Hey Alexa, Who Am I Talking to?: Analyzing Users' Perception and Awareness Regarding Third-party Alexa Skills

Aafaq Sabir
asabir2@ncsu.edu
North Carolina State University
Raleigh, North Carolina, USA

Evan Lafontaine
elafont@ncsu.edu
North Carolina State University
Raleigh, North Carolina, USA

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

ABSTRACT

The Amazon Alexa voice assistant provides convenience through automation and control of smart home appliances using voice commands. Amazon allows third-party applications known as skills to run on top of Alexa to further extend Alexa's capability. However, as multiple skills can share the same invocation phrase and request access to sensitive user data, growing security and privacy concerns surround third-party skills. In this paper, we study the availability and effectiveness of existing security indicators or a lack thereof to help users properly comprehend the risk of interacting with different types of skills. We conduct an interactive user study (inviting active users of Amazon Alexa) where participants listen to and interact with real-world skills using the official Alexa app. We find that most participants fail to identify the skill developer correctly (i.e., they assume Amazon also develops the third-party skills) and cannot correctly determine which skills will be automatically activated through the voice interface. We also propose and evaluate a few voice-based skill type indicators, showcasing how users would benefit from such voice-based indicators.

CCS CONCEPTS

• **Security and privacy** → *Human and societal aspects of security and privacy.*

KEYWORDS

Voice assistant; Third-party skills; Security indicators

ACM Reference Format:
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 *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3491102.3517510

1 INTRODUCTION

Voice-based user experience thrives on the ability that enables users to interact with devices and services through voice instead of keystrokes, clicks, or swipes. As a result, we have seen a rapid growth of voice-based services such as Amazon Alexa and Google Home. Amazon Alexa is the market leader in this space as it is compatible with over 7,400 smart home devices [3] — an important feature that attracts third-party developers to develop applications (a.k.a. *Skills* for the Alexa platform). These skills allow end-users to interact with numerous online services and smart home devices through Alexa-enabled devices, such as Amazon Echo or Echo Show. Amazon's Alexa ecosystem currently has over 100,000 skills, most of which are from third parties [22]. For example, the 'Restaurant Finder' skill [25] enables a user to search for restaurants based on a zip code. However, third-party skills can also create privacy concerns as developers can access sensitive user information. Amazon has various policies in place to prevent and limit unauthorized access to sensitive information while also vetting skills for inappropriate content before they are publicly available on the skill store; yet, recent studies have found gaps in the vetting process and have identified skills that are potentially non-compliant with various policies set forth by Amazon [38, 39, 52, 63]. Others have performed dynamic analysis of skill interaction to identify policy-violating skills [43, 62]. Researchers have also analyzed skills' privacy policies to identify those that do not fully disclose all types of sensitive data accessed [52, 53, 63]. Furthermore, researchers have started studying the privacy perceptions, concerns, and privacy-seeking behaviors [31, 32, 35, 50] associated with using voice assistants.

However, little research has evaluated the effectiveness of user-facing security indicators or interventions currently available on voice-based platforms like Alexa. Such security indicators include helping consumers distinguish which skills are third-party skills and inform what type of data a given skill requests (e.g., through the permission model). There has been a similar trend of research in the domain of web browsers [33, 41] and mobile apps [36, 54] that propose various visual cues (e.g., icons, color codes, or highlighted text) to inform/warn users of different security and privacy risks. However, the voice interface presents a unique set of challenges as there is typically no visual interface to display visual cues other than pushing information to the companion app, which would obstruct the seamless experience and under-utilize the convenience of a voice interface.

In this paper, we analyze users' perception of how skills are selected by Alexa and whether users can distinguish third-party skills from native skills. As any skill that matches the invocation phrase is *automatically* enabled (unless the skill requires special permissions) on Alexa, users can unknowingly enable the wrong skill and be exposed to malicious skills collecting sensitive data. Thus, studying the efficacy of existing security and privacy indicators or lack thereof is critical for the ecosystem to sustain. We conduct an interactive user study to assess users' knowledge about using skills and explore what controls can be implemented to enhance

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3517510

transparency in the Alexa skill ecosystem. To accomplish this, we study the following research questions: **RQ1:** ***Are users able to distinguish between third-party and native skills?*** We ask participants to visit the skill information page and also provide them with audio samples from real-world skills to determine whether participants can accurately distinguish skills that are native versus third-party. **RQ2:** ***With Alexa's auto-enable feature, can users predict which skill will be activated when invoked?*** We ask participants to use the Alexa app to search for skills and identify which skill they think will be activated when verbally invoked. Upon actual invocation, we can determine the extent to which participants can accurately predict which skills will be auto-enabled. **RQ3:** ***How can Alexa better inform its users about the ownership of the skills they interact with?*** We present participants with different audio-based templates to better convey information about the type of skill they are interacting with and ask them to rank them in terms of their preference. We also analyze open text responses to identify other forms of usable voice-based security indicators. In summary, we make the following contributions:

- We design an interactive user study that involves both listening to audio excerpts from real-world skills as well as invoking skills on the Alexa app. Furthermore, we recruit participants from different online forums and groups specializing in Amazon Alexa.
- Our analysis reveals that users cannot effectively distinguish third-party skills from native ones through the voice interface. Furthermore, we show that users cannot identify which skill will be activated when multiple similar skills exist.
- We prototype and evaluate a few voice-based interventions to help users better distinguish third-party skills from native skills.

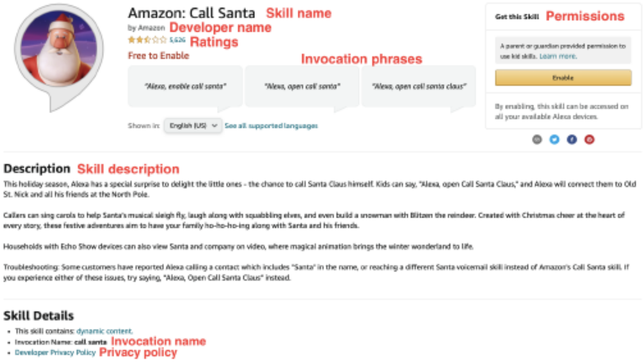
The remainder of the paper is structured as following. Section 2 provides background information on the Alexa skills ecosystem. Section 3 discusses related work in this field. Section 4 summarizes our study design and analysis approach. Sections 5, 6, and 7 provide the analysis of the user study data and answer to research questions 1, 2, and 3, respectively. Section 8 discusses our findings and section 9 describes limitations of our study. Section 10 concludes our research.

2 BACKGROUND

2.1 Publishing a Skill

There are currently over 100,000 skills in the Alexa skill store across different countries [22]. A skill essentially adds new functionality to Alexa that is not supported by native operations. A skill consists of many different meta-data, such as a skill name, invocation phrase, and invocation name (as shown in Figure 1). None of this information is required to be unique, causing many skills to have duplicate names and phrases. Invocation phrases are also used in order to actually enable and activate the skill on the Alexa-enabled device.

The majority of these skills are developed by third parties (referred to as "third-party skills") rather than by Amazon (referred to as "native skills"). In order to be published on the Amazon store, third-party skills need to go through various verification and approval measures. For instance, Amazon requires that the skill meets the Alexa policy guidelines, provides security requirements for hosting the skill, and passes functional tests [13]. If these requirements, among others, are met, then the skill may be published on the Amazon store for users to view and enable on their Alexa devices. Once a skill is approved, it is listed on the Amazon skill store. Users can search for skills and view their information page, which includes all information relevant to the skill (e.g., invocation phrase, skill name, and developer name). See Figure 1 for an example of a native skill on the Amazon store (meta-information is labeled in red).


Amazon: Call Santa

by Amazon

★★½☆☆ 5,626

Free to Enable

"Alexa, enable call santa"

"Alexa, open call santa"

"Alexa, open call santa claus"

Shown in: English (US) See all supported languages

Get this Skill

A parent or guardian provided permission to use kid skills. Learn more.

Enable

By enabling, this skill can be accessed on all your available Alexa devices.

Description

This holiday season, Alexa has a special surprise to delight the little ones - the chance to call Santa Claus himself. Kids can say, "Alexa, open Call Santa Claus," and Alexa will connect them to Old St. Nick and all his friends at the North Pole.

Callers can sing carols to help Santa's musical sleigh fly, laugh along with squabbling elves, and even build a snowman with Blitzen the reindeer. Created with Christmas cheer at the heart of every story, these festive adventures aim to have your family ho-ho-ho-ing along with Santa and his friends.

Households with Echo Show devices can also view Santa and company on video, where magical animation brings the winter wonderland to life.

Troubleshooting: Some customers have reported Alexa calling a contact which includes "Santa" in the name, or reaching a different Santa voicemail skill instead of Amazon's Call Santa skill. If you experience either of these issues, try saying, "Alexa, Open Call Santa Claus" instead.

Skill Details

- This skill contains: dynamic content.
- Invocation Name: call santa
- Developer Privacy Policy

Figure 1: The skill "Amazon: Call Santa" on the Amazon skill store (labeled with red text).

2.2 Enabling a Skill

When a user utters the invocation phrase or utters a sentence containing the invocation phrase, Alexa automatically opens the skill, whether native or third-party. Alexa will prioritize native skills over third-party skills. Alexa first checks whether the skill corresponding to the invocation phrase is already enabled and, if so, reaches out to the backend and responds with the output. However, if the skill is not already enabled, Alexa verifies whether there is a native skill available and, if so, utilizes that skill by default. If not, and if a third-party matching skill is found, it is automatically enabled and contacts the third-party backend for a response. Figure 2 highlights the overall skill selection process. A major issue with this workflow is whether a third-party skill is *automatically* enabled and used without user awareness. As skill developers can alter their backend code (responsible for maintaining the dialogue with the user) without requiring any new approval, there is the possibility to coax users into sharing more sensitive information than required for the functionality of the skill; this poses additional security and privacy risks [52].

