manual review by two researchers, with each variant receiving 40 or fewer responses (as this open-ended question was optional). Recurring themes were identified through discussion. This approach followed the methodology outlined in Section 5.1.3. The corresponding codebook is available in the OSF repository, as described in Section A.2 of Appendix A.

6.2 Results

We assessed participants' overall preference ratings to identify which notice variants were most preferred and then determined which factors contributed the most to the ratings.

6.2.1 Preferred Notice Design. First, we evaluate whether the specified parameters, such as the number of MCMC chains and iterations per chain, and the sample size, for the Bayesian Ordinal Regression (`brm`) model, are sufficient for the model to converge so the results can be considered reliable. For both of our models (relevant and irrelevant data), we found $\hat{R} = 1.0$ and $EffectiveSampleSize > 1000$ that shows the model converged well and the results are reliable. As model results were stable and converged, we did not have a need to adjust the number of chains or iterations in the `brm` model configuration. The `brm` model regression results are shown in Table 3, for relevant data and irrelevant data, respectively.

For the relevant data type where V00 is a reference notice design, most notices (V02, V04, V11, V12, V13, and V14) result in considerably low ratings with negative CI intervals at 95% confidence. However, some variants are rated lower but are uncertain (V01, V03, V10) with CIs containing 0. The average posterior probabilities (shown in Figure 3) shows that V01 has the highest probability ($p = 0.83$), which is closely followed by V00 ($p = 0.78$). From V03 ($p = 0.55$) onwards, a relatively larger reduction in probability is observed. This means that there is an 83% and 78% probability that notice V01 and V00 receive a higher rating than any other randomly chosen notice from the pool of all other notices.

For irrelevant data, V05 is the reference notice, which is clearly rated higher than most other notices (V07, V09, V15, V16, V17, V18, V19), and it has a 95% CI entirely negative. However, V06 and V08 are rated lower but close enough, with 95% CIs containing 0, which suggests that although these could be slightly worse than V05, the upper bound for the CI still shows a small positive effect. The posterior probabilities (Figure 3) shows that V05 has the highest posterior probability ($p = 0.83$) followed by V08 ($p = 0.68$) and V06 ($p = 0.66$). This means that there is an 82.9% probability that notice V05 receives a higher rating than any other randomly chosen notice from the pool of all other notices.

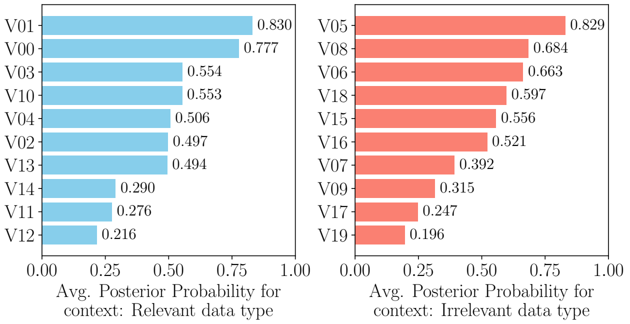


Figure 3: Average posterior probability of each notice being better than the others

In summary, for the relevant data type, notice V01 (non-blocking, before request with default voice) and V00 (blocking, before request with default voice) have the expectation of being rated as most preferred notices 83% and 78% of the time, respectively, whereas for the irrelevant data type, V05 (blocking, before request with default voice) has the expectation of being rated as the preferred notice 82.9% of the time.

6.2.2 Dissecting Factors. In this subsection, we address the second part of **RQ3** by examining the design factors. Each variant incorporates multiple design factors, so we analyzed the contribution of four key factors—*Control*, *Timing*, *Modality*, and *Voice*—to user ratings across both relevant and irrelevant verbal data request types. The mean ratings (shown in Table 5) show that participant preferences varied depending on the relevance of the data being requested. When the requested data was relevant to the skill and its flow, V01 (a non-blocking verbal warning in Alexa's voice before the request) received the highest mean rating of 4.03. Conversely, when the data was not relevant to the skill or its flow, V05 (a blocking verbal warning in Alexa's voice before the request) was rated highest, with a mean of 3.93.

The CLMM regression analysis reveals that each factor contributes differently to overall rating scores. As shown in Table 4, the *Voice* ($p < 0.001$) is the most important contributor, followed by *Timing* ($p < 0.01$), and both are statistically significant. Whereas *Control* ($p > 0.05$) and *Modality* ($p > 0.05$) are not significant. The `Estimate` sign indicates the preferred values for each factor (further elaborated in Table 8 in Appendix A). The estimate shows that for *Voice*, participants preferred the Default voice over a Different voice for both data types. For *Timing*, participants preferred the notice to be given before the information is requested rather than after for both data types. In terms of *Control*, we observe a slight overall preference for Blocking, but not significant, and for *Modality*, the preference is not as clear.

Orthogonality and Robustness Analysis. As the study followed a fractional factorial design, confounding among factors was possible, warranting robustness checks. To ensure the validity of our regression analysis, we conducted orthogonality checks, confounding factors analyses, and interaction tests. Orthogonality testing helps in verifying that the factors were independent, ensuring that each factor's effect could be estimated without being biased or inflated by correlations with others. Confounding analysis determines any overlap between predictors, as confounded variables can obscure the unique contribution of individual design factors and

Table 3: Bayesian Cumulative Logit Model Results for Relevant and Irrelevant Data. We used R call: `brm(formula = Rating ~ Notice + (1 | Participant), data = df, family = cumulative("logit"), chains = 4, iter = 2000, seed = 123)`

(a) Relevant Data

Parameter	Estimate	Est. Error	95% CI	$\hat{R}$	ESS
Random Effect (Participant)					
sd(Intercept)	1.19	0.18	[0.86, 1.56]	1.00	1943
Thresholds					
Intercept[1]	-3.59	0.38	[-4.36, -2.87]	1.00	1688
Intercept[2]	-2.07	0.35	[-2.78, -1.39]	1.00	1601
Intercept[3]	-0.92	0.34	[-1.59, -0.29]	1.00	1603
Intercept[4]	1.25	0.34	[0.59, 1.91]	1.00	1862
Regression Coefficients (Reference: V00)					
NoticeV01	0.57	0.45	[-0.32, 1.45]	1.00	2252
NoticeV02	-1.04	0.45	[-1.92, -0.18]	1.00	2367
NoticeV03	-0.72	0.43	[-1.55, 0.10]	1.00	2392
NoticeV04	-0.99	0.43	[-1.82, -0.15]	1.00	2240
NoticeV10	-0.75	0.44	[-1.59, 0.09]	1.00	2273
NoticeV11	-1.94	0.45	[-2.85, -1.08]	1.00	2155
NoticeV12	-2.26	0.46	[-3.19, -1.35]	1.00	2268
NoticeV13	-1.14	0.43	[-2.00, -0.32]	1.00	2303
NoticeV14	-1.81	0.43	[-2.65, -0.98]	1.00	2081

(b) Irrelevant Data

Parameter	Estimate	Est. Error	95% CI	$\hat{R}$	ESS
Random Effect (Participant)					
sd(Intercept)	0.95	0.16	[0.65, 1.28]	1.00	1171
Thresholds					
Intercept[1]	-3.95	0.39	[-4.75, -3.21]	1.00	1422
Intercept[2]	-2.12	0.34	[-2.82, -1.46]	1.00	1471
Intercept[3]	-1.16	0.33	[-1.81, -0.54]	1.00	1552
Intercept[4]	0.72	0.32	[0.08, 1.36]	1.00	1866
Regression Coefficients (Reference: V05)					
NoticeV06	-0.68	0.42	[-1.50, 0.13]	1.00	2058
NoticeV07	-1.82	0.43	[-2.67, -1.00]	1.00	2062
NoticeV08	-0.72	0.42	[-1.55, 0.09]	1.00	2182
NoticeV09	-2.04	0.44	[-2.96, -1.21]	1.00	1925
NoticeV15	-1.20	0.43	[-2.04, -0.39]	1.00	2132
NoticeV16	-1.36	0.43	[-2.23, -0.53]	1.00	1997
NoticeV17	-2.38	0.44	[-3.24, -1.51]	1.00	2098
NoticeV18	-0.97	0.42	[-1.79, -0.14]	1.00	2062
NoticeV19	-2.46	0.43	[-3.32, -1.61]	1.00	1779

lead to misleading inferences. In addition, we examined interaction effects to capture the potential influence of one factor (e.g., Timing) on another (e.g., Modality).

