Alexa, Who Am I Speaking To?: Understanding Users' Ability to Identify Third-Party Apps on Amazon Alexa

Project Website: https://sites.google.com/view/alexawhoamispeakingto/

DAVID MAJOR, Princeton University
DANNY YUXING HUANG, New York University
MARSHINI CHETTY and NICK FEAMSTER, University of Chicago

11

Many Internet of Things devices have voice user interfaces. One of the most popular voice user interfaces is Amazon's Alexa, which supports more than 50,000 third-party applications ("skills"). We study how Alexa's integration of these skills may confuse users. Our survey of 237 participants found that users do not understand that skills are often operated by third parties, that they often confuse third-party skills with native Alexa functions, and that they are unaware of the functions that the native Alexa system supports. Surprisingly, users who interact with Alexa more frequently are more likely to conclude that a third-party skill is a native Alexa function. The potential for misunderstanding creates new security and privacy risks: attackers can develop third-party skills that operate without users' knowledge or masquerade as native Alexa functions. To mitigate this threat, we make design recommendations to help users better distinguish native functionality and third-party skills, including audio and visual indicators of native and third-party contexts, as well as a consistent design standard to help users learn what functions are and are not possible on Alexa.

CCS Concepts: • **Human-centered computing** → **User studies**; • **Security and privacy** → **Usability in security and privacy;**

Additional Key Words and Phrases: Smart home, Internet of Things, network measurement, security, privacy

ACM Reference format:
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: Project Website: https://sites.google.com/view/alexawhoamispeakingto/. *ACM Trans. Internet Technol.* 22, 1, Article 11 (September 2021), 22 pages.
https://doi.org/10.1145/3446389

1 INTRODUCTION

Voice user interfaces (VUIs) are becoming more common as **Internet of Things (IoT)** devices proliferate. According to recent studies, 26% of U.S. adults own smart speakers, 61% of whom own

This work was partially supported by NSF awards CNS-1953740, CNS-1237265, and CNS-1518921, along with industry funding from Cable Labs (including in-kind donation of equipment plus funding), Amazon, Microsoft, Cisco, and Comcast.
Authors' addresses: D. Major, Princeton University, 35 Olden St, Princeton, New Jersey, 08540, USA; email: djmajor@princeton.edu; D. Y. Huang, New York University, 370 Jay St, Brooklyn, New York, 11201, USA; email: dhuang@nyu.edu; M. Chetty and N. Feamster, University of Chicago, 5730 S. Ellis Avenue, Chicago, Illinois, 60637, USA; emails: {marshini, feamster}@ uchicago.edu.

1533-5399/2021/09-ART11 $15.00
https://doi.org/10.1145/3446389

Amazon Echo (and related devices, e.g., the Echo Dot) and 24% of whom own Google Home [1]. Both Amazon's and Google's voice assistants allow users to use only voice to interact with the device's wide range of built-in and third-party functions. On Amazon Echo, the assistant is called *Alexa*, and these functions are known as "Alexa skills," analogous to applications or "apps" on mobile devices. Supported skills include setting an alarm clock, telling jokes, and sending money.

Although some skills and other *native* functionality ship built into Alexa, more than 50,000 skills are developed by third parties [2]. Users can invoke third-party skills to add functions to their Echo devices [3]. For reasons of both security and privacy, users need to be able to differentiate between native Alexa functionality and third-party skills. On conventional computing devices such as a phone with a **graphical user interface (GUI)**, it is relatively intuitive for users to differentiate between a device's native functionality (e.g., a homepage or settings screen on a phone) and apps they have actively downloaded from the Internet that are developed by a third party. The distinctive "look and feel" of native system functions provides a certain level of protection: if any third-party app could mimic a phone's homepage and native apps (e.g., calling, settings, or even a digital wallet), the security of any private information stored on that system would be at risk. In contrast to users of devices with GUIs, it is unclear if it is equally intuitive for Alexa users to differentiate between native functionality and third-party apps, especially considering that its VUI generally uses one voice for all interactions. This is critical to the adoption of VUIs as a whole, as users are specifically concerned about privacy and data collection with regard to third parties [4, 5]. Recent work has shown that users often hold incorrect mental models of the Alexa system generally [5, 6] and specifically with regard to third-party skills [7], further motivating a comprehensive study of how users perceive third-party skills and differentiate them from native functionality. Amazon created certain protections (e.g., a multicolored light on top of the Echo device) that can help users identify some native functions (e.g., the light flashing orange during setup), but the effectiveness of these methods is unknown. Furthermore, although past studies have explored users' mental models vis-a-vis third parties [7], no comprehensive study to date has addressed whether users can differentiate between native functionality and third-party skills even when they have accurate mental models of the Alexa ecosystem.

In this work, we seek to understand how the design decisions in the Amazon Alexa VUI affect users' ability to determine whether they are interacting with native functionality or with arbitrary third parties. In our hypothesis, users of Alexa's current VUI are unable to distinguish between functionality built into Alexa and functionality enabled by third parties—which highlights a human-computer interaction issue. To test this hypothesis, we conducted a survey from March to May 2019 with 237 new and existing users of Alexa, including 103 U.S. university students and 134 Amazon **Mechanical Turk (MTurk)** workers. Our work is distinct from previous work on VUIs that has focused more generally on how users interact with VUIs [8–13], and their ability to distinguish between third-party skills [14, 15]. In contrast, our work focuses on users' ability to distinguish third-party skills *from native functionality* (rather than distinguishing between third-party skills as in the previous work). Furthermore, our work also focuses on understanding users' mental models of VUI skills specifically—that is, learning *why* users are able or unable to distinguish third-party skills from native functionality. To study these questions, we presented participants with video and audio clips of interactions with three types of skills in the lab—(1) Alexa's native skills and functionality, (2) publicly available third-party skills, and (3) malicious skills that we developed and which are not publicly available—without revealing to the participant which type of skill is being presented. We tested whether participants could differentiate between the three categories and asked about their general impressions of Alexa and its third parties, both before and after presenting the videos.