3 RELATED WORK

3.1 Skill-based Attacks

As third-party skills are becoming increasingly popular, several studies have aimed to find vulnerabilities and exploits in Amazon Alexa skills. In a comprehensive literature review, Jide et al. [40] classified security and privacy vulnerabilities in Amazon Alexa skills into weak authentication, weak authorization, profiling, and adversarial AI. They discussed many studies that demonstrate skill-based attacks on smart voice assistants [34, 51, 65] and present countermeasures to some of the known skill-based attacks [37, 42,

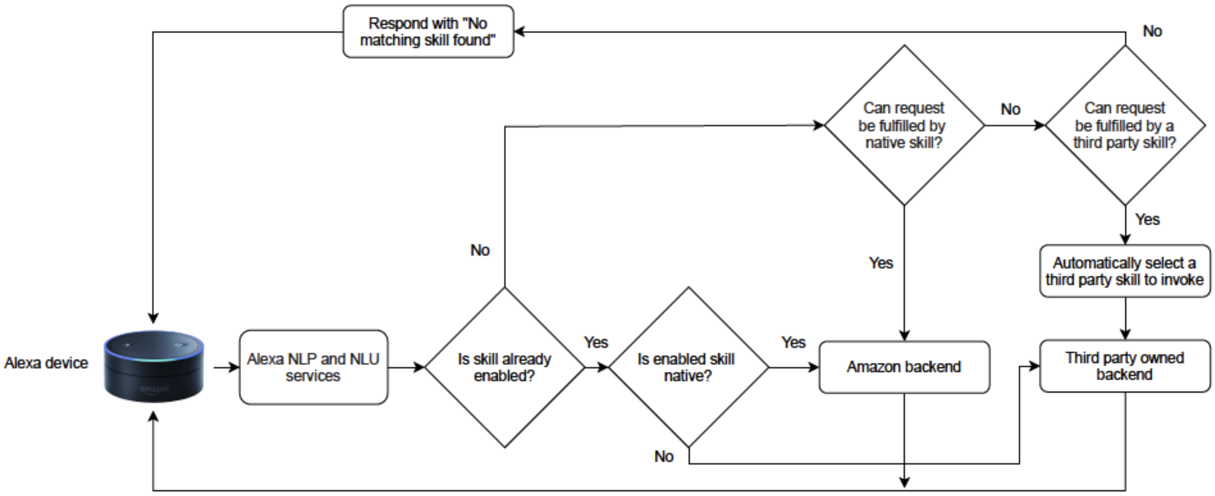

Respond with "No matching skill found"
No
Can request be fulfilled by native skill?
No
Can request be fulfilled by a third party skill?
Yes
No
Yes
Automatically select a third party skill to invoke
Alexa device
Alexa NLP and NLU services
Is skill already enabled?
Yes
Is enabled skill native?
Yes
Amazon backend
Third party owned backend
No

Figure 2: Alexa's skill invocation and enabling process.

49]. Malicious skills may also collude to aggregate personal data from multiple skills, as it is seen in smartphone apps [57], which allows for more powerful inferences based on user profiling from personal data.

Furthermore, skills can collect personal data, such as date of birth, age, or blood type, during user interaction without requiring any explicit permissions [59]. Third-party skills may also "request and collect personal data including user passwords" and "eavesdrop on users after they believe the smart speaker has stopped listening" [2]. These capabilities can enable malicious skills to be very intrusive, especially when a user mistakenly considers the skill to be native and fully trusts it.

With the ability to publish skills with similar or identical invocation phrases, multiple studies have looked at the feasibility of launching speech-based attacks. These attacks abuse existing weaknesses in the speech recognition system. For instance, prior studies demonstrate that it is possible to carry out voice-squatting attacks by leveraging the similarity between the invocation name of two different skills [48]. For example, two skills with the invocation names "Test Your Luck" and "Test Your Lock" will create confusion. Zhang et al. [65] presented a new variation of skill-squatting attacks where a paraphrased invocation name can hijack legitimate skills (e.g., the invocation "capital one please" would be able to hijack the voice command meant for the legitimate skill named "capital one"). Additionally, Zhang et al. [66] have utilized a fuzzing tool to systematically discover misinterpretation-prone voice commands that can activate the wrong skill. Lastly, Mitev et al. [58] demonstrate a security vulnerability in Alexa's voice interface that allows malicious attackers to redirect the user's voice towards a malicious skill, carrying out a form of the man-in-the-middle attack. The feasibility of such attacks motivates the need to make users are aware of which skills are activated when they interact with voice assistants.

3.2 Skill Vetting

An Alexa skill needs to pass a vetting process before it is published on the skill store; this ensures that it complies with Amazon's policies and security requirements [8]. However, researchers show that skills can bypass the skill vetting process and can pose security and privacy risks to users [38, 45, 52, 64]. Lentzsch et al. [52] performed a large-scale analysis of 90,194 unique skills in the skill store and found that not only can a malicious user publish a skill under any arbitrary developer/company name, but they can also make backend code changes after approval to coax users into revealing unwanted information. Cheng et al. [38] and Hu et al. [44] have manually vetted skills to identify policy-violating skills in the current store under different sensitive categories (e.g., kids' category). Furthermore, Guo et al. [43] have built an automated system named "SkillExplorer" to interact with skills to analyze their dynamic behavior. They analyzed 28,904 skills from the Amazon market and found that 1,141 skills requested users' private information without following developer specifications. They also discovered 68 skills that eavesdrop on users' private conversations, even after sending the 'stop' command. Similarly, Shezan et al. [62] built an automated system to dynamically analyze health-related skills. Researchers have also built tools to automatically analyze the privacy policies of skills to identify skills that do not fully disclose different types of sensitive data accessed [52, 53, 63]. These works demonstrate that malicious skills can bypass the vetting process and that users need to be more vigilant about what skills are actually enabled on their devices through the voice interface.

3.3 Analyzing Users' Perceptions

Researchers have also started looking at privacy perceptions, concerns, and even privacy-seeking behaviors associated with using voice assistants. Ammari et al. [35] studied how voice assistants are used by users and identified the most common use cases – streaming music, searching, and controlling smart home appliances. Others have highlighted similar usage patterns [55, 60]. Koshy et al. [47] compared the awareness and knowledge of personal assistants between pilot users (those who configure the device) and passenger users (those who use the already-configured device); they found that passenger users lack understanding about the functionality of virtual assistants compared to pilot users. They also found that the passenger/pilot relationship is very common in households, and passenger users primarily obtain information about the devices from pilot users. This might make them even more vulnerable to skill-based attacks. Huang et al. [46] found that participants were worried about their data, such as contact lists, being accessible by other household members and visitors. They found that participants were concerned by data collection by smart-speaker manufacturers and did not trust the smart speaker manufacturers.

Lau et al. [50] analyzed the privacy-seeking behaviors that users adopt when they interact with voice assistants and show that current privacy controls (e.g., physically pressing the mute button) are rarely used and do not align well with users' needs. Abdi et al. [31, 32] highlighted the privacy perceptions and norms surrounding the use of smart home personal assistants. They elicited participants' mental model by asking questions about how participants think the voice assistant works. Then, participants were asked specific questions regarding how and where they think the data is stored, with whom it is shared, and if they think the device learns about them. Abdi et al. found that participants were very concerned about the security and privacy of their smart home devices and believed that they were vulnerable to "hackers." However, none of the participants mentioned any threat based on third-party skills, which suggests a gap in awareness for participants' knowledge about skill-based attacks.

Given that malicious skills can bypass the vetting process, it is critical that platforms provide sufficient and usable indicators or interventions to help users make informed decisions about enabling third-party skills. Little research has been performed in this context to understand users' mental models and perceptions about how Alexa's skill ecosystem operates. Major et al. [56] compared Alexa users and non-users and found that even Alexa owners are more likely to presume skills are developed by Amazon. They performed the first user study to determine how well participants can distinguish third-party skills from native skills by showing participants videos of a simple interaction with different skills. They found that participants do not understand that third parties often operate skills and often confuse third-party skills with native Alexa functions.

3.4 Distinction from Prior Works

While existing studies have uncovered discrepancies in Alexa's skill vetting process, demonstrated attacks on voice assistants, and attempted to understand users' perception about how voice assistants handle user data, there is a lack in understanding which security and privacy indicators help users to better identify the skills they intend to interact with. There is also a lack of evaluation about how visual indicators present in a skill's information page compare to those present (or absent) in the voice interface.

As previously mentioned, Major et al. [56] first attempted to uncover the level of awareness users have regarding third-party skills and native functionalities; however, their study is based on participants feedback on videos and audios of skill interactions, and it did not involve interactions with the real-world voice assistant app to capture personal user experience. They also did not evaluate custom variations in skill interactions to evaluate feasible privacy indicators. We also believe existing works do not evaluate visual security indicators and how they are absent within the voice interface. In this paper, we analyze a broader set of research questions through a *interactive* user study that involves activating and interacting with real-world skills using the official Alexa app. We focus not only on the awareness to distinguish third-party skills from native skills but also elicit and test potential voice-based indicators/interventions to help users better distinguish different types of skills. For instance, in our user study, we ask participants about what skill they think will be activated once invoked by using the search and ranking function provided by Alexa. Next, we ask them to actually utter the invocation phrase and cross-check the actual activated skill with their initial predictions. We also attempt to understand participants' mental models by asking them about their level of confidence in their decision. We further elicit open-ended responses to obtain deeper insight into their thought process. Overall, we believe that we make novel contributions in understanding users' mental models regarding Alexa's skill selection process and the efficacy or lack thereof of existing security and privacy indicators.

4 DATA AND METHODOLOGY

We conduct an interactive user study to analyze the extent to which participants can distinguish which skills are automatically activated when a particular invocation phrase is uttered. Upon obtaining responses from our participants, we perform statistical analysis to find any statistically significant trends. We further explain the study design in Section 4.1, the recruitment process in Section 4.2, and the statistical analyses in Section 4.3.