Orthogonality analysis results indicate that *Control* moderately confounds *Timing* ($r = .406$) and *Modality* ($r = 0.413$), while *Timing* slightly confounds *Modality* ($r = 0.173$); in contrast, *Voice* remains fully orthogonal ($|r| \leq 0.005$) and is also a statistically significant factor for both data types. Therefore, we can explain the effect of *Voice* individually, but not of other factors. We perform robustness checks by measuring the effects of factors in the absence of others and by measuring interaction effects. We found that for both relevant and irrelevant data, *Modality* does not have a significant effect ($p > 0.05$), but *Voice* ($p < 0.001$) and *Timing* ($p < 0.001$) still do, if the effect of *Control* is removed. However, *Control* shows a significant effect ($p < 0.05$) only when the effects of *Modality* and *Timing* are removed (see Tables 9 and 10 in Appendix A). This suggests that the effect of *Control* is not conclusive and appears to be confounded, but *Modality* still has a negligible effect. We also studied the interaction effects of *Timing*Modality* and found that the effect of *Voice* ($p < 0.001$) still outweighs the interaction effect ($p > 0.05$), hence this effect is reliable (see Table 11 in Appendix A). This is further explained by the confidence intervals shown in Table 4. We cannot test the interaction effects of *Control*Timing* and *Control*Modality* as they do not exist in our data (as blocking is not possible after a request is made and we do not force interaction to implement blocking through the screen). *Voice* and *Timing* do not change direction (preferred values) with interaction effects. *Timing* changes direction, but it has a clear overall preference or skew towards a negative value (representing *Timing=Before*). *Modality* does not show a clear preference.

In summary, *Voice* is a statistically significant factor and is orthogonal to other factors, indicating a clear preference for Default voice. Timing and Control both show a moderate overall preference for a blocking notice before data request, but are confounded, so preference is not as clear. The Modality estimate is small and unstable, indicating no global preference.

Participants' Rationale On Overall Preference. Users found *V01* clear and effective, appreciated the advance notification, and favored Alexa's voice. One participant noted: *"This is the best option of the four, since it informs the user prior to giving the information and the voice sounds much better since it's similar to the original Alexa" (P108).* For *V05* (the blocking warning for irrelevant data), participants liked the ability to opt-out, with one participant commenting:*"It's a clear, concise warning about what is going to happen and occurs while the user can still opt out" (P112).* Some participants found *V01* slightly cumbersome but still necessary in the given context, with one stating: *"It took longer, which is annoying, but it is best for a user to know what the app developer is seeing when we use it. A necessary evil." (P116)*

Participants' Rationale On Inconvenience. Participants generally preferred *V00* and *V05* because they did not require using a phone and used Alexa's voice. One participant shared: *"I really like this. It adds an extra layer of consent, giving the user autonomy over their decision. I also like that the voice matches the rest of the dialogue" (P136).* A few participants offered suggestions for improvement. One suggested that the notification should specify the exact information being requested: *"I would like it even more if Alexa specified what information was going to be sent to a third party, like directly telling me it would be my home address" (P132).*

Note that here a rating of 5 indicated the highest inconvenience, and a rating of 1 indicated the lowest inconvenience. Interestingly, *blocking verbal warnings in Alexa's voice before the data request* (variants *V00* and *V05* respectively) were rated best in terms of convenience, with mean ratings of 2.2 and 2.3 respectively (see Table 5).

6.2.3 Frequency of Notices. Most participants (63) preferred notices only the first time they used a skill, while 60 preferred whenever personal information was requested. 31 opted for occasional

Table 4: CLMM Results for Relevant and Irrelevant Data. ** $p < 0.01$, *** $p < 0.001$. **R call used:** `clmm(Rating ~ Control + Timing + Modality + Voice + (1 | Participant), data = df)`

	Relevant Data					Irrelevant Data				
Factor	**Estimate (β)**	**std err.**	**p-value**	**95% CI [β]**	**OR**	**Estimate (β)**	**std err.**	**p-value**	**95% CI [β]**	**OR**
Control	-0.417	0.284	0.142	[-0.973, 0.139]	0.659	-0.319	0.279	0.253	[-0.865, 0.227]	0.727
Timing	-0.670	0.219	**0.0022****	[-1.100, -0.240]	0.512	-1.218	0.222	**<0.001*****	[-1.652, -0.783]	0.296
Modality	0.010	0.217	0.962	[-0.416, 0.436]	1.010	0.009	0.214	0.965	[-0.411, 0.429]	1.009
Voice	-1.069	0.196	**<0.001*****	[-1.454, -0.685]	0.343	-0.605	0.192	**0.0016****	[-0.981, -0.230]	0.546
Model Summary:						**Model Summary:**				
Log-Likelihood		-582.51	AIC		1183.02	Log-Likelihood		-586.80	AIC	1191.60
N obs		400	BIC		1218.94	N obs		400	BIC	1227.52

Table 5: Mean of rating for all variants across relevant and irrelevant data types. The ↑ represents the higher value, the better, and ↓ represents the opposite. Standard deviations are provided in parentheses. We have an average standard deviation of 1.17 (relevant) and 1.2 (irrelevant), which is within the medium range for 5-point Likert-type scales [77] and is common in user opinion-based studies [71].

Data Type	Variant	Preference ↑	Inconvenience ↓
Relevant Data	V00	3.93 (± 0.69)	2.20 (±0.94)
	V01	4.03 (± 1.02)	2.35 (±1.21)
	V02	3.18 (±1.32)	2.90 (±1.29)
	V03	3.35 (±1.14)	2.83 (±1.26)
	V04	3.24 (±1.04)	2.73 (±1.1)
	V10	3.33 (±1.38)	3.03 (±1.31)
	V11	2.62 (±1.35)	2.88 (±1.32)
	V12	2.51 (±1.35)	3.15 (±1.33)
	V13	3.20 (±1.22)	3.10 (±1.18)
	V14	2.68 (±1.16)	3.08 (±1.27)
	Mean	3.21 (±1.17)	2.82 (±1.22)
Irrelevant Data	V05	3.93 (±1.0)	2.30 (±1.2)
	V06	3.55 (±1.1)	2.53 (±1.24)
	V07	2.93 (±1.24)	2.73 (±1.15)
	V08	3.78 (±1.0)	2.45 (±0.96)
	V09	2.78 (±1.35)	2.78 (±1.25)
	V15	3.30 (±1.18)	2.73 (±1.36)
	V16	3.25 (±1.15)	2.83 (±1.34)
	V17	2.60 (±1.42)	3.08 (±1.3)
	V18	3.38 (±1.27)	2.45 (±1.2)
	V19	2.48 (±1.22)	2.63 (±1.19)
	Mean	3.18 (±1.2)	2.65 (±1.22)

notifications. A few participants selected “Other”: two suggested notices only for a new type of information, and one deemed them unnecessary if the information was clearly relevant to the task.