This article's main contribution is the illustration of a fundamental conflict between providing users with a seamless VUI experience and simultaneously indicating to users with which third parties they are communicating. Our results are as follows.

First, we quantitatively and qualitatively found that many participants—regardless of their demographic background—did not know that skills could be developed by third parties at all, nor that interactions with Alexa can be directly with third parties (when the skill is created by a third-party developer). This finding contrasts against previous qualitative work [7] with a smaller participant population ($n = 17$). Furthermore, we identified a novel finding where Alexa owners were collectively *more* likely to hold this incorrect mental model of Alexa.

Second, we discovered that Alexa users, even when educated as to the existence of third-party skills, generally cannot differentiate native skills from third-party skills. Much to our surprise, we found that users who have *more* familiarity and experience with Alexa are in fact more likely to mistakenly assume that a third-party skill is in fact native Amazon functionality. Alexa users could also not differentiate between real and fake native system messages, including a fake message that prompted the user to execute a hard reset of the device. Participants' responses indicated that the characteristics of Alexa they primarily used to differentiate between native/third-party and real/fake responses was how the response sounded and whether the functionality made sense, suggesting that Amazon's current measures to differentiate some native functionality from third-party skills are not effective.

Third, we found that many participants did not understand what functionality could be executed on Alexa verbally (some system commands, like setting up WiFi, can only be done physically with buttons or through an accompanying app) nor that nearly any phrase can invoke a third-party skill on Alexa (a majority of participants incorrectly thought benign phrases like "Alexa, please go away" could *not* invoke a third-party skill). This set of results, given the previous two, suggests that a malicious third-party skill could trick a user by imitating a native function (e.g., enabling parental controls) that, although the user might think exists, actually does not.

Throughout the article, we tie each of these sets of results to three key design principles we find lacking in the design of Alexa and, by extension, VUIs generally: conceptual model, feedback, and discoverability [16]. We make two recommendations to better incorporate these principles into Alexa's design and thereby help users differentiate between native functionality and third-party skills.

In summary, although previous work has explored various aspects of VUIs and voice assistants (including users' mental models of VUI ecosystems and third-party skills mimicking one another), our research is the first to illustrate that VUIs present new challenges for differentiating between native and third-party functionality. Malicious third-party applications are well studied in the context of the web, less well understood in the context of IoT, and even less well understood in the context of VUIs. Our research motivates an important future design challenge for VUIs: how do we design VUI ecosystems that clearly differentiate native functionality from third-party applications without disrupting the natural, conversational experience?

Section 2 overviews the Alexa skill ecosystem, third-party skills, and certain protections Alexa has in place to help users differentiate native functionality from third-party skills. Section 3 reviews related work in designing VUIs and other security issues with regard to malicious third-party skills on Alexa. In Section 4, we review the survey structure and participant characteristics. In Section 5, we present the results of the survey, grouped into the preceding three major findings. In Section 6, based on our findings, we present two recommendations to help users differentiate between native functionality and third-party skills while still giving users a seamless user experience. In Section 7, we discuss limitations and future work.

2 BACKGROUND: ALEXA SKILLS

Recent years have seen a proliferation of voice-enabled IoT devices, ranging from phones to voice assistants to microwaves to car navigation systems. This article focuses on one specific type of voice-enabled device that can host third-party applications. In this sector, Amazon is the dominant player with 61% market share across its Alexa-enabled Echo devices (Google has the second highest with 17%) [17]. To further spread Alexa, Amazon has built the Alexa Voice Service, which can configure other smart devices (not made by Amazon) to run Alexa software [18]. Thus, Alexa can be seen as the clear leader in the field and a useful case study for understanding how users interact with VUIs for virtual assistants. We provide an overview of Alexa's skill ecosystem and a description on how users invoke and interact with skills.

2.1 Native and Third Party

Alexa supports two types of skills: (1) native skills and (2) third-party skills. Native skills come built in by Amazon and thus only involve code and dialog developed by Amazon. For example, users can ask for the time, set an alarm, or play music from Amazon Music. As Amazon is the sole developer for these skills, we assume that all information collected from users flows only to Amazon.

To support a broader range of functions, Amazon allows third-party developers to build skills for Alexa using the Alexa Skills Kit. Using the skills kit, developers can configure Alexa to communicate with their own services, create custom Alexa responses, and run custom code on their own servers [19]. Third-party developers have built at least 47,000 skills, including a wide variety of functions such as playing podcasts, locking doors, checking credit card balances, and telling jokes, which are publicly available on the Amazon Skill Store [3]. Since the code of these skills could be on third-party servers, we assume that some of the information collected from users may flow to the third-party developers (in addition to Amazon).