4.1 User Study Design

Our user study was divided into multiple segments, each focusing on eliciting participants' mental models regarding one specific skill-related task at hand. Following is a brief description of different segments in *chronological* order. A link to the full version of the user study is provided in Appendix A.

(1) **General information about skills**: We did not expect participants to know about third-party skills and how they work, even if they already use Alexa. Thus, we explained what third-party skills are and the difference between third-party and native skills before asking any question. The following explanation was provided: "Amazon's voice-based assistant, Alexa, enables users to directly interact with various web services through natural languages dialogues. It provides developers with the option to create third-party applications (known as Skills) to run on top of Alexa. These applications ease users' interaction with smart devices and bolster a number of additional services that native skills developed by Amazon might not otherwise offer. For example, when you say "Alexa, open Jeopardy," Alexa starts the popular game known as "Jeopardy!" — this is an example of a skill that extends Alexa's basic functionality."

(2) **Differentiating third-party skills using visual cues**: This task asked participants to distinguish third-party skills from native skills by visiting the skill's information page on the

Amazon skill store. The purpose of this task was to contrast the efficacy of visual cues with auditory cues currently available on the Alexa skill ecosystem. The participants were provided with the links for three skills – Restaurant Finder [25] (third-party), MyFitnessPal Lite [24] (third-party), and Amazon Storytime [9] (native). They were provided with three options to select one from "Third-party," "Native," and "Note sure". A complete description of this task is provided in Task 1 of the survey available in Appendix A.

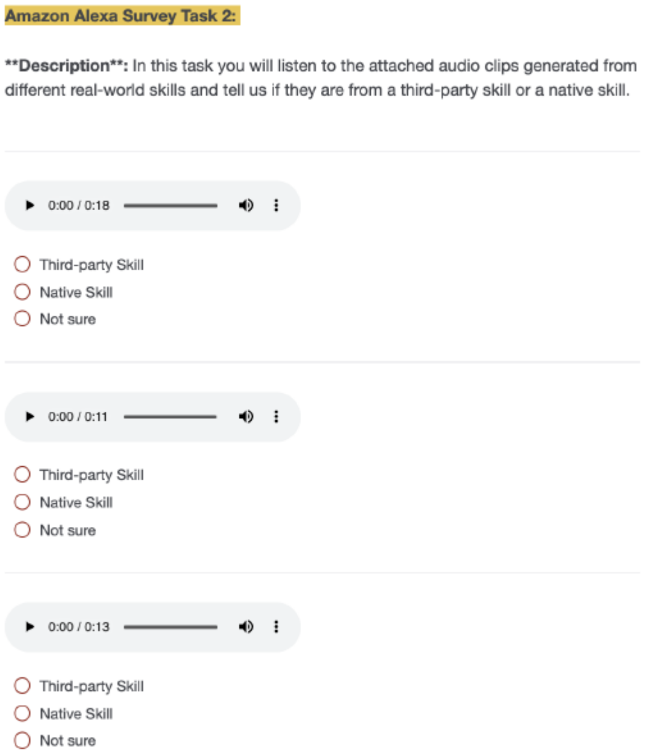
(3) **Differentiating third-party skills using auditory examples**: In this task, we asked participants to listen to auditory excerpts from real skills. Participants were presented with three audio files containing interactions with three real-world skills – Song Quiz [28] (third-party), Translated [29] (third-party), and Call Santa [4] (native). Participants had to identify which samples originated from a third-party skill versus a native skill. Participants had three options to choose from: 'Native', 'Third party', and 'Not sure'. A screenshot of the task is shown in Figure 3.

(4) **Skill selection process**: In this task, participants utilized the official Amazon Alexa app to search for certain skill invocation names. Participants were provided with the skill name in text so that they could search the skill by entering the skill in the search bar. Next, upon receiving the list of skills ranked by Alexa, they were asked to predict which skill would be activated if verbally invoked. Participants were asked to enter the skill name and its developer name for the skill they thought would be activated in the text boxes provided. Participants were guided through the whole process using example screenshots. Task 3 of the survey in Appendix A shows the complete task description.

(5) **Interactive skill invocation**: This task asked participants to verbally invoke the same set of invocation phrases as previously searched (in the task for "Skill selection process") and report which skills are actually activated through the app. This allowed participants to determine whether their predictions (from the previous task) were correct. The participants were thoroughly guided using screenshots about how to invoke a skill and how to find the skills that Alexa enables. Participants were provided with the invocation phrases and text boxes to report each question's skill name and corresponding developer name. A complete description of this task is shown in Task 4 of the survey available in Appendix A.

(6) **Alternative voice-based indicators**: This task focused on obtaining insights from participants regarding how they think Alexa's voice interface can be improved, allowing users to better understand when they interact with third-party skills. Participants were asked to brainstorm different ways on how Alexa could do a better job in differentiating third-party skills from native ones. We provided participants with a free-text box to record their responses (a complete description is shown in Task 5 of the survey available in Appendix A). The purpose of this task was to elicit alternative options from participants to further explore the design of such voice-based indicators as future work. This section also contained one additional task where participants were presented with three different audio templates for a given skill to assess their usability as an indicator for distinguishing third-party skills. We derived these templates through a separate small pilot study of five participants.

We asked participants if any of the three variations helped them better understand the origin of the skill compared to the original version. Participants could select "Yes," "No," or "Maybe" as shown in Figure 4. 49/52 participants answered "Yes," and 3/52 participants answered "Maybe." If the participants selected "Yes" or "Maybe," we asked them to rank the three templates in order of preference and provide reasons for such choices. Figure 4 shows the task description and Figure 5 shows the prompt where participants were asked to rank our custom voice models.



Amazon Alexa Survey Task 2:

Description: In this task you will listen to the attached audio clips generated from different real-world skills and tell us if they are from a third-party skill or a native skill.

○ Third-party Skill
○ Native Skill
○ Not sure

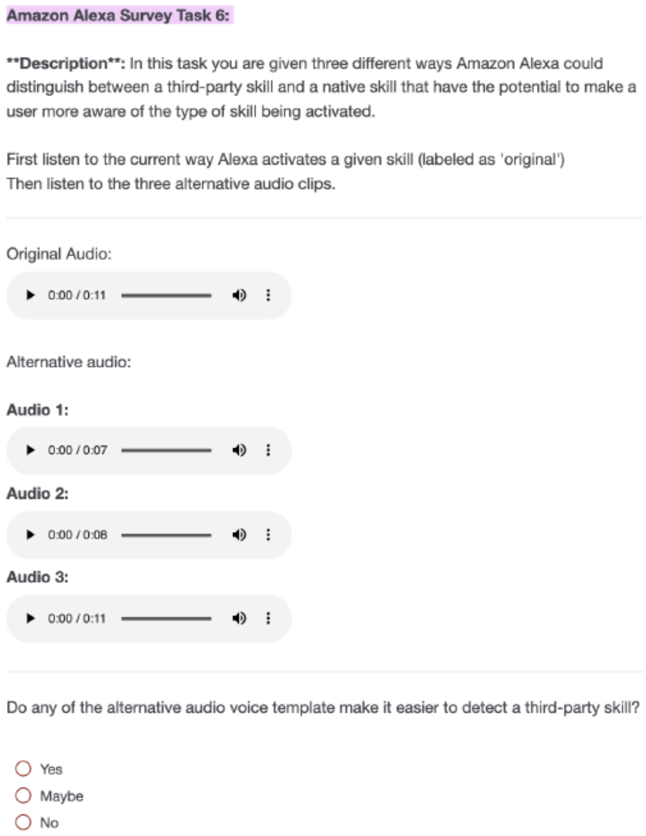
○ Third-party Skill
○ Native Skill
○ Not sure

○ Third-party Skill
○ Native Skill
○ Not sure

Figure 3: Task 2: Participants listen to the given audio files and answer accordingly.

4.2 Recruitment Process

We wanted to recruit participants who had either interacted with the Alexa app or device in the past or actively used it on a regular basis. Thus, after obtaining IRB approval from our institution, we recruited participants via posts on groups and forums specialized for Alexa users (e.g., "EchoTalk" forum [20], "Amazon Alexa Users" Facebook group [7], and Reddit forums [5, 6, 21, 23, 27]) as well as through newsletters and postings to campus mailing lists. In total, we recruited 52 participants. As our study had many interactive components, we wanted to ensure that participants could ask questions in case they did not fully understand any of the tasks. To accomplish this, we scheduled participants to join a virtual Zoom



Amazon Alexa Survey Task 6:

Description: In this task you are given three different ways Amazon Alexa could distinguish between a third-party skill and a native skill that have the potential to make a user more aware of the type of skill being activated.

First listen to the current way Alexa activates a given skill (labeled as 'original')
Then listen to the three alternative audio clips.

Original Audio:

Alternative audio:

Audio 1:

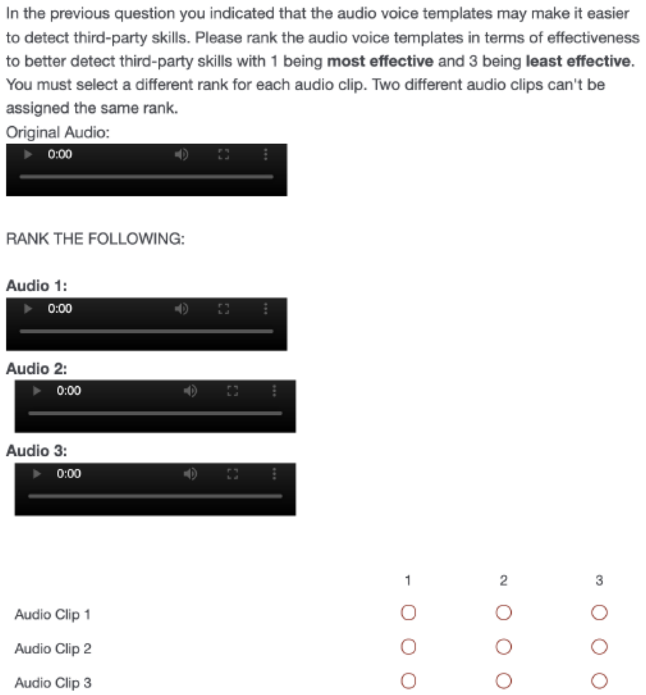
Audio 2:

Audio 3:

Do any of the alternative audio voice template make it easier to detect a third-party skill?