7 Discussion

Study-I (interviews) was conducted to understand the voice assistant users’ perspective on personal data collection, its current privacy mechanisms and shortcomings as well as their suggestions to solve the identified problems (RQ1). Participants identified the need for verbal privacy notices with specific attributes (RQ2). Study-II evaluated various privacy notice prototypes that were designed based on the user-defined factors and contexts elicited from the results of Study-I, and identified the preferred notice designs (RQ3).

7.1 Privacy Notices in Voice vs. Visual Interfaces

In Study-I, our participants expressed a lack of knowledge about third-party voice apps even if they had used them (see Section 5.2.1), which may likely be a result of voice interface and seamless integration of voice apps—an aspect that did not exist in visual interface studies by prior works [79, 87]. Prior studies have also shown that many users incorrectly assume Amazon, rather than a third-party, is requesting access to their personal data [92]. This misconception fosters a false sense of security, potentially leading users to share sensitive data more readily, trusting Amazon’s brand over potentially less secure third parties. The results from Study-I also suggest that although the frameworks defined by prior works (Privacy Notices Design Space [79] and VR Security Indicators Design Space [87]) partially apply to the voice assistant domain (factors like *Control*, *Timing*, and *Modality*), voice-specific interfaces still notice design require further research covering unique voice interface-specific aspects (such as *Voice* and varying *Context*) that have not been studied before. In addition to adding voice interface-specific factors such as *Voice*, some factors from existing frameworks that were applied needed to be modified, such as combining visual and auditory modalities (from Privacy Notices Design Space [79]) and removing haptic and machine-readable feedback. Despite not being fully applicable directly, frameworks defined by prior works remarkably contributed to the design of our notices.

Study-II results show that the design of the voice-based notice makes a significant difference in users’ overall experience. Therefore, voice interfaces require significant intricacies in notice design compared to visual interfaces. In voice-only contexts, traditional visual privacy notices—such as consent forms, terms of service, or on-screen alerts—are unavailable. This makes verbal notifications essential for meaningful consent, as they can explicitly inform users about which entity is collecting their data and guide them toward making informed choices. Our proposed privacy notices are designed to function independently of screens, seamlessly operating on voice assistants with or without displays.

7.2 Regulatory Considerations

Smart voice assistants are subject to privacy regulations such as the GDPR and CPRA, which are designed to protect user privacy by enforcing requirements for informed consent and giving users control over their personal data [64, 72]. Participant’s lack of knowledge about third-party apps (from Study-I) mean that when these third-party voice apps collect user’s personal data, users will not have correct mental model and information about who is collecting or handling their data which is a violation of the article 13 of the GDPR and Section 3(A)(1) of CPRA that requires the data handling entity to provide the subject with its contact information [2, 3].

Article 13 of the GDPR explicitly states that the data controller should provide this information "at the time when personal data are obtained," which is enforced by our proposed run-time privacy notices [2]. Our participants of Study-I notably desired that the third parties should state *who* is requesting the data and *why* it is needed. Article 13 of the GDPR and Section 7(A)(3) of the CPRA also mandate data handlers to state the purposes of the processing for which the personal data are intended" [2, 4]. The CPRA's Section 3(A)(3) further allows consumers to request correction or deletion of previously collected data [3], which is also a desire expressed by our participants of Study-I to have transparency and control, including the ability to view and delete data shared with third parties.

As the law requires explicit consent for data collection, it is the platform's responsibility to ensure compliance when third-party voice apps collect user data, enforcing standardization. Our recommendation of providing automated verbal notice helps streamline and standardize this process without requiring individual skill developers to make additional modifications. Voice assistant manufacturers have a strong incentive to develop standardized privacy notification systems, as it can enhance consumer trust.

7.3 Recommendations

We make recommendations for voice assistant platforms as well as voice app developers based on the results from Study-I and Study-II. The Study-I highlighted the need for privacy notices that help users make informed decisions when personal data is requested, along with other privacy features that let users take control of their data on the platform. Study-II evaluated variants for privacy notices under the context of relevant and irrelevant data requests and determined notices that are most likely to be preferred.

Recommendations for Platforms. Based on our findings, we recommend that voice assistant platforms implement interactive verbal privacy notices before a third-party app requests personal information using the default voice. Platforms should implement privacy notices at the cloud level, automatically injecting appropriate notices whenever a voice app initiates verbal data requests. Implementing such notifications at the platform level (i.e., not leaving responsibility to individual developers) ensures fairness, standardization, and a consistent user experience. Study-I participants desired a data dashboard that would display personal data shared with third-parties—a commonly desired feature [32, 34]. Participants also wanted customization in notice settings, allowing adjustment of frequency and details of notices based on the sensitivity of the data. This approach would allow users to establish rules that apply across apps, minimizing unnecessary notifications—for instance, when verbal data requests come from trusted entities or involve less sensitive data. Our recommendations can adapt to a range of platforms, such as smart glasses (e.g., Meta Ray-Ban, SesameAI), in-car infotainment systems (e.g., Alexa-enabled vehicles), and other emerging voice-enabled technologies.

While our study offers insights into four design factors and their preferred combinations (constituting a notice), we observed some confounding and non-linear factors limiting us from measuring the individual effect of such factors. Moreover, a larger participant pool could have an even better statistical power to derive clearer factor preferences. Despite the challenges, our work elicits concrete findings showing the potential of improving user privacy through such efforts. Therefore, we suggest that platforms could run large-scale experiments (e.g., A/B testing) to further identify optimal notice designs involving more factors specific to the platform.

Recommendations for Developers. We suggest that skill developers offer privacy notices tailored to their app based on the design choices our work found to be preferred: not switching the voice to give privacy-related information and giving interactive notices beforehand. Notices can explain why a specific piece of data was being collected (e.g., *"To give you a personalized workout routine, I'll need to know your age."*). Additionally, data minimization is crucial for user privacy: developers should avoid requesting more personal data than is necessary to perform the intended task (see Section 5.2.3). Voice apps should provide users with mechanisms to access, delete, and control all data — shared explicitly or inferred through profiling features, such as Ad-ID [22].

Future Work. This work offers a preliminary investigation into notification design for voice interfaces. Future research could examine how preferences vary across voice application types and scenarios, and explore factors such as optimal notification frequency and framing. Unmoderated usability tests and A/B tests can be conducted to determine the ideal content, length, and frequency of the privacy notices catering for various demographics of users. While this study focuses on users, studies can be conducted from a developer's perspective on mechanisms to enhance users' privacy.

7.4 Limitations

Study-I relies on self-reported data from a limited, US-based English-speaking population, which may introduce social desirability bias and may not represent the diverse voice assistant community. Future studies may include a wider range of cultural backgrounds and underserved populations to address this. Study-II's survey sample was also exclusively from the US, which limits the generalizability of the findings to other regions with different privacy preferences and concerns. Additionally, while using observation-based evaluation via interaction videos was the most suitable method given privacy concerns, a more accurate assessment of usability would involve participants directly interacting with the skills. CLMM regression and orthogonality tests further revealed that not all design factors were independent: while the Voice factor was fully orthogonal, statistically significant, and distinguishable, Timing and Control showed aliasing due to fractional factorial design, which limited our ability to isolate the exact effects. Moreover, Study-II could have benefited from more participants for factor analysis, as it is underpowered; however, the number of participants was sufficient for the overall preference analysis via `brm` that showed convergence.

8 Conclusion

Our study revealed that participants preferred a verbal notice before personal information is verbally requested by a third-party voice app, especially when the requested information is not related to the task at hand. We also found that voice and timing are the most important factors impacting notice rating. Our work establishes a foundation for research into designing effective and usable privacy notices for voice interfaces.

Acknowledgments

We thank our anonymous reviewers and shepherd for their valuable feedback. We also like to thank Harshita Gupta, who helped with the initial design and making of the educational videos used in our user study. This research is partially supported by the National Science Foundation (NSF) under grant number CNS-2350075. The opinions, findings, conclusions, or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of the funding organization.