2.2 Invoking Skills

Whether a skill is native or third party, a user can invoke (i.e., verbally enable) it by saying the corresponding *invocation phrases*. These phrases follow the form of "Open <invocation name> for <optional action>" where the invocation name is often the name of the skill. Examples include "Alexa, open Jeopardy" (i.e., a game shown in the United States) and "Alexa, ask Daily Horoscopes about Taurus."

However, Alexa allows some flexibility in invoking skills. For some native skills such as the alarm clock, a user can invoke it via either "Alexa, set an alarm for 8 am" or "Alexa, wake me up at 8 am." For third-party skills, users replace "Open" with one of 13 words such as "play" and "launch." If none of these phrases are present, Alexa automatically parses the user's statement for an invocation name and responds with the corresponding skill [20]. However, invocation names do not appear to be unique, as we have found skills with the same invocation names. It is unclear how Alexa chooses which skill to invoke given two skills with the same invocation name.

2.3 Interacting with Skills

Once a user invokes a skill, Alexa enters what we call the skill's *context*. At the time of writing, Alexa does not verbally confirm which third-party context a user is in; in fact, Alexa's voice sounds exactly the same. Once Alexa is in a skill's context, Alexa accepts only voice commands predefined by that skill, along with phrases such as "cancel" or "quit" that allow users to leave the skill's context and invoke a different skill. A user cannot invoke a second skill until the user leaves the first skill's context.

2.4 Built-in Signifiers to Differentiate between Native and Third-Party Contexts

Amazon has developed several methods that might inhibit a third-party skill's ability to mimic native functionality. First, Amazon Echo devices have a built-in light that flashes when users are using the device. The light can flash at least six different colors to indicate certain contexts—for example: blue in third-party skills (and some native skills), red when the microphone is off or a specific error message plays, and orange when the device is in setup mode [21]. Third-party skills cannot control the light—it exclusively flashes blue during third-party skill execution—making it more difficult for third-party skills to mimic native functionality in which another color light flashes. Later, we discuss the success of this method, and whether it should be extended, based on our results.

Second, certain invocation phrases are reserved for native Alexa functionality, ensuring that third-party skills cannot be built to be invoked by these phrases (and then pretend to be the system executing these functionalities). For example, when one says "Alexa, Wi-Fi"to Alexa, the device responds with Wi-Fi connection information, suggesting that such a phrase is reserved for the system.

Third, Alexa's developer guide states the following in its skill naming requirements: "The invocation name must not create confusion with existing Alexa features" [22]. Rejecting to publish skills that violate this rule could prevent skills that might attempt to imitate native functionality. However, it is unclear if Amazon would reject a skill in practice. The developer guide seems to leave the possibility open of a third-party skill responding to an invocation phrase for native functionality; in the case of the weather skill, for example, the guide states: "If your invocation name is too similar to the built-in "weather" command, Alexa may sometimes respond with your skill and sometimes respond with the built-in weather feature" [22]. In addition, some currently published skills respond to phrases that could be meant for the native system. For example, *Home Wi-Fi* can be invoked by the phrase "Alexa, ask home Wi-Fi what's the wireless password?" [23].

3 RELATED WORK

There is a large body of work on security and privacy attacks on voice assistants, ranging from inaudible voice commands [24] to inferring Alexa activities based on encrypted network traffic [25]. Many of these attacks are possible due to the attackers' *technical* capabilities. This article, in contrast, focuses on the *human* issues; we seek to understand the human-computer interaction problem where the design of Alexa's VUI appears to affect users' ability to distinguish between native and third-party capabilities.

In this section, we focus on this human problem. We provide a literature review of user interfaces, focusing on differentiating third parties on GUIs, how to design VUIs, how humans interact with VUIs, and security/privacy concerns regarding VUIs.

3.1 Differentiating Native Functionality and Third Parties on GUIs

Although less common, there are specific cases where manufacturers of GUIs also had to take precautions to stop third-party applications from mimicking the native system. On Mozilla Firefox, for example, developers are limited in automatically making pages go full screen and users cannot type input while in full screen mode in order to prevent phishing attempts where a full screen Internet page might mimic a browser with the URL of a different website (and thereby steal passwords or other private data) [26]. Google Chrome similarly gives users a notification (that cannot be disabled) when going into full screen mode. Similarly for iPhones, phishing attacks have been devised whereby a fake popup within a third-party app, mimicking a nearly identical native popup, asks for private information [27]. In such a case, however, the user can always press the home

button and return to the home screen, limiting the ability for a third-party app to mimic the native system.

3.2 Designing VUIs

We consider Alexa as a conversational agent with a voice-only interface. Design for conversational agents can be dated back to Weizenbaum [28]; similarly, design patterns for GUIs are a well-established field [29]. However, paradigms for VUI design are scarce to our knowledge, presumably because voice assistants and other voice-enabled technologies have only taken off in recent years. One example of literature in VUI designs is Cathy Pearl's *Designing Voice User Interfaces: Principles of Conversational Experiences* [30], which covers design principles such as always confirming users' voice input or handling ambiguous voice commands. However, the authors assume that only the first party (i.e., the device manufacturer) engages in conversation with users without considering third-party capabilities such as skills. Similarly, López et al. [31] evaluated the usability of popular VUIs such as Alexa and Apple Siri in terms of correctness of responses and how natural the responses sound; again, this work did not consider third-party functionalities. In fact, we are unaware of any literature in VUI design that incorporates third-party apps, and we are among the first to discuss third-party-enabled VUI design in the research community.