○ Yes
○ Maybe
○ No

Figure 4: Task 6: Participants are shown three alternate audio responses that we created incorporating suggested changes.



In the previous question you indicated that the audio voice templates may make it easier to detect third-party skills. Please rank the audio voice templates in terms of effectiveness to better detect third-party skills with 1 being **most effective** and 3 being **least effective**. You must select a different rank for each audio clip. Two different audio clips can't be assigned the same rank.

Original Audio:

RANK THE FOLLOWING:

Audio 1:

Audio 2:

Audio 3:

	1	2	3
Audio Clip 1	○	○	○
Audio Clip 2	○	○	○
Audio Clip 3	○	○	○

Figure 5: Task 6 (continued): Participants rank three alternate audio responses that we created incorporating suggested changes.

session where each participant was assigned an individual breakout room. The research team members then helped answer any questions participants might have (participants used the raise hand button to request help from the research team members). However, to preserve participants' privacy, no personal information was recorded. On average, it took around 26 minutes for participants to complete all tasks, and each participant was compensated $10 for their time. The whole data collection process lasted for about three months (April 2021 to July 2021). Details on the participants' demographics are shown in Table 1.

4.3 Analytic Methods

For basic summarization, we report the percentage of participants for each unique response to the questions. To compare and contrast alternative voice-based indicators, we conducted Chi-Square tests with pairwise comparisons to test for statistical significance. We consider $\alpha = .05$ as an indicator of statistical significance. The null hypothesis H_0 represents no statistical difference or relationship between the tested factors, whereas the alternate hypothesis H_a indicates a statistical difference between the factors. If the $p-value$ is less than .05, we reject the H_0; otherwise, we do not reject the H_0. We mention $p - value$, χ^2 statistic, and df (degree of freedom) for each respective analysis. Bonferroni's correction was applied to all posthoc analyses to adjust for the risk of a Type I error [1]. We performed Fisher's exact tests to test for significance if conditions for the Chi-Square test are not met, such as low sample size and approximation [61]. To analyze free-text responses, two independent researchers labeled the text responses. We then calculated Cohen's Kappa (κ) to calculate similarities before resolving the conflicts between labels. Kappa scores are reported in their respective sections.

Table 1: Demographics of our 52 participants.

Attribute	Value (count)
Age	18-24 (38), 25-34 (7), 35-44 (2), 45-54 (4), 55-64 (1), 65 or older (0)
Gender	Male (17), Female (35), Prefer not to answer (0)
Employment Status	Employed Full-Time/Part-Time (12), Student (38), Retired (1), Seeking Opportunities (1)

5 IDENTIFYING THIRD PARTY SKILLS

In this section, we address the research question **RQ1:** ***Are users able to distinguish between third-party and native skills?*** This section attempts to answer to what extent users are aware of third-party skills running in Alexa. It also attempts to discover if Alexa users are able to identify third-party skills through voice and visual interfaces. In the following experiments, participants were asked to identify whether a skill is native or third-party, first through the visual web/app interface and then through the voice interface. This allowed us to determine if Alexa's audio and visual interfaces provide sufficient information to distinguish third-party skills from native. We compared the presence and efficacy of security indicators in the visual and voice interface.

5.1 Distinguishing Skill Type through the Visual Interface

The Alexa skill store allows users to browse for skills and offers a similar interface as Google Play Store or Apple App Store for mobile apps. A user can search for skills and open any skill to obtain additional information, such as invocation phrases, permissions, and developer names. An example of a skill information page is shown in Figure 1. We first test if the current skill information page provides sufficient visual/textual indicators to help participants distinguish third-party skills from native skills. For this purpose, we asked participants to visit three web links in the skill store; two third-party skills (MyFitnessPal Lite [24], and Restaurant Finder [26]) and one native skill (Amazon Storytime [9]). Next, we ask them to identify whether each skill was developed by Amazon or a third-party vendor. There were 156 responses from 52 participants (three per participant): 104 were for third-party skills and 52 for a native skill.

5.1.1 Evaluating accuracy. When identifying the skill type through the skill store, the majority of the participants correctly identified the skill types. 80.76% (42/52) of participants correctly identified the skill type for all three skills; we also found that participants were very confident about their decisions. A heat map with accuracy and confidence levels is shown in Figure 6. Furthermore, it is evident that 59.25% (34/52) of the participants were extremely or fairly confident while correctly identifying the skill type for the three test skills.

86/104 (82.69%) of third-party skill tests were identified correctly by the participants. 49/52 (94.23%) of the native skill tests were also correctly identified. If we consider this experiment as classifying third-party skills using visual cues, then the precision and recall are 96.63% and 82.69%, respectively. The confusion matrix is shown in Figure 7b. This suggests that visual indicators were very effective in the case of both native and third-party skill identification.

5.1.2 Correlating confidence levels with accuracy. We further tested to determine if participants' confidence levels are correlated with their accuracy using Fisher's exact test. We found a statistically significant result ($p = .0017$), demonstrating that these two factors are correlated. This matches with Figure 6, which depicts that more confident participants were likely to identify the correct skill type with higher accuracy for the three test skills. This signifies that the skill information page enabled participants to distinguish third-party skills from native skills. Given that the skill store hosts similar metadata about skills like other app stores, it is not surprising that participants correctly distinguished third-party skills from native skills.

5.1.3 Reasoning behind levels of confidence. Participants were also asked to articulate why they felt confident or not confident when distinguishing skills through the skill store information page. Two independent coders labeled the free-text responses; the inter-rater reliability was $\kappa = .50$, and discrepancies were resolved to finalize the labels. We found that the most common reason was participants could clearly see the "developer name" on the skill web page, as stated by 73.08% (38/52) of participants. Other participants described that they recognized the skill type due to its complexity (5.76%, 3/52), branding (3.85%, 2/52), context clues (1.92%, 1/52), and because they were already familiar with the skills (3.85%, 2/52). However, 11.54% (6/52) of the participants did not state a clear reason. This suggests that the 'developer name' is the primary indicator that allows participants to distinguish different types of skills.

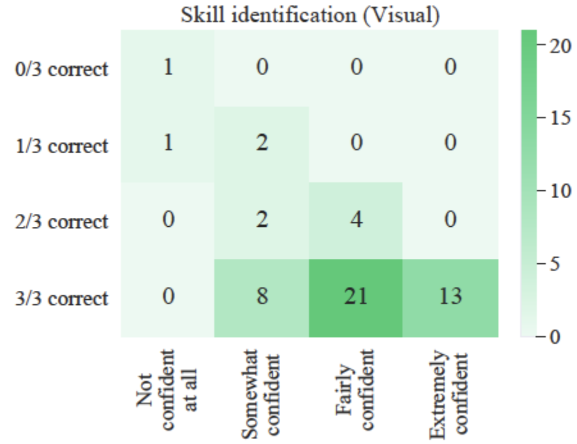


Figure 6: Distribution of correct answers in identifying skill type through the skill information page in respective of the self reported confidence levels.

Following are some responses from our participants.

> P20: There was no specific distinction in the product information section on whether a third party was used. I based my answer off of ... how complex the skill was.

> P35: The developer information at the top of the page states clearly whose skill it is.

> P48: It was easy to differentiate between the third-party skills and the native skills based on the who made the skill. The skill information page included the publisher/creator of the skill.

5.2 Distinguishing Skill Type through the Voice Interface

We asked participants to listen to pre-recorded audio samples from three real-world skills; one native (Call Santa [4]) and two third-party (Translated [29] and Song Quiz [28]). We then asked them to identify whether it was a native or third-party skill. Participants then noted their confidence levels for each recording. Again there were a total of 156 responses: 104 were for third-parry skills and 52 for a native skill. The transcriptions of each audio recording are presented below:

> **Audio Recording 1 - Command:** Alexa, Call Santa
> **Response:** [Music ...] Check the halls with holiday cheer hotline calls. If you like to call Santa and his friends ...

> **Audio Recording 2 - Command:** Alexa, Open Translated **Response:** Okay, here is Translated. Welcome to the translated skill. For instructions say help.

> **Audio Recording 3 - Command:** Alexa, Open Song Quiz **Response:** Welcome to Song Quiz. How many people want to play Song Quiz? You can say one to four.

5.2.1 Evaluating accuracy. Participants listened to the audio interactions from three different Alexa skills and chose whether it was a native or third-party skill. The audio interaction started from the invocation phrase and the first response from the skill. Two out of the three skills were developed by third parties. In this experiment, we investigate the presence and efficacy (or lack thereof) of security indicators that can help users distinguish third-party skills from native skills from a user's point of view. Currently, Alexa does not enforce a defined set of indicators on the voice interface of third-party or native skills that can be used to distinguish skill types.

In the absence of any obvious auditory cue, one would presume participants would randomly guess and correctly identify the skill-type 50% of the time. However, our experiments show that participants perform worse than randomly guessing when identifying third-party skills. Only 31.73% (33/104) of the time participants identified the third-party skills correctly, which suggests that the vast majority of the time, third-party skills lacked appropriate indicators that could help participants distinguish third-party skills from native skills. Moreover, participants underestimated the scope of native skills as 84.61% (44/52) of the times native skills were wrongly identified as third-party. If we consider this experiment as classifying/identifying third-party skills using auditory cues, then the precision and recall are 42.85% and 31.73%, respectively. Figure 7a highlights the confusion matrix. The low precision and recall suggest that the boundary between native and third-party skills is very blurred over the voice interface. It is important to note that this is unlike the result for distinguishing skill type using visual cues, where apparent indicators such as "developer name" helped participants distinguish the skill type with high precision (96.63%) and recall (82.69%).