References

[1] 2015. https://press.aboutamazon.com/2015/6/amazon-introduces-the-alexa-skills-kit-a-free-sdk-for-developers

[2] 2016. Article 13 – Information to be provided where personal data are collected from the data subject. https://gdpr-info.eu/art-13-gdpr/. Regulation (EU) 2016/679 (General Data Protection Regulation).

[3] 2020. Section 3: Purpose and Intent. https://www.caprivacy.org/cpra-text/#section3. California Privacy Rights Act of 2020 (Prop. 24) — Text of the law (un-annotated) as published by the CPRA Resource Center.

[4] 2020. Section 7: Consumers' Right to Know What Personal Information Is Being Collected. Right to Access Personal Information. https://www.caprivacy.org/cpra-text/#section7. California Privacy Rights Act of 2020 (Prop. 24) — amendment to CCPA.

[5] 2020. The Smart Audio Report. https://www.nationalpublicmedia.com/uploads/2020/04/The-Smart-Audio-Report-from-NPR-and-Edison-Research.pdf

[6] 2021. 5 Ways Consumers Interact With Smart Speakers. https://mindstreammediagroup.com/introduction-smart-speakers-voice-search-brand-advertisers/

[7] 2021. Incredible Amazon Alexa Statistics You Need to Know in 2021. https://safeatlast.co/blog/amazon-alexa-statistics/

[8] 2022. BMW to build its next-generation voice experience on Alexa technology. https://www.aboutamazon.com/news/devices/amazon-bmw

[9] 2022. Ford's Alexa Built-In Rollout Continues. https://media.ford.com/content/fordmedia/fna/us/en/news/2022/04/07/ford-alexa-rollout-continues.html

[10] 2024. Create intents, utterances, and slots | Alexa skills kit. https://developer.amazon.com/en-US/docs/alexa/custom-skills/create-intents-utterances-and-slots.html

[11] 2024. Meet the People Bringing Your Digital Life into Your Car. https://www.gm.com/stories/alexa-team-spotlights

[12] 2024. The Smart Audio Report. https://www.nationalpublicmedia.com/uploads/2020/04/The-Smart-Audio-Report-from-NPR-and-Edison-Research.pdf

[13] 2024. Vehicles with Alexa. https://www.amazon.com/alexa-auto/b?ie=UTF8&node=17744356011

[14] 2025. https://www.prolific.com/

[15] 2025. Android Police. https://www.androidpolice.com/google-shutting-down-assistant-conversational-actions-app-actions-for-android/

[16] 2025. Free text to speech online with realistic AI voices. https://www.naturalreaders.com/online/

[17] 2025. Functional testing for a custom skill – Alexa skills kit. https://developer.amazon.com/en-US/docs/alexa/custom-skills/functional-testing-for-a-custom-skill.html

[18] 2025. What is the alexa skills kit? | Alexa skills kit. https://developer.amazon.com/en-US/docs/alexa/custom-skills/create-the-interaction-model-for-your-skill.html

[19] 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 15th Symposium on Usable Privacy and Security (SOUPS)*.

[20] Noura Abdi, Xiao Zhan, Kopo M. Ramokapane, and Jose Such. 2021. Privacy Norms for Smart Home Personal Assistants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI)*. Article 558, 14 pages.

[21] Mohammed Aldeen, Jeffrey Young, Song Liao, Tsu-Yao Chang, Long Cheng, Haipeng Cai, Xiapu Luo, and Hongxin Hu. 2024. End-Users Know Best: Identifying Undesired Behavior of Alexa Skills Through User Review Analysis. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 3, Article 89 (Sept. 2024), 28 pages. doi:10.1145/3678517

[22] Amazon Alexa Documentation. 2024?. About Alexa Advertising ID. https://developer.amazon.com/en-US/docs/alexa/advertising-id/overview.html. Last updated: approximately 1.2 years ago.

[23] Tawfiq Ammari, Jofish Kaye, Janice Y Tsai, and Frank Bentley. 2019. Music, search, and IoT: How people (really) use voice assistants. *ACM Transactions on Computer-Human Interaction (TOCHI)* 26, 3 (2019), 1–28.

[24] Reza Ghaiumy Anaraky, Tahereh Nabizadeh, Bart P. Knijnenburg, and Marten Risius. 2018. Reducing Default and Framing Effects in Privacy Decision-Making. In *SIGHCI 2018 Proceedings*. https://aisel.aisnet.org/sighci2018/20/ Paper 20.

[25] Chittaranjan Andrade. 2018. Internal, External, and Ecological Validity in Research Design, Conduct, and Evaluation. *Indian Journal of Psychological Medicine* 40, 5 (2018), 498–499. doi:10.4103/IJPSYM.IJPSYM_334_18

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

[27] Richard E Boyatzis. 1998. *Transforming qualitative information: Thematic analysis and code development.* sage.

[28] Stephen P. Brooks and Andrew Gelman. 1998. General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics* 7, 4 (1998), 434–455. doi:10.1080/10618600.1998.10474787

[29] Paul-Christian Bürkner and Matti Vuorre. 2019. Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science* 2, 1 (2019), 77–101.

[30] Stuart K. Card, Allen Newell, and Thomas P. Moran. 1983. *The Psychology of Human-Computer Interaction.* L. Erlbaum Associates Inc., USA.

[31] Janet X Chen, Allison McDonald, Yixin Zou, Emily Tseng, Kevin A Roundy, Acar Tamersoy, Florian Schaub, Thomas Ristenpart, and Nicola Dell. 2022. Trauma-informed computing: Towards safer technology experiences for all. In *Proceedings of the 2022 CHI conference on human factors in computing systems.* 1–20.

[32] Cheng Cheng and Kopo M. Ramokapane. 2025. "Erasing the Echo": The Usability of Data Deletion in Smart Personal Assistants. In *Proceedings on Privacy Enhancing Technologies (PoPETs)*, Vol. 2025. 76–93. doi:10.56553/popets-2025-0120

[33] Long Cheng, Christin Wilson, Song Liao, Jeffrey Young, Daniel Dong, and Hongxin Hu. 2020. Dangerous Skills Got Certified: Measuring the Trustworthiness of Skill Certification in Voice Personal Assistant Platforms *(CCS '20)*. Association for Computing Machinery, New York, NY, USA, 1699–1716. doi:10.1145/3372297.3423339

[34] Eugene Cho, S. Shyam Sundar, Saeed Abdullah, and Nasim Motalebi. 2020. Will Deleting History Make Alexa More Trustworthy? Effects of Privacy and Content Customization on User Experience of Smart Speakers *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376551

[35] Rune Haubo B. Christensen. 2022. *Cumulative Link Models for Ordinal Regression with the R Package ordinal.* Vignette / Technical Report. Technical University of Denmark & Christensen Statistics. https://cran.r-project.org/web/packages/ordinal/vignettes/clm_article.pdf

[36] Richard E Clark, David F Feldon, Jeroen JG Van Merrienboer, Kenneth A Yates, and Sean Early. 2008. *Cognitive task analysis.* Routledge. 577–593 pages.

[37] Benjamin R. Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "What can i help you with?": infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Vienna, Austria) *(MobileHCI '17)*. Association for Computing Machinery, New York, NY, USA, Article 43, 12 pages. doi:10.1145/3098279.3098539

[38] Lorrie Faith Cranor. 2012. Necessary but not sufficient: Standardized mechanisms for privacy notice and choice. *J. on Telecomm. & High Tech. L.* 10 (2012), 273.