Despite the apparent lack of literature, there are general design principles that could apply to our case. Don Norman's *The Design of Everyday Things* [16] introduces seven fundamental principles of design, three of which are especially relevant to this study of VUIs: (1) discoverability, which, when applied to skills, suggests that Alexa should let users know what voice commands are available and can invoke third-party skills; (2) feedback, which suggests that Alexa should inform users of which skills they are currently interacting with; and (3) the conceptual model, which would require Alexa to help users understand that skills are developed by third parties. As we will show in the survey results, Alexa's design appears inconsistent with these principles, exposing users to security and privacy risks. We leave for future work to evaluate Alexa's design against the remaining four design principles: affordances, signifiers, mappings, and constraints.

3.3 Human Interactions with VUIs

A large body of work studies how humans interact with VUIs and what kind of relationship is developed as a result. For instance, researchers found that some users personified their VUI devices and treated the devices with emotion as if the devices were family members or friends [8–10]. Past work has also found that interactions with VUIs were integrated with activities or conversations involving the entire household, including children [11, 12]. However, some researchers identified privacy concerns for VUIs in the public space, resulting in greater user caution when users transmitted sensitive information than non-sensitive information [13]. In this article, we also study how users interact with a VUI (i.e., Alexa), but we specifically focus on how users could be confused by Alexa's design vis-a-vis third parties and how users might leak sensitive information due to this confusion.

3.4 Security and Privacy Risks

Users face multiple security and privacy risks that originate from a number of actors. First, manufacturers of voice assistants (i.e., the first parties) may collect potentially sensitive recordings of users without the users' knowledge, such as through the always-on microphones. This design may lead to accidentally recording sensitive conversations and sharing the data with manufacturers [6, 32]. In addition to manufacturers, third-party skills (or "actions" on Google Home) could also present security and privacy risks to users. In particular, a third-party malicious skill could

effectively phish a user by pretending to be another benign skill. As demonstrated in a proof of concept by Kumar et al. [33] and Zhang et al. [34], a malicious skill could use an invocation name that sounds similar to a benign skill, such as "Capital One" (legitimate banking skill) and "Capital Won" (malicious skill).

This work is different from that of Kumar et al. [33] and Zhang et al. [34], who focus exclusively on third-party skills mimicking other third-party skills. In contrast, this work is focused specifically on the problem of third-party skills mimicking the native system—a much more pernicious problem. Since all developers have the same toolkits available, it makes sense that any developer can design one skill to mimic another (this is true for third parties on all platforms: anyone can make a website homepage identical to Google's). However, developers *should not* be able to design skills that mimic native system functionality, which this work demonstrates is possible (and is largely outside the scope of Kumar et al. [33] and Zhang et al. [34]).

In addition, both of those works showed that skills have the *potential* to deceive users by mimicking each other and that users displayed behaviors that could be exploited by such skills. Our research is the first to empirically demonstrate that users are *actually* deceived by their inability to distinguish third-party skills from native functions (see Section 5). The jump from possible (past work) to actual (this work) deception is an important result and a critical step in the context of consumer protection, where enforcement relies on demonstration of actual (not just hypothetical) deception.

We do believe that some of our quantitative findings (Section 5) offer important insights absent in previous work. For example, much to our surprise, Alexa owners and frequent users are more likely to have an incorrect mental model of third-party skills (e.g., see Figure 2) and are often worse at distinguishing third-party skills than people who were less familiar with Alexa and non-owners (e.g., see Figure 5), perhaps even suggesting that familiarity with the device introduces additional risks.

3.5 Perceptions, Attitudes, and Mental Models

In the face of multiple security and privacy threats, much attention has been given to users' various perceptions, attitudes, and mental models toward voice assistants and third-party skills. In an earlier work based on analyzing product reviews online, consumers already expressed concern about how voice assistants—being always on—collected data and the scope of this data collection [4]. A in-depth diary-based study reveals that some of these concerns arose because users lacked an understanding of how voice assistants collected data, and that few users actually review the audio logs [5]. This finding is also echoed in another work by Malkin et al. [6], which shows that real-world users of voice assistants (as opposed to lab users) were unaware that their recordings were permanently stored and could be reviewed and/or deleted.

In a highly related effort, Abdi et al. [7] explored users' mental models with regard to third-party skills. They found that participants rarely considered third-party skills as relevant agents in the VUI ecosystem that could process and store user data or present a threat to user privacy. Our work builds on Abdi et al. [7] to show that incorrect mental models are only part of the problem vis-a-vis third-party skills and user privacy. Although we confirm their finding that users are often unaware skills can be developed by third parties (Section 5.1), we show that even when aware of this fact, users *still* cannot differentiate between native and third-party functionality (Section 5.2). Thus, "awareness and transparency mechanisms" to improve users' mental models, as Abdi et al. [7] rightfully suggest, must be complemented with design changes to enable users to differentiate between native functionality and third-party skills. We make specific recommendations toward these goals in Section 6.