We found that 44.23% (23/52) of participants responded incorrectly for all three skills, while only 36.53% (19/52) answered one skill correctly. 15.38% (8/52) of participants answered two out of three answers correctly, and only 3.84% (2/52) responded with all correct answers. Thus, it is evident that the voice interface does not provide sufficient auditory cues to allow participants to accurately distinguish third-party skills from native skills. Since skills have dynamic content and can obtain sensitive and personal information from users, they ought to be informed about what types of skills are being activated [52]. Moreover, since data used in third-party skills are stored and handled by third-party servers, it is important that users are aware when they interact with a third-party skill.

5.2.2 Correlating confidence levels with accuracy. We also asked participants to rate their confidence levels about identifying the skill on a scale of 1 to 4 (1 being not confident at all and 4 being extremely confident). We found that most of the participants were "somewhat confident" (59.61%, 31/52) and "not confident at all" (23.07%, 12/52). Only 17.31% (9/52) of participants were "fairly confident" and none were "extremely confident." We compared the confidence levels with the levels of accuracy and found that the majority of participants were not only incorrect but also not confident about their responses. This heatmap is depicted in Figure 8.

We also performed Fisher's exact test to determine whether the participants' levels of accuracy varied by confidence levels. We found that the participants' accuracy was not statistically significantly different across the confidence levels ($p = .4788$). Thus, a participant's confidence levels did not correlate with correctly identifying third-party skills. This demonstrates that Alexa does not clearly convey whether users are interacting with a native or a third-party skill.

5.2.3 Reasoning behind levels of confidence. Given that there are no obvious and consistent indicators present on the voice interface, we cannot pinpoint a ground-truth indicator that participants could have used to identify/guess skill type. However, after participants marked their confidence levels for accurately distinguishing third-party skills from native ones, they were asked for relevant reasoning behind their confidence ratings. To analyze these free-text responses, two independent researchers labeled the responses based on their content and then compared both labels. We calculated the Cohen Kappa score to find the similarity in labeling and found $\kappa = .730$. The conflicting labels were resolved after a discussion, which led to the final distribution of labels. We found that the most common reason was that participants underestimated the capability of native skills. 34.64% (18/52) of participants thought that Alexa's native skills could not perform complicated tasks, so it must be a third-party skill. 32.69% (17/52) participants were falsely confident that a change of voice means that a third-party skill was activated, but native skills can also utilize different voices. Additionally, 7.69% (4/52) and 3.85% (2/52) of participants thought that they were confident because of different wording within the audio output and because they were already familiar with skills, respectively. Yet, 21.15% (11/52) of participants did not provide any clear reason because they were simply not sure.

Following are some responses from our participants about their reasoning.

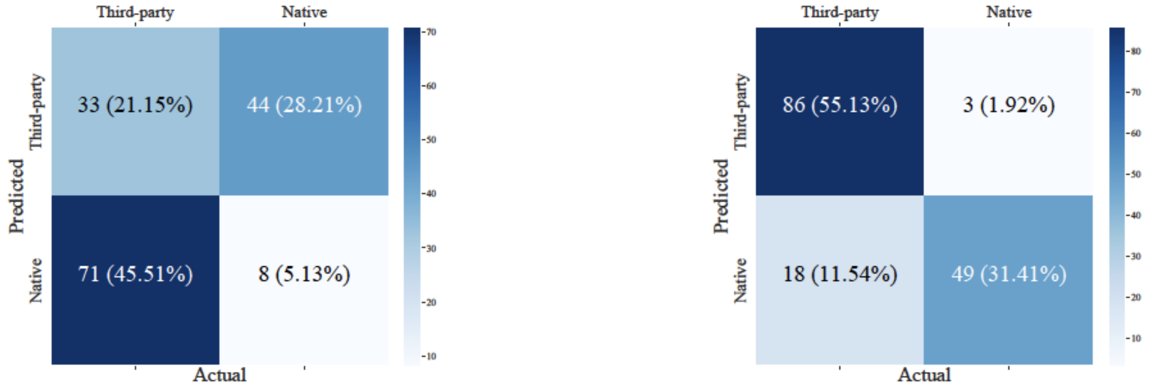
> P13: The loudness of the voice makes it confusing. The louder the voice, the more native it sounds to me at least.

> P23: I picked based on if the voice changed from Alexa to something else, but I have no idea whether that's how it works.

> P39: I feel that Amazon only makes skills for basic things, such as the weather and other simple programs. The ones in the examples felt like they were more from other developers.

6 DIFFERENTIATING SIMILAR SKILLS

In this section, we seek an answer to the following research question, RQ2: ***With Alexa's auto-enable feature, can users predict which skill will be activated when invoked?*** Skill invocation names are not unique – there can be multiple skills with similar or same invocation names. For instance, Lentzsch et al. found 9,948 skills that share the same invocation name with at least one other skill in US skill store [52].



(a) Confusion matrix for voice interface (precision = 42.85%, recall = 31.73%)

(b) Confusion matrix for visual interface (precision = 96.63%, recall = 82.69%)

Figure 7: Confusion matrices for voice and visual interfaces while distinguishing between native and third-party skills.

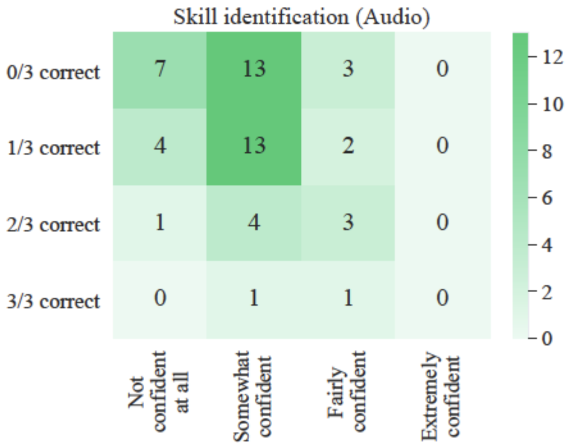


Figure 8: Distribution of correct answers in identifying skill type through the voice interface in respective of the self reported confidence levels.

If duplicate invocation phrases exist between a native and a third-party skill, Alexa will prioritize the native skill. However, if no native skill exists, there is no publicly known algorithm that determines which skill will be activated (Amazon uses an internal selection scheme). Nevertheless, the skill selection process is consistent as the same skill is activated across multiple rounds of activation [52]. Researchers, however, have performed different blackbox tests to demystify the skill selection process. Zhang et al. showed that if a skill's invocation name is contained in another skill's invocation name, Alexa prefers the longest match [65]. However, the system's behavior is not known when two skills share the same invocation name. Lentzsch et al. analyzed different potential attributes such as "number of ratings," "average rating," and "age of skill"; however, none of their results implied causation [52]. This further emphasizes the need for auditory interventions to ensure the desired skill is being activated.

The purpose of this experiment is to evaluate if the skill selection process aligns with users' expectations. As Alexa by default enables a matching skill (the details of the skill selection process is unknown), it is unclear if the right/desired skill will be activated; this might give rise to privacy concerns because users can enable an unintended skill, leading to potential privacy leaks and consequences.

6.1 Experimental Setup

In order to analyze whether participants can correctly identify what skills they interact with, the user study includes an experiment with the Amazon Alexa app. Participants were asked to download the latest version of this app (either on Android or iOS phones) and were instructed to log into their accounts. Once on the app's home page, participants could begin the experiment. Participants were asked to report the previously enabled skills on their account (if any). This is to avoid cases where participants might already have some of the test skills enabled. We found no such cases in our participant pool.

6.1.1 Participants' predictions. The participants were asked to search for three distinct invocation names — "daily horoscope," "baby names," and "currency converter" — using the search feature of the Alexa app under the "Skill & Games" section from the app menu. The skill chosen for this experiment had at least *three* other alternatives with the same skill name and invocation name. The three skills used in this experiment and three alternatives that appeared in the search results for each of the three skills are provided as follows: Daily Horoscope (by a.myers.inc [17], by marks_matters [18], by GV Skills [19]), Baby Names (by stringfree [10], by Piperal Technology [11], by Hatem Elseidy [12]), and Currency Converter (by implemica [14], by Sam Sepiol [15], by Logical Enigma [16]). Moreover, we found the search results (including the order) to be consistent. We tested across multiple accounts and devices (five in total) and found the same consistent search results throughout the data collection process. We, therefore, believe the participants all saw the same search results.

Participants were prompted to search for a particular skill by retyping the invocation name from the survey into the search bar on their phones. Then, participants were asked to select one skill (among all returned matching skills) that they thought would be activated when verbally invoked. The participants then opened the selected skill and entered the skill name and developer name from the skill information page to the respective survey fields. This process was repeated for all three skills. Participants were also asked how confident they were that the skills they chose would be the ones that will be activated when verbally invoked. Furthermore, participants were asked what made them feel confident or not confident about their decision.

6.1.2 Live skill invocation. After manual predictions, participants were asked to verbally utter the skills' invocation phrases to Alexa and report which skills were actually activated. The participants were provided instructions on how to invoke the skills using the Alexa app. Once a skill was invoked, we prompted participants to navigate to the "Your Skills" tab (in the "Skill & Games" menu option) to report the name and developer of the skill that was recently enabled (Task 6 of the survey provided in Appendix A contains instructions provided to participants about how to identify what skill was actually activated). This allowed us to compare whether participants were accurate with their skill predictions. Note that in all three cases, the actual skill that was activated also appeared in the search result (i.e., it was one of the skills returned by the search result). If participants observed any differences between their predicted and activated skills, we also asked what steps could be taken to resolve this discrepancy.