[39] Jide S. Edu, Jose M. Such, and Guillermo Suarez-Tangil. 2021. Smart Home Personal Assistants: A Security and Privacy Review. *Comput. Surveys* 53, 6 (2021), 1–36. doi:10.1145/3412383

[40] Pardis Emami-Naeini, Janarth Dheenadhayalan, Yuvraj Agarwal, and Lorrie Faith Cranor. 2022. An Informative Security and Privacy "Nutrition" Label for Internet of Things Devices. *IEEE Security & Privacy* 20, 2 (2022), 31–39. doi:10.1109/MSEC.2021.3132398

[41] Peer Eyal, Rothschild David, Gordon Andrew, Evernden Zak, and Damer Ekaterina. 2021. Data quality of platforms and panels for online behavioral research. *Behavior research methods* (2021), 1–20.

[42] Patricia I Fusch Ph D and Lawrence R Ness. 2015. Are we there yet? Data saturation in qualitative research.

[43] Jonah Gabry, Ben Goodrich, Martin Lysy, and Andrew Johnson. 2024. *rstantools: Tools for Developing R Packages Interfacing with Stan.* Stan Development Team. https://www.rdocumentation.org/packages/rstantools/versions/2.4.0/topics/posterior_linpred Function 'posterior_linpred': Generic function for accessing the posterior distribution of the linear predictor.

[44] Christine Geeng, Mike Harris, Elissa Redmiles, and Franziska Roesner. 2022. " Like Lesbians Walking the Perimeter": Experiences of {US}.{LGBTQ+} Folks With Online Security, Safety, and Privacy Advice. In *31st USENIX Security Symposium (USENIX Security 22)*. 305–322.

[45] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian Data Analysis* (3rd ed.). Chapman & Hall/CRC, Boca Raton, FL. https://sites.stat.columbia.edu/gelman/book/BDA3.pdf Electronic version.

[46] Lea Gröber, Rafael Mrowczynski, Nimisha Vijay, Daphne A Muller, Adrian Dabrowski, and Katharina Krombholz. 2023. To Cloud or not to Cloud: A Qualitative Study on {Self-Hosters}'Motivation, Operation, and Security Mindset. In *32nd USENIX Security Symposium (USENIX Security 23)*. 2491–2508.

[47] Zhixiu Guo, Zijin Lin, Pan Li, and Kai Chen. 2020. Skillexplorer: Understanding the behavior of skills in large scale. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*. 2649–2666.

[48] Hana Habib and Lorrie Faith Cranor. 2022. Evaluating the usability of privacy choice mechanisms. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*. 273–289.

[49] Hana Habib, Yixin Zou, Yaxing Yao, Alessandro Acquisti, Lorrie Cranor, Joel Reidenberg, Norman Sadeh, and Florian Schaub. 2021. Toggles, dollar signs, and triangles: How to (in) effectively convey privacy choices with icons and link texts. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–25.

[50] Camille Harris, Amber Gayle Johnson, Sadie Palmer, Diyi Yang, and Amy Bruckman. 2023. " Honestly, I Think TikTok has a Vendetta Against Black Creators": Understanding Black Content Creator Experiences on TikTok. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2, 1–31.

[51] Sandra Höltervennhoff, Philip Klostermeyer, Noah Wöhler, Yasemin Acar, and Sascha Fahl. 2023. {"I} wouldn't want my unsafe code to run my {pacemaker"}: An Interview Study on the Use, Comprehension, and Perceived Risks of Unsafe Rust. In *32nd USENIX Security Symposium (USENIX Security 23)*. 2509–2525.

[52] Yue Huang, Borke Obada-Obieh, and Konstantin (Kosta) Beznosov. 2020. Amazon vs. My Brother: How Users of Shared Smart Speakers Perceive and Cope with Privacy Risks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376529

[53] Athina Ioannou, Iis Tussyadiah, Graham Miller, Shujun Li, and Mario Weick. 2021. Privacy nudges for disclosure of personal information: A systematic literature review and meta-analysis. *PLOS ONE* 16, 8 (Aug. 2021), e0256822. doi:10.1371/journal.pone.0256822

[54] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W. Reeder. 2009. A "nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security* (Mountain View, California, USA) *(SOUPS '09)*. Association for Computing Machinery, New York, NY, USA, Article 4, 12 pages. doi:10.1145/1572532.1572538

[55] Tina Khezresmaeilzadeh, Elaine Zhu, Kiersten Grieco, Daniel Dubois, Konstantinos Psounis, and David Choffnes. 2025. Echoes of Privacy: Uncovering the Profiling Practices of Voice Assistants. In *Proceedings on Privacy Enhancing Technologies (PETS)*, Vol. 2025. 71–87. doi:10.56553/popets-2025-0050

[56] Shari Kraber. 2022. What's Behind Aliasing in Fractional-Factorial Designs. Statease Blog. https://www.statease.com/blog/whats-behind-aliasing-in-fractional-factorial-designs/ Accessed: 2025-09-11.

[57] John K. Kruschke. 2021. Bayesian Analysis Reporting Guidelines (BARG). *Nature Human Behaviour* 5, 10 (2021), 1282–1291. doi:10.1038/s41562-021-01177-7

[58] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, are you listening? Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on human-computer interaction* 2, CSCW (2018), 1–31.

[59] 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. doi:10.1145/3274371

[60] Sophie-Charlotte Lemmer. 2020. *Alexa, Are You Friends With My Kid? Smart Speakers and Children's Privacy Under the GDPR*. Graduate Student Research Paper 2018/9_6. King's College London, Law School. 56 pages. doi:10.2139/ssrn.3627478

[61] Christopher Lentzsch, Sheel Jayesh Shah, Benjamin Andow, Martin Degeling, Anupam Das, and William Enck. 2021. Hey Alexa, is this skill safe?: Taking a closer look at the Alexa skill ecosystem. *Network and Distributed Systems Security (NDSS) Symposium2021* (2021).

[62] Jingjin Li, Chao Chen, Mostafa Rahimi Azghadi, Hossein Ghodosi, Lei Pan, and Jun Zhang. 2023. Security and privacy problems in voice assistant applications: A survey. *Computers & Security* 134 (2023), 103448.

[63] Bin Liu, Mads Schaarup Andersen, Florian Schaub, Hazim Almuhimedi, Shikun Aerin Zhang, Norman Sadeh, Yuvraj Agarwal, and Alessandro Acquisti. 2016. Follow my recommendations: A personalized privacy assistant for mobile app permissions. In *Twelfth symposium on usable privacy and security (SOUPS 2016)*. 27–41.

[64] Natasha Lomas. 2019. Google ordered to halt human review of Voice AI recordings over privacy risks. https://techcrunch.com/2019/08/02/google-ordered-to-halt-human-review-of-voice-ai-recordings-over-privacy-risks/

[65] Mohd Aliff Abdul Majid, Mohhidin Othman, Siti Fatimah Mohamad, Sarina Abdul Halim Lim, Aziz Yusof, et al. 2017. Piloting for interviews in qualitative research: Operationalization and lessons learnt. In *International Journal of Academic Research in Business and Social Sciences*, Vol. 7. 1073–1080.

[66] David Major, Danny Yuxing Huang, Marshini Chetty, and Nick Feamster. 2021. Alexa, Who Am I Speaking To?: Understanding Users' Ability to Identify Third-Party Apps on Amazon Alexa. 22, 1, Article 11 (Sept. 2021), 22 pages. doi:10.1145/3446389

[67] David Major, Danny Yuxing Huang, Marshini Chetty, and Nick Feamster. 2021. Alexa, who am I speaking to?: Understanding users' ability to identify third-party apps on Amazon Alexa. *ACM Transactions on Internet Technology (TOIT)* 22, 1 (2021), 1–22.