Additionally, users have expressed various perceptions and attitudes toward third-party skills. For example, Tabassum et al. [35] explore how users expect voice assistants to behave, what functionalities they expect voice assistants to demonstrate, and whether the users are comfortable sharing sensitive data with these functionalities. We find that these inherent expectations may lead users to not understand that voice commands can invoke skills and what can be done with Alexa verbally (Section 5.3). This problem is compounded by the design of the voice interface of Alexa (i.e., having the same voice for native and third-party functionalities); this design, as discussed in a review work by Edu et al. [36], might bring confusion to users as to which functionality—native or third party—the users are talking to and thus create opportunities for attacks (e.g., Voice Masquerading). In contrast to these previous works that also discussed the *potential* for users to get confused and face attacks, our work *actually* shows that our participants were confused by real attack skills (albeit in a lab environment to limit the damage to real-world users).

4 SURVEY METHOD

To understand how users conceptualize and interact with Alexa and its skills, we conducted surveys of both Alexa owners and non-owners in two populations: 103 undergraduate and graduate university students ("University survey") and 134 participants through Amazon MTurk ("MTurk survey"). Having both surveys enabled us to survey a wide swath of participants [37]. We tested both owners and non-owners to better understand whether previously owning or using an Alexa affected a participant's familiarity with the device and how skills operate. Both surveys were approved by our university's Institutional Review Board.

4.1 Recruitment

Recruitment method and differences between populations. We conducted the University survey between March 27 and April 10, 2019. We recruited 103 U.S. university student participants by email through our university's **survey research center (SRC)**. The SRC randomly selected students and emailed them a link to the survey hosted on Qualtrics. We incentivized participation by awarding approximately 1 in 10 participants with an Amazon Echo device. We did not require Alexa ownership or experience in the recruitment criteria, although participants who decided to take the survey were presumably aware of or interested in Alexa.

Based on our initial findings, we expanded the survey through an Amazon MTurk survey between April 19 and May 9, 2019. Through MTurk, we recruited 134 English-speaking participants with at least a bachelor's degree. Participants were paid at least minimum wage for the 10-minute survey. To ensure quality responses, we shared the survey only with MTurk users with approval ratings over 95% and who were MTurk "Masters," a special designation Amazon gives only to the top-performing MTurk users. Again, Alexa ownership or experience was not a part of the recruitment criteria. As discussed in the following, the quality of the MTurk participants was ensured through multiple attention checks, prevention of re-tries, and a time limit.

The reason for the two populations was to supplement the (original) University survey to make the total sample population more diverse. We conducted chi-squared tests on the participants' responses question across various demographic groups (including the survey group) and did not find evidence that most answers were independent of the survey group. We note wherever answers were independent of survey group throughout the article.

Removing participants. To exclude low-quality responses and their participants, we added three attention checks and tests that participants understood the core ideas throughout Survey Sections 3 and 4 in random locations and orders (after we described key concepts like native/third-party skills in Survey Section 2), such as "what is the website of Amazon," "who builds native Alexa skills," and

"who builds third-party Alexa skills," with "amazon.com," "Amazon," and "non-Amazon" as the choices. These attention check questions appear in random order and locations, and they could potentially reduce the likelihood of "survey straight-lining" (i.e., respondents entering random answers). In doing so, we hoped to remove participants who failed to understand the concept of third-party skills. In a further effort to filter out low-quality participants, we also prevented re-try attempts and added a time limit to the survey.

Characteristics of participants. Of the participants reached through our university's SRC, all were current undergraduate or graduate students. Although participants came from a wide range of specific majors, about 40.4% were in an engineering-related subject, and 51.0% identified as male, 47.0% as female, and 2.0% as other. In addition, 19.6% of participants own an Alexa. We did not ask how often the participants used their Alexa devices.

For the MTurk survey, all participants have a bachelor's degree (as verified by MTurk), a 95% approval rating or higher, and MTurk "Masters" status. In addition, 51.9% identified as male and 48.1% as female, and 75.2% of participants own an Alexa. Of those who own an Alexa, 34.0% have owned it for less than a year, 40.0% for 1 to 2 years, and 26.0% for more than 2 years. A majority (56.1%) use it several times a day, whereas 9.2% use it less than once a week, 7.1% once a day, and 27.6% several times a week.

Limitations. Our survey respondents include 103 college students (40.4% of whom are in engineering-related majors) and 135 MTurk users (at least 95% approval ratings), with an approximate half-half split in gender. As we will show in Figure 1, both demographic groups include owners and non-owners of Alexa, as well as frequent and non-frequent users of Alexa. Although we did not collect demographic details such as age or occupation from the respondents, we assume that our overall subject population is likely more tech-savvy than the general population. This is likely necessary; our survey includes an education section after which participants must understand the difference between native functionality and third-party skills. A more tech-savvy population is more likely to understand these concepts. For example, 86% of MTurk participants owned either an Alexa or some other smart device (with a smart device being defined as any Internet-connected device other than a computer or phone). Nevertheless, we believe that our results are generally applicable; if anything, the general population that is less tech-savvy would highlight a general lack of awareness about Alexa skills than our survey results already do with the tech-savvy population.

4.2 Survey Questions

We designed a Qualtrics survey to achieve the following goals:

(1) to test whether participants are originally aware that Alexa skills can be built by third parties before we alert them to this fact (Survey Section 1);
(2) to help participants learn about the existence of skills built by third parties to enable the next set of questions (Survey Section 2);
(3) to test whether participants can distinguish between native functionality and third-party skills now that they are aware these third-party skills exist (Survey Section 3); and
(4) to test participants' beliefs as to what verbal commands can invoke third-party skills and what actions can be executed verbally on Alexa, which outlines some possible ways a malicious skill can try to trick users (Survey Section 4).