6.2 Analysis

We evaluated the extent to which participants were able to correctly predict the right skill by comparing their predictions with the actual enabled skills. We found that many participants predicted the wrong skill.

6.2.1 Evaluating accuracy. 44.23% (23/52) of the participants wrongly predicted all three skills, 36.54% (19/52) correctly guessed only one out of three skills, 15.38% (8/52) predicted two out of three correctly, and only 3.84% (2/52) guessed all three skills correctly. This demonstrates that participants are likely unaware about what skill Alexa is actually interacting with, potentially resulting in sharing their personal information with an unknown entity, especially one that may not be malicious in nature.

75% (39/52) of the participants reported that they found a difference in the skills that were actually activated from what they thought would be activated. Only 13.46% (7/52) said that they did not see any difference, and 11.54% (6/52) participants were not sure.

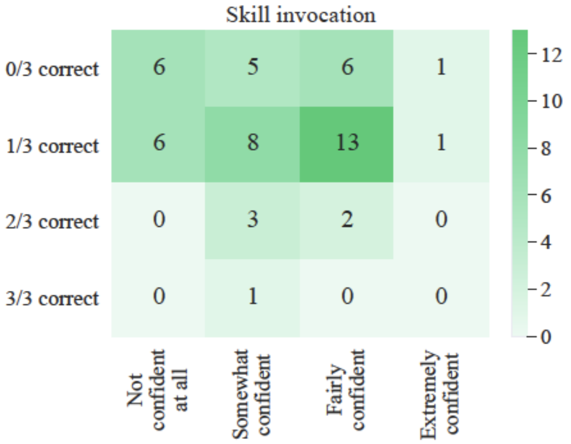


Figure 9: Distribution of correct answers in identifying which skill will be activated in respective of the self reported confidence levels.

6.2.2 Correlating confidence levels with accuracy. We asked participants how confident they were about their answers on a four-point scale (1 being not confident at all and 4 being extremely confident). We found that only 3.84% (2/52) participants were "extremely confident" and 40.38% (21/52) were "fairly confident," while the majority of participants (55.78%, 29/52) were "somewhat confident" or "not at all confident." Figure 9 depicts the distribution of accuracy across different confidence levels. We see that most of the responses, regardless of levels of confidence, were mostly incorrect. This points towards participants' lack of ability to correctly predict the skill to be activated.

We conducted Fisher's exact tests to evaluate the correlation between correctness and confidence; the result is not significant at $p = .6900$. Thus, no correlation was found between accuracy and confidence, and as shown in Figure 9, that confidence level can vary while still having mostly incorrect answers. This also points to a flaw in Alexa's voice interface where users cannot determine if they are talking to and sharing data with the correct service. Thus, more research is required in designing appropriate voice-based interventions to create user awareness.

6.2.3 Reasoning behind levels of confidence. We asked participants to describe why they were confident or not confident about their predictions. Two independent researchers labeled the free-text responses and calculated the similarity in labels; Cohen's Kappa was found to be $\kappa = .547$. The conflicting labels were then discussed and resolved. We found that the skills' "ranking" in the search results was the most common element participants took into account. 21.15% (11/52) of the participants stated that they were confident that skills that appeared on top of search results would be invoked. Other participants looked at other pieces of skill information: 15.38% (8/52) were confident because of the "invocation phrase," 9.62% (5/52) due to "Good reviews," 7.69% (4/52) based on "relevance," and 3.84% (2/52) because they were already familiar with skills. However, we also found that 38.46% (20/52) of the participants did not understand why certain skills were selected. 3.84% (2/52) of participants did not respond to this question.

Following are some responses from our participants.

> P8: There were many options of practically the same skill so it wasn't super clear which one would default open first.

> P27: These were the first ones pulled up when I searched in the Alexa app, so they should be the first ones pulled by the phrases.

> P44: I picked the first skill or the one with the best ratings.

7 IMPROVING ALEXA'S VOICE INTERFACE

One common underlying issue we found from our analysis in Sections 5 and 6 was that participants had little awareness regarding the ownership of the skills they interacted with. In this section, we seek answer to the research question, RQ3: ***How can Alexa better inform its users about the ownership of the skills they interact with?***

7.1 Experimental Setup

We leverage participants' fresh experience with Alexa skills to obtain insights into how Alexa can better inform users about skills. We asked participants to brainstorm different ways in which Alexa could do a better job at distinguishing native and third-party skills. Participants were provided with a free-text box to write their views. After this question, we asked participants to rank three different alternative audio templates in terms of effectively conveying ownership of a skill ("wikiHow" [30]). We also provide the default audio response as a baseline so that participants can easily compare them before ranking them from 1 (the best) to 3 (the worst).

Before launching the main user study, we conducted a short pilot study with five participants to obtain some initial insights into their suggestions for improved skill interactions. The pilot study had the free-text question asking them to suggest changes in the interaction model that would help them better identify the skill type. We analyzed the free-text responses and picked the most suggested recommendations. Moreover, in the main study, we asked participants to brainstorm ideas to improve privacy indicators before presenting our suggested variations to avoid any priming effect. We, however, found that participants' most frequent suggestions were similar to those in the pilot study. The three variations consisted of the characteristics that participants from our pilot study had pointed out. While various combinations of the three templates are possible, we only considered the three basic templates to limit the amount of time participants would spend on the task (and thereby reduce user fatigue). A full-fledged analysis of all possible auditory interventions is left as future work.

The characteristics of the three voice templates are as follows: i) A warning phrase saying "You are about to enable a third-party skill" followed by the developer name "by wikiHow," ii) Change in voice tone only when the skill was invoked, and iii) A warning phrase saying "You are about to enable a third-party skill" followed by the developer name "by wikiHow" and the skill starts with a different voice. The three audio templates are summarized below:

> **Model 1:** Command: Alexa open wikiHow.
> Response: You are about to enable a third-party skill developed by wikiHow.

> **Model 2:** Command: Alexa open wikiHow.
> Response: [In different voice] Okay. Here is wikiHow, Hi! Ask me anything.

> **Model 3:** Command: Alexa open wikiHow.
> Response: [In different voice] You are about to enable a third-party skill developed by wikiHow. Okay. Here is wikiHow, Hi! Ask me anything.

Table 2: Ranking of alternative voice-based interventions.

Voice-based Intervention	Rank 1	Rank 2	Rank 3
Model 1 (Warning Phrase + Developer information)	31	14	7
Model 2 (Change in voice only)	6	12	34
Model 3 (Warning + Change in voice + Developer information)	15	26	11

We asked participants to rank these three choices. We found that most participants liked the first option with a warning phrase with the developer's name. The second preference was given to the third option with a warning phrase with the developer's name information and a change in voice tone. The least preferred option was the second model, which only had a voice tone change. Table 2 shows a distribution of the ranking across the three options.

7.2 Ranking Explanations

We asked participants to explain their decision for ranking via a free-text field (shown in Task 6 of the survey available in Appendix A). We found that the majority of participants (59.62%, 31/52) liked the warning phrase because it explicitly warns users about using third-party skills. Participants also preferred including the developer's name in the skill's response to easily denote a third-party skill. However, changing the voice tone was the least preferred option for participants as it seemed like "overkill" and did not naturally integrate well with Alexa. Thus, we find that including the developer's name and a warning phrase about using a third-party skill best notified participants about using a third-party skill.

Following are some responses we obtained.

> P34: The change in voice is a good clue, but also having it specifically state that it is a skill developed by someone else makes it very clear.

> P44: Saying that it is third-party helps differentiate it.

> P47: It was a bit weird when the voice changed, I would rather it just said it was opening a 3rd party skill.

7.3 Analyzing Statistical Difference in Rankings

To analyze if the different interventions differ significantly in rankings, we ran Chi-Square tests with the significant boundary of $\alpha = .05$. To perform Chi-Square comparisons, we divided the votes of all three variations into three groups of Rank 1, 2, and 3. Since all participants ranked all variations and no two variations can be ranked the same by a participant, we had 52 votes for each variation (represented by rows in the contingency table 2), and similarly, we had 52 votes in each rank group. This comparison helped us determine if any of the three variations are ranked somewhat similar by participants (e.g., if two variations are ranked 1 by a similar number of participants) or significantly different, which helped us determine the most popular choice. We found that all our voice-based interventions were ranked significantly different from each other where $\chi^2(df = 4, n = 156) = 49.6154$ and $p < 0.00001$. We also performed posthoc analysis and tested every pair of interventions;

Table 3: Chi-Square tests for *rank* comparisons. $df = 2$ for all tests. * represents significant $p - value$.

Rank Pairs	χ^2 *statistic*	$p - value$
Rank 1 vs. Rank 2	11.3734	*0.010173
Rank 2 vs. Rank 3	18.9362	*0.000231
Rank 1 vs. Rank 3	35.3733	*<0.00003

we found that all pairs are significantly different from one another. Table 3 shows the χ^2 statistic and $p - values$ for all possible pairs of rankings. The significant statistical differences imply that the user preferences were clear and distinct, and no two rankings were similar. Thus, we can confidently say that providing a warning message with the developer's name was the most preferred voice-based intervention for conveying ownership of skill.

7.4 Users' Suggestions on Indicators

We asked participants to describe what changes in Alexa's voice interface would make it easier to help users differentiate third-party skills. The participants responded using a free-text field. To analyze the free-text responses, two independent researchers labeled the responses based on the underlying theme, and then both labels were compared. We calculated the Cohen Kappa score and found it to be $\kappa = .503$. We next resolved the differences through discussion and calculated the final distribution of labels. We found the most common response was that Alexa should say the "Developer name" of the skill that it is invoking, such as "Opening <skill name> by <developer name>." 25% (13/52) of the participants stated that verbally saying the developer's name would serve the purpose of making people aware of the skill. 9.62% (5/52) participants stated that using a different keyword for third-party skill would be helpful; the same number of participants (9.62%, 5/52) advocated for using a warning phrase and playing a different tone or flashing light on the Alexa device when a third-party skill is invoked. Another 9.62% (5/52) stated that a warning should be added before each third-party invocation. 7.69% (4/52) of the participants said that Alexa should ask for confirmation if the user wants to open a third-party skill. 15.38% (8/52) participants were not very clear or certain about the suggestions. Additionally, 23.07% (12/52) of the participants discussed more unique options, such as stating Amazon's name before invoking native skills or providing a warning on only the first invocation of the third-party skill.