[68] Nathan Malkin, Joe Deatrick, Allen Tong, Primal Wijesekera, Serge Egelman, and David Wagner. 2019. Privacy Attitudes of Smart Speaker Users. In *Proceedings on Privacy Enhancing Technologies (PoPETs)*, Vol. 2019. 250–271. doi:10.2478/popets-2019-0068

[69] Hiroaki Masaki, Kengo Shibata, Shui Hoshino, Takahiro Ishihama, Nagayuki Saito, and Koji Yatani. 2020. Exploring Nudge Designs to Help Adolescent SNS Users Avoid Privacy and Safety Threats. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3313831.3376666

[70] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-Rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 72 (nov 2019), 23 pages. doi:10.1145/3359174

[71] Laura Naismith, Mike Sharples, and Jeffrey Ik Ting. 2005. Evaluation of CAERUS: A Context Aware Mobile Guide. https://www.researchgate.net/figure/Mean-and-Standard-Deviation-of-Responses-to-a-5-point-Likert-Scale_tbl1_32231646.

[72] Nicole Olsen. 2022. Voice assistants and privacy issues. https://www.privacypolicies.com/blog/voice-assistants-privacy-issues/

[73] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. https://cdn.openai.com/papers/whisper.pdf

[74] Juniper Research. 2020. Number of Voice Assistant Devices in Use to Overtake World Population by 2024, Reaching 8.4 bn, Led by Smartphones. Business Wire press release. https://www.businesswire.com/news/home/20200427005609/en/Juniper-Research-Number-of-Voice-Assistant-Devices-in-Use-to-Overtake-World-Population-by-2024-Reaching-8.4bn-Led-by-Smartphones.

[75] Jacqueline M. Roehl and Darci J. Harland. 2022. Imposter Participants: Overcoming Methodological Challenges Related to Balancing Participant Privacy with Data Quality When Using Online Recruitment and Data Collection. *The Qualitative Report* 27, 11 (2022), 2469–2485. doi:10.46743/2160-3715/2022.5475

[76] Aafaq Sabir, Evan Lafontaine, and Anupam Das. 2022. Hey alexa, who am I talking to?: analyzing users' perception and awareness regarding third-party alexa skills. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.

[77] Jeff Sauro. 2023. How to Estimate the Standard Deviation for Rating Scales. https://measuringu.com/rating-scale-standard-deviations/.

[78] Abhinaya SB, Abhishri Agrawal, Yaxing Yao, Yixin Zou, and Anupam Das. 2025. " What are they gonna do with my data?": Privacy Expectations, Concerns, and Behaviors in Virtual Reality. *Proceedings on Privacy Enhancing Technologies* 2025, 1 (2025), 58–77.

[79] Florian Schaub, Rebecca Balebako, Adam L Durity, and Lorrie Faith Cranor. 2015. A design space for effective privacy notices. In *Eleventh symposium on usable privacy and security (SOUPS 2015)*. 1–17.

[80] Eric Hal Schwartz. 2022. Alexa beats google assistant and Siri in voice assistant popularity as voice AI market expands. https://voicebot.ai/2022/06/24/alexa-beats-google-assistant-and-siri-in-voice-assistant-popularity-as-voice-ai-market-expands/

[81] Anselm Strauss and Juliet Corbin. 1990. *Basics of qualitative research*. Sage publications.

[82] Madiha Tabassum, Tomasz Kosinski, and Heather Richter Lipford. 2019. "I don't own the data": End User Perceptions of Smart Home Device Data Practices and Risks. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*. USENIX Association, Santa Clara, CA, 435–450. https://www.usenix.org/conference/soups2019/presentation/tabassum

[83] Mohammad Tahaei, Ruba Abu-Salma, and Awais Rashid. 2023. Stuck in the Permissions With You: Developer & End-User Perspectives on App Permissions & Their Privacy Ramifications. In *Proceedings of the 2023 ACM Conference on Human Factors in Computing Systems (CHI '23)*. ACM, 1–24. doi:10.1145/3544548.3581060

[84] Cayetano Valero, Jaime Pérez, Sonia Solera-Cotanilla, Mario Vega-Barbas, Guillermo Suarez-Tangil, Manuel Alvarez-Campana, and Gregorio López. 2023. Analysis of Security and Data Control in Smart Personal Assistants from the User's Perspective. *Future Generation Computer Systems* 144 (2023). doi:10.1016/j.future.2023.02.009

[85] Yang Wang, Pedro Giovanni Leon, Alessandro Acquisti, Lorrie Faith Cranor, Alain Forget, and Norman Sadeh. 2014. A field trial of privacy nudges for facebook. 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, 2367–2376. doi:10.1145/2556288.2557413

[86] Dominik Wermke, Noah Wöhler, Jan H Klemmer, Marcel Fourné, Yasemin Acar, and Sascha Fahl. 2022. Committed to trust: A qualitative study on security & trust in open source software projects. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1880–1896.

[87] Maximiliane Windl, Anna Scheidle, Ceenu George, and Sven Mayer. 2023. Investigating security indicators for hyperlinking within the metaverse. In *Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023)*.

[88] Rebecca Wong. 2021. Guidelines to Incorporate Trauma-Informed Care Strategies in Qualitative Research. Urban Institute.

[89] Fuman Xie, Chuan Yan, Mark Huasong Meng, Shaoming Teng, Yanjun Zhang, and Guangdong Bai. 2024. Are Your Requests Your True Needs? Checking Excessive Data Collection in VPA App. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering* (Lisbon, Portugal) *(ICSE '24)*. Association for Computing Machinery, New York, NY, USA, Article 205, 12 pages. doi:10.1145/3597503.3639107

[90] Fuman Xie, Yanjun Zhang, Chuan Yan, Suwan Li, Lei Bu, Kai Chen, Zi Huang, and Guangdong Bai. 2023. Scrutinizing Privacy Policy Compliance of Virtual Personal Assistant Apps *(ASE '22)*. Association for Computing Machinery, New York, NY, USA, Article 90, 13 pages. doi:10.1145/3551349.3560416

[91] Jeffrey Young, Song Liao, Long Cheng, Hongxin Hu, and Huixing Deng. 2022. SkillDetective: Automated Policy-Violation Detection of Voice Assistant Applications in the Wild. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association, Boston, MA. https://www.usenix.org/conference/usenixsecurity22/presentation/young

[92] Yangyong Zhang, Raj Vardhan, Phakpoom Chinprutthiwong, and Guofei Gu. 2023. Do Users Really Know Alexa? Understanding Alexa Skill Security Indicators. In *Proceedings of the 2023 ACM Asia Conference on Computer and Communications Security (ASIA CCS '23)*. 870–883.

[93] Serena Zheng, Noah Apthorpe, Marshini Chetty, and Nick Feamster. 2018. User Perceptions of Smart Home IoT Privacy. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 200 (Nov. 2018), 20 pages. doi:10.1145/3274469

[94] Noé Zufferey, Kavous Salehzadeh Niksirat, Mathias Humbert, and Kévin Huguenin. 2023. 'Revoked just now!' Users' Behaviors Toward Fitness-Data Sharing with Third-Party Applications. In *Proceedings on Privacy Enhancing Technologies (PoPETs)*, Vol. 2023. 47–67. doi:10.56553/popets-2023-0004

A Appendix

A.1 Ethical Considerations

We obtained approval for both of our studies from the Institutional Review Board (IRB) of our University. Specific ethical considerations for both studies are presented separately below.

A.1.1 *Study-I.*

For Study-I, no personally identifiable information was collected from participants; all audio recordings were transcribed and de-identified immediately after the interviews, and the recordings were deleted after transcription. Participants could skip any questions or quit the interview anytime without penalty [31, 88]. During the initial part of the interview, as a verification step, we asked the participants to turn on their webcam so that their faces and their VA devices were visible. We did not record this portion of the interview to respect participants' privacy. We obtained the interview transcripts using a local version of the Whisper model that did not upload any participant data and did not use the data for further training of the model. Participants were informed that their interviews would be transcribed.