4.2.1 Survey Section 1: Pre-definition Questions about Alexa. Through this section, we aimed to understand users' privacy expectations of skills *before* we defined the terms *native*/*third-party*

skills for them. Per the design principles of Norman [16], our goal was to check whether users' conceptual model of third-party skills was consistent with reality.

To this end, we first asked a general true/false question, "Everything Alexa says is programmed by Amazon," and we counted the number of respondents with each answer "yes" or "no" (correct answer).

Additionally, we produced and presented the following videos in order, in which a member of the research team engaged in a conversation with Alexa in the lab. We then asked users where they thought the data from the conversation was sent: to "Only Amazon," "Only Third Parties," or "Both." Interested readers can view all of our videos (including those in later survey sections) at our anonymized project webpage: https://sites.google.com/view/alexawhoamispeakingto/:

- *Video 1A: Add Rubber Ball to Shopping Cart*: A user (i.e., a member of the research team) asks Alexa to add a rubber ball to his cart, and Alexa responds, "Ok, I've added a choice for rubber ball to David's Amazon Cart." This is an actual interaction that occurs when a user asks Alexa to add an item to the cart. In the survey, we asked each participant, "Immediately as result of the following conversation, what parties do you think know David added a rubber ball to his Amazon cart?" The correct response is "Only Amazon."
- *Video 1B: Bedtime Story*: The following conversation occurs:
  User: "Alexa, open 'I'm going to bed."'
  Alexa: "Time for a bedtime story! First, what's your name?"
  User: "Benji."
  Alexa: "Ok, Benji! Here's your story."
  This skill is an example third-party bedtime story skill we built (which we did not publicly release). In the survey, we asked each participant, "Immediately as result of the following conversation, what parties do you think know your name is 'Benji'?" The correct response is "Both."

After showing Video 1B, we asked each of the participants to provide a free-text open-ended response to explain their rationale for their answer.

4.2.2 Survey Section 2: Defining Key Alexa Concepts/Terms. In this survey section, we briefly described to participants what an Alexa skill is and what native and third-party skills mean. The goal was to ensure, to our best effort, that the participants understood these concepts in later sections, as we would test whether participants could distinguish between native and third-party skills and capabilities. Later attention checks tested whether participants did indeed understand these concepts; participants who failed a single test were excluded.

4.2.3 Survey Section 3: Differentiating Native and Third-Party Skills. To test whether participants could differentiate between native third-party skills—effectively whether Alexa was able to offer Feedback (per the design principles of Norman [16]) on which skills a user was interacting with—we embedded five video clips and five audio clips in this section of the Qualtrics survey and asked for the participants' response. Similar to Videos 1A and 1B, we produced the following clips ourselves and presented them to the participants in order.

The video clips show a member of the research team interacting with a native or third-party skill. After we showed each clip, we asked participants whether the participant had interacted with a native or third-party skill:

- *Video 3A: Tell a Joke (native)*: A user asks Alexa for a joke and Alexa responds with a joke.
- *Video 3B: Jeopardy (third party)*: A user asks Alexa to play Jeopardy (a U.S. game show) and the game begins with the voice of Alex Trebek (Jeopardy's host).

- *Video 3C: Baseball Scores (third party)*: A user asks Alexa about the Astros (a baseball team) and Alexa responds with the latest scores.
- *Video 3D: Rain Sounds (third party)*: A user asks Alexa to play rain sounds and Alexa responds with the sound.
- *Video 3E: Parental Controls (third party)*: A user asks Alexa to enable parental controls and Alexa responds confirming the user would like to do that. Although Videos 3B through 3D feature real skills available on the Amazon Skill Store, the parental control skill in this video is not public; in fact, we developed this skill ourselves using the Alexa Skill Kit and made it available only to the Amazon Echo in our lab. We designed this skill to sound as if it could configure parental controls on Alexa, although in reality parental controls cannot be configured verbally with Alexa.

We also showed audio messages that we recorded from a native skill or third-party skill. The third-party skill could be from the Skill Store [3], or it could be developed by us and not released publicly. We asked each participant to respond whether the message was a real system message (i.e., native skill) or a fake one (i.e., third-party skill masquerading as a native skill):

- *Audio 3A: Wi-Fi (fake)*: "You seem to be disconnected from Wi-Fi. Please hold down the circle button in order to reconnect." Similar to the skill in Video 3E, we developed a private skill that hard coded the preceding message. Following the instructions would initiate a hard reset of the device.
- *Audio 3B: Problem with Response (real)*: "There was a problem with the requested skill's response." Alexa generates this verbal message when a third-party skill's response is not configured correctly.
- *Audio 3C: Link (fake)*: "Sorry, something is wrong with this device. Please restart or go to amazon.com/alexa for more information." Again, we developed this private skill ourselves. A malicious third-party skill could say this message (e.g., in the middle of other activities of the skill, thus giving the illusion that this is a system-generated message), replacing the URL with that of a phishing website.
- *Audio 3D: Sorry (real)*: "Sorry, I'm not sure about that." Alexa generates this message when it cannot understand the user's voice commands.
- *Audio 3E: Amazon Account (fake)*: "Sorry, before using this device you need to connect your Amazon account." We developed this private skill ourselves.

4.2.4 Survey Section 4: Voice Commands That Alexa Understands. Finally, we aimed to test whether Alexa offers users discoverability, per Norman's design principles [16], or whether users know what voice commands can be understood by Alexa.