We found some interesting insights from the free-text responses. For example, some participants explicitly mentioned that a warning phrase should be added only when a skill is invoked/installed for the first time, while others mentioned annoyance caused by it (further discussed in the discussion section 8). However, participants suggested the use of developer names with every invocation and did not mention any annoyance with such an approach. We, therefore, suggest using an explicit warning phrase with the first invocation/activation, while the developer name should be mentioned with every invocation.

Following are some interesting responses from participants.

> P2: It could be as simple as "Opening x skill by so-and-so." But it might be easier long-term to verify that the user is aware that they are adding a skill by a third-party when they first "install" or "enable" the skill and not with every invocation.

> P7: She could say "okay- opening [skill name] by [developer name]" so that it's clear it's opening a third-party software.

> P21: I feel like Amazon Alexa could say something along the lines of "I am fetching a skill from another source to share with you. Would you like to continue using it?" Something along those lines to differentiate would be good.

8 DISCUSSION

We conducted an interactive user study to analyze Amazon Alexa users' perceptions and awareness about third-party skills. Our user study focused on asking participants to identify whether skills were developed by Amazon (native skills) or by a third-party (third-party skills) both through the skill's visual interface (e.g., skill store page) and through the voice interface. We also evaluate how easy it is to predict which skill Alexa will auto-enabled when an invocation phrase is actually uttered. The purpose of these experiments was to understand the gap that exists between a user's mental model and the actual working of the system. In this section, we discuss how this gap can make users more vulnerable to skill-based attacks and privacy risks. We also suggest recommendations based on feedback obtained from our study and highlight their implications. Lastly, we layout future directions in this space.

8.1 Potential Reasons for Incomplete Mental Model

We found that the majority of participants successfully identified a skill's developer using information displayed on the skill's information page, as the 'developer name' was clearly displayed. However, 96.15% (50/52) of participants were unable to correctly identify at least one skill with Alexa's voice interface, citing many different reasons for their confusion. *This suggests that while there are sufficient usable security/privacy indicators on the skill information page, such indicators are lacking on the voice interface.*

Furthermore, we asked participants to predict what skills would be invoked for a given invocation name by utilizing the Amazon Alexa app. We asked participants to search for a skill and then invoke them through voice command to see if their expectations matched with what Alexa actually did. This enabled us to help participants better contextualize the gap in their mental model regarding how Alexa automatically enables skills. Many participants predicted a skill based on its ranking in the search results and yet found that Alexa's chosen skill differed from their choices, creating ambiguity in terms of what skill would automatically activate. *Due to the lack of transparency on how Amazon auto-enable skills with similar invocation names, users can easily activate the wrong skill, which can lead to unwanted information disclosure.*

8.2 Implications of Incorrect Mental Model

As third-party skills are difficult to distinguish from native skills, an attacker can exploit this discrepancy to make users think that they are interacting with a different skill and coax them into disclosing personal information [59]. Similarly, a malicious third-party skill can fake skill termination and continue to run and listen to a user's conversation stealthily, as shown by Bräunlein et al. [2]. A malicious skill can also use intents that are not evaluated during the skill vetting process but later activated to obtain sensitive data from users that are typically protected through permission APIs [52]. Given these privacy and security risks associated with Alexa skills, it is essential to introduce security and privacy indicators on the voice interface to help users become more aware of the applications they are interacting with.

The lack of effective indicators on the voice interface is further exacerbated when such devices are shared among multiple users, as some users are more likely to be unaware of the third-party skills enabled on the devices. Khoshy et al. [47] found that passenger users (who do not configure device themselves) have less knowledge about smart home devices' functionalities compared to pilot users (who configure the device for household use), which suggests that it would be even harder for passenger users to distinguish third-party skills from native skills. Their findings further motivate the need for explicit privacy indicators in the Alexa skill ecosystem. Similarly, Huang et al. [46] found that users were uncomfortable with sharing certain information with household members even when such information is readily available in the cloud. Although they did not explicitly ask participants regarding third-party skills, instead, voice assistants in general, they found participants did not even trust the first-party manufacturers of the voice assistants, which further raises concerns regarding third-party services enabled on the devices.

8.3 Recommendations and Future Work

As existing works have shown that third-party skills can involve numerous privacy and security risks [2, 52, 59], Alexa's failure to provide users with proper notification of third-party skills poses potential risks to consumers. According to our results, the platform currently lacks effective voice-based indicators/interventions to inform users about third-party skills. Major et al. [56] made some suggestions to incorporate privacy indicators for third-party skills. However, they did not test any specific voice template in their study. Our user study identified and tested three voice-based interventions that incorporate a warning message with the developer name, changed the voice tone, or a combination of both. We found that most participants preferred the simple warning message with the developer name as a good indicator for interaction with third-party skills. We, therefore, suggest explicitly including a warning message about a third-party vendor in Alexa's initial response to activating a third-party skill. Upon analyzing free-text responses from participants, we found that participants would be fine with mentioning the developer name at the start of each session, while the warning message regarding the activation of third-party skills may be limited to only the first occurrence of such skills.

While the free-text suggestions made by our participants resonated with the most highly ranked interventions tested, there is still room for studying other ways of introducing privacy interventions and raising awareness; we leave such efforts as future work. In general, we believe that more work is needed to improve data transparency in the Alexa skill ecosystem and understand the tradeoff between transparency and usability. In this study, we only focused on the effectiveness of different variations of voice-based interventions for the Alexa interface, but a more thorough evaluation of how such interventions affect a user's seamless experience over time requires further investigation. Furthermore, the frequency and timing of such interventions for the voice interface are still open research questions. Even in our study, some participants raised the question of repetitive explanations as potentially a source of annoyance. More research is therefore required to develop models of when and how frequently users would be receptive to interventions on emerging voice interfaces. Such models can further foster the design of customized interventions (in terms of both message and frequency) when tested across populations of different cultures or technical backgrounds.

9 LIMITATIONS

In order to maximize the quality of user responses, we attempted to keep the average survey time between 20 and 30 minutes. This allowed us to ask participants questions about only three skills, yet including more diverse skills could provide additional insights. Similarly, a larger pool of participants (currently 52) could be obtained; however, a small sample size enabled us to obtain high-quality responses by hosting live virtual sessions over Zoom with every participant. However, our participant pool is skewed towards young females. Most of them were university students, which might have introduced unwanted bias in the result; nevertheless, we believe that our findings still hold for a tech-savvy population. Some of the participants for our study were recruited from Reddit forums for smart home enthusiasts, which may not represent an average Alexa user; however, it is notable that even tech-savvy participants had confusion regarding skill identification which suggests our results are lower-bound estimates of the skill identification problem. Although our study assumes that participants may not have complete and correct knowledge regarding third-party skills, some might have formed their mental models from the initial explanations provided at the beginning of the survey. Lastly, we evaluate the effectiveness of our proposed voice-based interventions by playing audio clips. A more realistic deployment could involve deploying our own skills and asking participants to directly interact with such skills. We leave a full-fledged study on the usability and effectiveness of different voice-based interventions as future work.

10 CONCLUSION

Our user study concludes that the Alexa voice interface lacks proper auditory interventions that are critical for minimizing the security and privacy risks imposed by third-party applications running on top of Alexa. We found that participants failed to differentiate third-party skills from native skills through the voice interface; however, participants could make such distinction through the skill information pages (due to the presence of visual cues). We also found a significant gap in participants' mental model and how Alexa selects and auto-enables skills. Lastly, we briefly explored the design of plausible voice-based interventions and their efficacy using prototype auditory templates. Based on popular choice by participants, we suggest that Alexa should utter a warning phrase informing participants when they first activate a third-party skill, while the developer name can be mentioned at the start of each subsequent session. However, we believe this is only the first step towards designing practical interventions for voice interfaces, and a more thorough analysis of the usability and effectiveness of different voice-based interventions is required.

ACKNOWLEDGMENTS

We thank our user study participants for undertaking our interactive experiments and providing insights to improve Amazon Alexa's voice-based interface. We also thank our anonymous reviewers for their valuable feedback. This material is based upon work supported in parts by the National Science Foundation under grant number CNS-1849997 and Amazon Alexa Research award. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] 2015. Handbook of Biological Statistics. http://www.biostathandbook.com/chiind.html

[2] 2019. Smart spies: Alexa and Google Home expose users to vishing and eavesdropping. https://www.srlabs.de/bites/smart-spies

[3] Review.com 2020. *The Best Voice Assistants*. Review.com. https://www.reviews.com/home/smart-home/best-voice-assistant/

[4] 2021. Alexa, Call Santa. https://www.amazon.com/Amazon-Call-Santa/dp/B07Z9KCZSL

[5] 2021. Amazon Alexa. https://www.reddit.com/r/alexa/

[6] 2021. Amazon Echo | A voice command system that brings the Internet Of Things to your home. https://www.reddit.com/r/amazonecho/

[7] 2021. Amazon Echo (Alexa) users. https://www.facebook.com/groups/ECHOBYAMAZON

[8] 2021. Amazon Skill Certification Requirements. https://developer.amazon.com/en-US/docs/alexa/custom-skills/certification-requirements-for-custom-skills.html

[9] 2021. Amazon Storytime. https://www.amazon.com/Amazon-Education-Consumer-Team-Storytime/dp/B073X5FYVF

[10] 2021. Baby Names. https://www.amazon.com/stringfree-Baby-Names/dp/B07SXR3D1V