A.1.2 *Study-II.*

In this study, no personally identifiable information was collected, and participants were contacted anonymously through the Prolific platform. Ideally, to enable participants to run skill variants on their own Alexa devices, we would need to publish all 20 variants of our skills to Alexa's skills store. This approach has ethical concerns for the participants as well as for the platform, as our skills request sensitive personal information (phone number and home address) as part of the experiment.

For participants and other Alexa users, there are risks involved, such as:

- Participants might inadvertently share their actual information, even though we do not actually store any data.
- Other users on the Alexa platform may access our skills who do not have knowledge of our study and share their personal information, as we do not have any control over who can install our skills once published.
- Participants can invoke incorrect skills during the experiment, which is possible as multiple skills share the same or similar invocation phrases, which may lead to confusion or missing data.

and **for the platform**, ethical concerns involve:

- Publishing multiple copies of experimental/mock skills on the live store. All skills need to pass Alexa's certification requiring platform's resources, and can be an abuse of the platform.
- Publishing skills that mock the functionality but do not actually perform the task. Publishing non-functional skills violates Alexa's skill policy [17].

Due to these ethical complications, we conducted our experiment using a safer option for participants and without adversely affecting the platform by using videos of the skill interaction. We deploy the skills in developer mode in our own accounts and make the videos.

A.2 Study Artifacts

Anonymized research artifacts are available online at https://osf.io/h7wkm/?view_only=baedc79e9f304172ae000f8392c30c9e. The videos used for Study-II are available in Table 12 in Appendix A.

A.3 Brief Interview Guide: Study-I

Following is the brief outline of the Study-I interview script. Complete version of the interview script can be found at https://osf.io/h7wkm/?view_only=baedc79e9f304172ae000f8392c30c9e.

1. **Introduction & Screening**
   - Consent to record audio
   - Confirm participant's Alexa skill usage.
   - If "no" but skills were listed in the screening survey: ask to recall; if there is a discrepancy, terminate interview without compensation.
   - If no skills in sign-up survey and none used, continue.
2. **Awareness of Skills & Personal Information**
   - Explain what "skills" are if needed (*EDU1*).
   - Explain how skills request "personal information" if needed.
   - Ask if any skills requested personal information.
     - If yes: details on what was requested, how provided (voice, permissions), and other types they think skills may request.
     - If no: ask perceptions of possible personal information requests; correct misconceptions (*EDU2*).
3. **Privacy Features in Alexa**
   - For those aware of personal information requests:
     - Ask if they know of Alexa privacy features related to skill requests.
     - For each feature: evaluate efficiency, likes, and shortcomings.
     - Present scenario with sensitive information request; assess if opinion changes.
   - If unaware: educate about privacy features (*EDU3*) before evaluation.
4. **Non-Skill Users**

   - Educate about skills (*EDU1*) and personal information collection (*EDU2*).
   - Educate about privacy features (*EDU3*) and follow same evaluation process as above.
5) **Suggestions**
   - Ask for recommended new features or changes to improve privacy for personal information requests.
   - Record any additional concerns.
6) **Closing**
   - Thank participant, share post-interview survey link.

A.4 Participant demographics

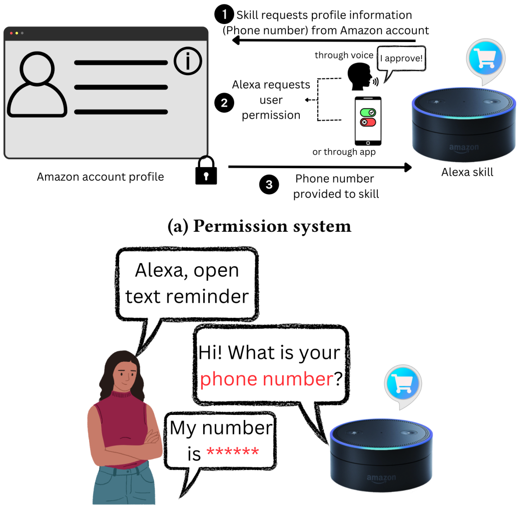
Table 6: Demographics of Study-I participants.

P#	Race	Age	Gender	Education
P0	African American	25-34	Woman	Bachelor's degree
P1	African American	18-24	Woman	High school or equiv.
P2	Caucasian	35-44	Man	Doctoral degree
P3	African American	25-34	Man	Bachelor's degree
P4	African American	25-34	Man	Bachelor's degree
P5	Caucasian	45-54	Man	Associate's degree
P6	Caucasian	18-24	Woman	Bachelor's degree
P7	Caucasian	65+	Man	Doctoral degree
P8	Asian	18-24	Man	Bachelor's degree
P9	Asian	18-24	Man	Bachelor's degree
P10	Asian	18-24	Man	Master's degree
P11	Hispanic + Asian	18-24	Man	Bachelor's degree
P12	Asian	25-34	Man	Bachelor's degree
P13	Prefer not to answer	18-24	Woman	Master's degree
P14	Asian	25-34	Woman	Bachelor's degree
P15	Asian	25-34	Woman	Master's degree
P16	Asian	25-34	Woman	Master's degree
P17	Latino	35-44	Woman	Some college
P18	African American	18-24	Man	High school or equiv.
P19	Latino	18-24	Woman	High school or equiv.
P20	Caucasian	55-64	Man	Bachelor's degree

Table 7: Demographics of Study-II participants.

Demographic	Value	Count	Percentage
Age Group	18-24	27	17%
	25-34	60	38%
	35-44	45	28%
	45-54	17	10%
	55-64	11	6%
Gender	Man	67	42%
	Woman	90	56%
	Non-binary	2	1%
	Prefer not to answer	1	1%
Education	High school or equivalent	18	11%
	Some college	22	14%
	Trade, technical or vocational training	5	3%
	Associate's degree	17	11%
	Bachelor's degree	66	41%
	Master's degree	28	18%
	Doctorate degree	1	1%
Race	Caucasian	112	70%
	Black or African American	24	15%
	Hispanic, Latinx, or Spanish origin	20	13%
	Asian	18	11%
	American Indian or Alaskan Native	2	1%
	Middle Eastern or North African	2	1%
	Prefer not to answer	1	1%

A.5 Data request methods


1 Skill requests profile information (Phone number) from Amazon account
through voice
I approve!
2 Alexa requests user permission
or through app
Amazon account profile
3 Phone number provided to skill
Alexa skill
Alexa, open text reminder
Hi! What is your phone number?
My number is ******

(a) Permission system

(b) Verbal data request

Figure 4: Comparing data request through Alexa account vs. requesting users verbally.

A.6 CLMM Models for Robustness Checks

Table 8: Interpretation of the Estimate sign in CLMM regression

Factor	Value	Code	Estimate (β) interpretation
Control	Blocking Non-blocking	0 1	Estimate will be negative if Blocking is preferred (increases ratings) and will be positive if Non-blocking is preferred.
Timing	Before request After request	0 1	Estimate will be negative if Before request is preferred and will be positive if After request is preferred.
Modality	Audio Audio+App	0 1	Estimate will be negative if Audio only is preferred and will be positive if Audio+App is preferred.
Voice	Default Different	0 1	Estimate will be negative if Default voice is preferred and will be positive if Different voice is preferred.