In particular, we asked participants whether the following invocation phrases could open skills on Alexa: "Open Grubhub," "What's the NY Times report," "Find my iPhone," "Quit," "Please go away," and "There's a bug over there." With the exception of "Quit" (which lets users leave a particular skill), all of these phrases can open actual Alexa skills on the Skill Store or those we developed in private (e.g., "Please go away" and "There's a bug over there.").

We also asked whether certain actions can be accomplished with Alexa verbally: changing device volume, muting the device, checking the Wi-Fi connection, changing the Amazon password, ordering items on Amazon, turning off the device, and turning on/off parental controls. At the time of this writing, the only actions that Alexa can accomplish are changing the device volume and ordering Amazon items. These questions are relevant, as participants' expectations of what can be done on Alexa and what invokes third parties on Alexa can influence their ability to differentiate between native and third-party skills.

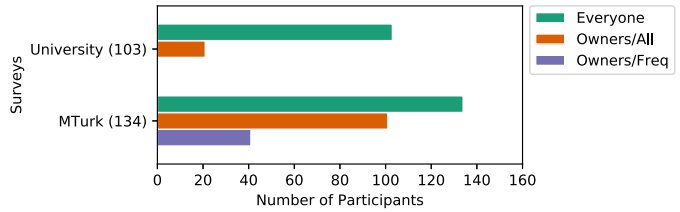

Fig. 1. Number of participants. The numbers in parentheses indicate the sample size.

4.3 Data Analysis

Preparing data for analysis. We downloaded survey responses from Qualtrics as CSV files and analyzed the data in Python Pandas. We removed three university participants and 48 MTurk participants for failing attention checks. Despite the relatively high ratings of the MTurk participants, one reason for their removal, as we suspect, could be that our study teaches participants about the existence of third-party skills and our attention checks test their understanding of this concept. It is possible participants skimmed this education section and then later failed attention checks. Another reason could be the low compensation we provide ($2.10 for the entire survey). The median time of completion for the surveys is 464 seconds (7.7 minutes).

Labeling participant groups. As we discuss in the Findings section, we analyzed the responses in terms of different levels of familiarity and experience with Alexa. To facilitate this analysis, we created three participant groups: (i) "Everyone," which refers to all 237 participants; "Owners/All," which is a subset of "Everyone" that refers to those who own Alexa devices, including 21 and 101 participants in the University and MTurk surveys, respectively; and "Owners/Freq," which is a subset of "Owners/All" that includes owners of Alexa who had owned the device for at least a year and indicated usage "multiple times a day or more." We used these labels to denote users with potentially increasing levels of familiarity with Alexa. Since we did not ask how often university participants used Alexa, all 41 "Owners/Freq" participants were from the MTurk survey. We provide a summary in Figure 1.

Coding free-text responses. For each free-text open-ended survey question, one member of the research team coded all responses using qualitative techniques [38]. Example codes tagged phenomena of interest related to a participant's mental model of Alexa—for instance, whether Amazon alone handled the interaction, or whether a third party was involved. Another member of the team then individually reviewed the codes, and we discussed final themes as a research team. For both free-text survey questions, the second team member was able to validate all codes/responses without disagreement.

5 RESULTS

Our survey results yield three major themes:

(1) Many participants were unaware that Alexa skills can be developed by third parties.
(2) Even when informed that Alexa skills can be developed by third parties, most participants could not differentiate between native functionality and third-party skills, nor between real and fake Alexa system messages. Interestingly, frequent users were even less able to distinguish native from third-party skills.
(3) Alexa users often do not understand the standards of how the Alexa system functions nor what is possible and not possible on Alexa.

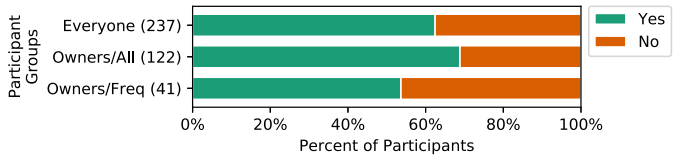

Fig. 2. Responses to the question, "Everything Alexa says is programmed by Amazon." Correct answer: 'No."

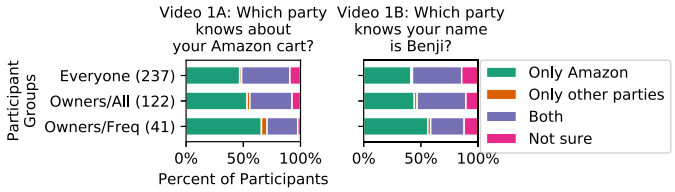

Fig. 3. Responses to two videos. Video 1A: What parties do you think know David added a rubber ball to his Amazon cart? Correct answer: "Only Amazon." Video 1B: What parties do you think know your name is Benji? Correct answer: "Both."

5.1 Finding 1: Participants Are Unaware That Skills Are Developed by Third Parties

Our results showed that participants' conceptual models [16] of who develops skills and who could see the users' data are inconsistent with the reality, where third parties can build skills and thereby have access to user behavioral data [39]. This finding contrasts against previous qualitative work with a smaller participant population ($n = 17$) [7]. We highlight novel insights where applicable.