[11] 2021. Baby Names. https://www.amazon.com/Piperal-Technology-Baby-Names/dp/B07QYW7LHX

[12] 2021. Baby Names. https://www.amazon.com/Hatem-Elseidy-Baby-Names/dp/B07L1KFZ6Q

[13] 2021. Certification Requirements. https://developer.amazon.com/en-US/docs/alexa/custom-skills/certification-requirements-for-custom-skills.html#submission-checklist

[14] 2021. Currency Converter. https://www.amazon.com/implemica-Currency-Converter/dp/B083Q24TVR

[15] 2021. Currency Converter. https://www.amazon.com/Sam-Sepiol-Currency-Converter/dp/B07MY49DQX

[16] 2021. Currency Converter. https://www.amazon.com/Logical-Enigma-Currency-Converter/dp/B01MS27WLR

[17] 2021. Daily Horoscope. https://www.amazon.com/marks$_m$atters-Daily-Horoscope/dp/B073ZQV61R

[18] 2021. Daily Horoscope. https://www.amazon.com/a-myers-inc-Daily-Horoscope/dp/B074WMR3M2

[19] 2021. Daily Horoscope. https://www.amazon.com/GV-Skills-Daily-Horoscope/dp/B0872SDHY5

[20] 2021. Echo & Alexa User Discussions and Support Forums. https://www.echotalk.org/index.php

[21] 2021. HomeAutomation. https://www.reddit.com/r/homeautomation/

[22] 2021. Incredible Amazon Alexa Statistics You Need to Know in 2021. https://safeatlast.co/blog/amazon-alexa-statistics/

[23] 2021. Let's Discuss Alexa Skills! https://www.reddit.com/r/Alexa$_s$kills/

[24] 2021. MyFitnessPal Lite. https://www.amazon.com/Under-Armour-Inc-MyFitnessPal-Lite/dp/B07QN179C5

[25] 2021. Restaurant Finder. https://www.amazon.com/TheHumbleOne-Restaurant-Finder/dp/B074K9MPNX

[26] 2021. Restaurant Finder. https://www.amazon.com/Garrett-Vargas-Restaurant-Finder/dp/B01N76G9H5

[27] 2021. SmartHome. https://www.reddit.com/r/smarthome/

[28] 2021. Song Quiz. https://www.amazon.com/Volley-Inc-Song-Quiz/dp/B06XWGR7XZ

[29] 2021. Translated. https://www.amazon.com/Translated-Labs/dp/B01N9BZJPZ

[30] 2021. wikiHow. https://www.amazon.com/wikiHow/dp/B01NAI70T7

[31] 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)*.

[32] 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.

[33] Devdatta Akhawe and Adrienne Porter Felt. 2013. Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness. In *Proceedings of the 22nd USENIX Security Symposium (USENIX Security)*. 257–272.

[34] Efthimios Alepis and Constantinos Patsakis. 2017. Monkey Says, Monkey Does: Security and Privacy on Voice Assistants. *IEEE Access* 5 (2017), 17841–17851. https://doi.org/10.1109/ACCESS.2017.2747626

[35] Tawfiq Ammari, Jofish Kaye, Janice Y. Tsai, and Frank Bentley. 2019. Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Transactions on Computer-Human Interaction* 26, 3 (2019), 1–28.

[36] Chaitrali Amrutkar, Patrick Traynor, and Paul C. van Oorschot. 2015. An Empirical Evaluation of Security Indicators in Mobile Web Browsers. *IEEE Transactions on Mobile Computing* 14, 5 (2015), 889–903.

[37] Si Chen, Kui Ren, Sixu Piao, Cong Wang, Qian Wang, Jian Weng, Lu Su, and Aziz Mohaisen. 2017. You Can Hear But You Cannot Steal: Defending Against Voice Impersonation Attacks on Smartphones. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. 183–195. https://doi.org/10.1109/ICDCS.2017.133

[38] 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. https://doi.org/10.1145/3372297.3423339

[39] Jide S Edu, Xavier Ferrer-Aran, Jose M Such, and Guillermo Suarez-Tangi. 2021. SkillVet: Automated Traceability Analysis of Amazon Alexa Skills. arXiv:2103.02637

[40] Jide S. Edu, Jose M. Such, and Guillermo Suarez-Tangil. 2020. Smart Home Personal Assistants: A Security and Privacy Review. *ACM Comput. Surv.* 53, 6, Article 116 (dec 2020), 36 pages. https://doi.org/10.1145/3412383

[41] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. 2008. You've Been Warned: An Empirical Study of the Effectiveness of Web Browser Phishing Warnings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 1065–1074.

[42] Huan Feng, Kassem Fawaz, and Kang G. Shin. 2017. Continuous Authentication for Voice Assistants. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking* (Snowbird, Utah, USA) *(MobiCom '17)*. Association for Computing Machinery, New York, NY, USA, 343–355. https://doi.org/10.1145/3117811.3117823

[43] 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.

[44] Hang Hu, Limin Yang, Shihan Lin, and Gang Wang. 2020. A case study of the security vetting process of smart-home assistant applications. In *2020 IEEE Security and Privacy Workshops (SPW)*. IEEE, 76–81.

[45] Hang Hu, Limin Yang, Shihan Lin, and Gang Wang. 2020. Security Vetting Process of Smart-home Assistant Applications: A First Look and Case Studies. arXiv:2001.04520 [cs.CR]

[46] 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. https://doi.org/10.1145/3313831.3376529

[47] Vinay Koshy, Joon Sung Sung Park, Ti-Chung Cheng, and Karrie Karahalios. 2021. *"We Just Use What They Give Us": Understanding Passenger User Perspectives in Smart Homes.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445598

[48] Deepak Kumar, Riccardo Paccagnella, Paul Murley, Eric Hennenfent, Joshua Mason, Adam Bates, and Michael Bailey. 2018. Skill squatting attacks on Amazon Alexa. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*. 33–47.

[49] Veton Këpuska and Gamal Bohouta. 2018. Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home). In *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*. 99–103. https://doi.org/10.1109/CCWC.2018.8301638

[50] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, Are You Listening?: Privacy Perceptions, Concerns and Privacy-seeking Behaviors with Smart Speakers. *ACM Transactions on Computer-Human Interaction* 2, CSCW

(Nov. 2018), 102:1–102:31.

[51] Xinyu Lei, Guan-Hua Tu, Alex X Liu, Kamran Ali, Chi-Yu Li, and Tian Xie. 2017. The Insecurity of Home Digital Voice Assistants–Amazon Alexa as a Case Study. *arXiv preprint arXiv:1712.03327* (2017).

[52] 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. In *28th Annual Network and Distributed System Security Symposium (NDSS 2021). The Internet Society.*

[53] Song Liao, Christin Wilson, Long Cheng, Hongxin Hu, and Huixing Deng. 2020. Measuring the effectiveness of privacy policies for voice assistant applications. In *Annual Computer Security Applications Conference (ACSAC)*. 856–869.

[54] Jialiu Lin, Shahriyar Amini, Jason I. Hong, Norman Sadeh, Janne Lindqvist, and Joy Zhang. 2012. Expectation and Purpose: Understanding Users' Mental Models of Mobile App Privacy through Crowdsourcing. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp)*. 501–510.

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

[56] 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 Trans. Internet Technol.* 22, 1, Article 11 (sep 2021), 22 pages. https://doi.org/10.1145/3446389

[57] Atif M. Memon and Ali Anwar. 2015. Colluding Apps: Tomorrow's Mobile Malware Threat. *IEEE Security Privacy* 13, 6 (2015), 77–81. https://doi.org/10.1109/MSP.2015.143

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

[59] Atsuko Natatsuka, Ryo Iijima, Takuya Watanabe, Mitsuaki Akiyama, Tetsuya Sakai, and Tatsuya Mori. 2019. Poster: A First Look at the Privacy Risks of Voice Assistant Apps. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* (London, United Kingdom) *(CCS '19)*. Association for Computing Machinery, New York, NY, USA, 2633–2635. https://doi.org/10.1145/3319535.3363274

[60] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I. Hong. 2018. "Hey Alexa, What's Up?": A Mixed-Methods Studies of In-Home Conversational Agent Usage. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 857–868. https://doi.org/10.1145/3196709.3196772

[61] Guogen Shan and Shawn Gerstenberger. 2017. Fisher's exact approach for post hoc analysis of a chi-squared test. *PloS one* 12, 12 (2017).

[62] Faysal Hossain Shezan, Hang Hu, Gang Wang, and Yuan Tian. 2020. VerHealth: Vetting Medical Voice Applications through Policy Enforcement. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–21.

[63] Faysal Hossain Shezan, Hang Hu, Jiamin Wang, Gang Wang, and Yuan Tian. 2020. Read between the lines: An empirical measurement of sensitive applications of voice personal assistant systems. In *Proceedings of The Web Conference 2020*. 1006–1017.

[64] Dan Su, Jiqiang Liu, Sencun Zhu, Xiaoyang Wang, and Wei Wang. 2020. "Are you home alone?" "Yes" Disclosing Security and Privacy Vulnerabilities in Alexa Skills. arXiv:2010.10788 [cs.CR]

[65] Nan Zhang, Xianghang Mi, Xuan Feng, XiaoFeng Wang, Yuan Tian, and Feng Qian. 2019. Dangerous skills: Understanding and mitigating security risks of voice-controlled third-party functions on virtual personal assistant systems. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1381–1396.

[66] Yangyong Zhang, Lei Xu, Abner Mendoza, Guangliang Yang, Phakpoom Chinprutthiwong, and Guofei Gu. 2019. Life after speech recognition: Fuzzing semantic misinterpretation for voice assistant applications. In *Proc. of the Network and Distributed System Security Symposium (NDSS'19)*.

A APPENDIX

The survey is available through the following link:
https://ncsu.qualtrics.com/jfe/form/SV$_6$rmzhbNcIsl2F2S