Table 9: CLMM Results for Relevant Data under Different Model Specifications

(a) Without Control Factor
`(Rating ∼ Timing + Modality + Voice + (1 | Participant))`

Factor	Estimate (β)	Std. Error	z value	p-value
Timing	-0.7771	0.2072	-3.751	0.000176***
Modality	-0.1141	0.2001	-0.570	0.5686
Voice	-1.0711	0.1962	-5.459	4.8e−08***

Random Effects:
Participant (Intercept): Variance = 1.191, Std. Dev. = 1.091

$^{***}p < 0.001$

(b) Without Modality and Timing
`(Rating ∼ Voice + Control + (1 | Participant))`

Factor	Estimate (β)	Std. Error	z value	p-value
Voice	-1.0428	0.1953	-5.338	9.38e−08***
Control	-0.7149	0.2414	-2.961	0.00306**

Random Effects:
Participant (Intercept): Variance = 1.160, Std. Dev. = 1.077

$^{**}p < 0.01$, $^{***}p < 0.001$

Table 10: CLMM Results for Irrelevant Data under Different Model Specifications

(a) Without Control Factor
`(Rating ∼ Timing + Modality + Voice + (1 | Participant))`

Factor	Estimate (β)	Std. Error	z value	p-value
Timing	-1.3102	0.2069	-6.333	2.41e−10***
Modality	-0.0827	0.1985	-0.417	0.6770
Voice	-0.6017	0.1916	-3.140	0.00169**

Random Effects:
Participant (Intercept): Variance = 0.730, Std. Dev. = 0.854

$^{**}p < 0.01$, $^{***}p < 0.001$

(b) Without Modality and Timing
`(Rating ∼ Voice + Control + (1 | Participant))`

Factor	Estimate (β)	Std. Error	z value	p-value
Voice	-0.5968	0.1897	-3.145	0.001658**
Control	-0.8843	0.2359	-3.749	0.000178***

Random Effects:
Participant (Intercept): Variance = 0.637, Std. Dev. = 0.798

$^{**}p < 0.01$, $^{***}p < 0.001$

Table 11: CLMM Results for Relevant and Irrelevant Data. Model: `Rating ∼ Timing * Modality + Control + Voice + (1 | Participant)`

(a) Relevant Data

Factor	Estimate (β)	Std. Error	z value	p-value
Timing	-0.9146	0.3199	-2.859	0.00426**
Modality	-0.2137	0.3044	-0.702	0.48250
Control	-0.2969	0.3058	-0.971	0.33156
Voice	-1.0729	0.1964	-5.462	4.7e−08***
Timing * Modality	0.4541	0.4314	1.052	0.29258

Random Effects:
Participant (Intercept): Variance = 1.211, Std. Dev. = 1.101

$^{**}p < 0.01$, $^{***}p < 0.001$

(b) Irrelevant Data

Factor	Estimate (β)	Std. Error	z value	p-value
Timing	-1.0649	0.3060	-3.480	0.00050***
Modality	0.1558	0.2953	0.528	0.59772
Control	-0.3935	0.2972	-1.324	0.18555
Voice	-0.6021	0.1918	-3.139	0.00170**
Timing * Modality	-0.3062	0.4250	-0.721	0.47114

Random Effects:
Participant (Intercept): Variance = 0.7408, Std. Dev. = 0.8607

$^{**}p < 0.01$, $^{***}p < 0.001$

A.7 Links to All Educational Videos

Table 12: The table contains the urls for all the Educational videos used in study phase 1 and also for the video interactions for all the skill with privacy notice variants

Study-I: Educational Videos	
Video description	**URL**
EDU Video 1: Skills	https://www.youtube.com/watch?v=K6GdgwHA0yc
EDU Video 2: Personal information collection	https://www.youtube.com/watch?v=xCcNnTSExpc
EDU Video 3: Skill permissions	https://www.youtube.com/watch?v=oCuiG2r5OZg

Study-II: Skill intervention variants		
Video description	**URL**	**Intervention text**
Base variant for relevant data (Base-V0X)	https://www.youtube.com/watch?v=SNi1VHtrPtE	None.
Base variant for irrelevant data (Base-V1X)	https://www.youtube.com/watch?v=O7I1oHI1LQs	None.
V00 (Relevant data, Blocking, Before request, Audio, Default-Voice)	https://www.youtube.com/watch?v=212alm_q1k4	Alexa wants to inform you that the task to do skill is going to request some personal information which will be sent to the third-party developers of this skill... Would you like to continue?
V01 (Relevant data, Non-blocking, Before request, Audio, Default-Voice)	https://www.youtube.com/watch?v=WEsUYlCUywU	Alexa wants to inform you that the task to do skill is going to request some personal information which will be sent to the third-party developers of this skill... The skill will continue now.
V02 (Relevant data, Non-blocking, After response, Audio, Default-Voice)	https://www.youtube.com/watch?v=-ObA6dvFggI	Alexa wants to inform you that the personal information that you provided is sent to the third-party developers of the task to do skill... The skill will continue now.
V03 (Relevant data, Non-blocking, Before request, App, Default-Voice)	https://www.youtube.com/watch?v=P_c6OBexL_M	Same as V02 with popup on Alexa app.
V04 (Relevant data, Non-blocking, After response, App, Default-Voice)	https://www.youtube.com/watch?v=DO9I0yR8QS0	Same as V03 with popup on Alexa app.
V05 (Irrelevant data, Blocking, Before request, Audio, Default-Voice)	https://www.youtube.com/watch?v=IP7lyylS8RY	Same as V00 for irrelevant data request.
V06 (Irrelevant data, Non-blocking, Before request, Audio, Default-Voice)	https://www.youtube.com/watch?v=hR_Q3Dhm04k	Same as V01 for irrelevant data request.
V07 (Irrelevant data, Non-blocking, After response, Audio, Default-Voice)	https://www.youtube.com/watch?v=THpxVycT108	Same as V02 for irrelevant data request.
V08 (Irrelevant data, Non-blocking, Before request, App, Default-Voice)	https://www.youtube.com/watch?v=uRtwNt090Go	Same as V03 for irrelevant data request.
V09 (Irrelevant data, Non-blocking, After response, App, Default-Voice)	https://www.youtube.com/watch?v=9VtCWP_mF9k	Same as V04 for irrelevant data request.
V10 (Relevant data, Blocking, Before request, Audio, Different-Voice)	https://www.youtube.com/watch?v=gtSFVBgffDw	Same as V00 but with different voice.
V11 (Relevant data, Non-blocking, Before request, Audio, Different-Voice)	https://www.youtube.com/watch?v=jv3qqeGujjA	Same as V01 but with different voice.
V12 (Relevant data, Non-blocking, After response, Audio, Different-Voice)	https://www.youtube.com/watch?v=NFVWUj7jDpo	Same as V02 but with different voice.
V13 (Relevant data, Non-blocking, Before request, App, Different-Voice)	https://www.youtube.com/watch?v=nF0Vy7ZcuT8	Same as V03 but with different voice.
V14 (Relevant data, Non-blocking, After response, App, Different-Voice)	https://www.youtube.com/watch?v=_sOWPJsB-QI	Same as V04 but with different voice.
V15 (Irrelevant data, Blocking, Before request, Audio, Different-Voice)	https://www.youtube.com/watch?v=9n9Std1sOlI	Same as V05 but with different voice.
V16 (Irrelevant data, Non-blocking, Before request, Audio, Different-Voice)	https://www.youtube.com/watch?v=cZgF8XIFyYI	Same as V06 but with different voice
V17 (Irrelevant data, Non-blocking, After response, Audio, Different-Voice)	https://www.youtube.com/watch?v=UxFMlI2jzto	Same as V07 but with different voice
V18 (Irrelevant data, Non-blocking, Before request, App, Different-Voice)	https://www.youtube.com/watch?v=9ewdDNri3SM	Same as V08 but with different voice.
V19 (Irrelevant data, Non-blocking, After response, App, Different-Voice)	https://www.youtube.com/watch?v=uXU81MLQ2G8	Same as V09 but with different voice.