5.1.1 Some Participants Assume All Alexa Contents/Capabilities Are Handled by the First Party. The participants' conceptual model of the device, particularly with regard to who builds skills, runs counter to reality. As shown in Figure 2, when asked whether "Everything Alexa says is programmed by Amazon" in Survey Section 1 (with "No" as the correct answer), 62.4% of all participants ("Everyone") thought the statement was true. In particular, 68.9% of "Owners/All" answered "Yes," which suggests that familiarity with Alexa may not always correspond to a more accurate conceptual model.

Furthermore, the participants' conceptual model appears independent of demographic groups—for instance, (a) whether the participant was recruited in the university or MTurk group, (b) whether the participant owns an Alexa device, and (c) whether the participant indicated being a frequent user of the device (i.e., interacting with Alexa "multiple times a day"). We conducted chi-squared tests on the participants' responses this question across groups (a), (b), and (c). The resultant p values are 0.0000186, 0.0496, and 0.000486, respectively, all less than 0.05. For example, for (a), the answers to "Everything Alexa says is programmed by Amazon" are independent of whether the participants were recruited from the university or MTurk, with $p = 0.0000186$. These p values suggest that the responses are independent of the demographic groups with statistical significance.

Our finding confirms previous results [7], which show that Alexa users are unaware of third-party skills, but across a larger sample size—that is, 237 participants in our study versus 17 participants in previous work [7]. In addition, we identify a novel insight, where familiarity with Alexa may not always correlate with a more accurate conceptual model.

5.1.2 Some Participants Were Unaware That Third-Party Skills Could Collect Data. In Survey Section 1, some participants were unaware that third parties could collect user data through Alexa

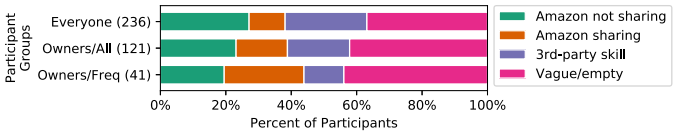

Fig. 4. User explanation (coded) for their responses to Video 1B.

skills. Figure 3 shows participants' responses to Videos 1A and 1B that were meant to gauge whether the participant understood that Amazon had third-party skills, and that the third parties had access to user responses. In particular, 46.8% of "Everyone" understood that only Amazon had access to the cart information (Video 1A), and this percentage increased as the level of familiarity and experience increased; in fact, some 65.9% of "Owners/Freq" answered correctly. For Video 1B, however, the more experienced participants were associated with a lower rate of correctness. For example, 56.1% of "Owners/Freq" incorrectly believed only Amazon knew the name was Benji, compared with only 41.4% of "Everyone." Again, although our finding is consistent with previous work [7], our novel insight is that familiarity with Alexa may not always correspond to a more accurate conceptual model.

In an open-ended free-text question after Video 1B, we asked why each participant answered in a certain way. Three themes emerged from the responses: (i) Amazon originally having the user data and subsequently sharing it with third parties ("Amazon sharing"), (ii) Amazon originally having the data but not sharing it with third parties ("Amazon not sharing"), and (iii) understanding that a third-party skill could directly have access to the data ("3rd-party skill"). For empty responses or vague responses, we used the code "Vague/empty."

We present a distribution of these codes in Figure 4, which suggests that only 25.0% of "Everyone" understood that skills had direct access to the data, rather than relying on Amazon to share the data. This percentage decreases as the level of familiarity and experience with Alexa increases; in fact, only 12.2% of "Owners/Freq" made the same choice. An example of a response showing this understanding included *S1R13*, *"Data is shared by the third party developer of the app,"* and another participant, *S2R8*, who wrote: *"I think it is a skill developed by another party, and they will have access to this data."*

For participants who were not aware of the skills, many believed that their interactions with Alexa were strictly with Amazon. Overall, 27.1% of "Everyone" were coded "Amazon not sharing." For instance, *S1R30* wrote: *"As far as I'm aware, Amazon doesn't sell any data to other companies, it only uses it privately (I could be wrong but I think this is true)."* Similarly, *S2R23* responded: *"Alexa is connected to Amazon and I think most info is stored and shared only with Amazon."* In contrast, 11.0% of "Everyone" believed Amazon did share data with third-parties (as opposed to skill developers having direct access to the data). For instance, *S1R5* wrote: *"There have been enough reports of information sharing across 'The Internet of Things' for me to presume that any information given to a smart device, especially one belonging to the Amazon company, is shared with other parties and services.,"* and *S2R24* wrote: *"I don't trust anyone to not sell or share data. They all do it."*

In summary, most of these responses (38.1% of "Everyone") centered around whether or not Amazon shared the data rather than the interaction being with a third party itself. Although participants' opinions on data sharing is irrelevant to this article, their responses shed light on their conceptual model of Alexa. For a majority of participants, an Alexa user interacts directly with Amazon alone, and only Amazon possesses data from the exchange as a direct consequence. This conceptual model contrasts greatly with the reality, where a skill can be built by any developer and anything a user says in such an interaction can go directly to the developer.

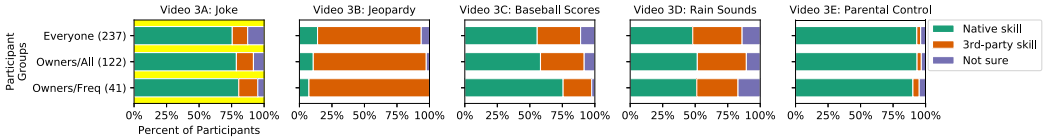

Fig. 5. Differentiating between native and third-party skills. Only Video 3A (highlighted) shows a native skill.

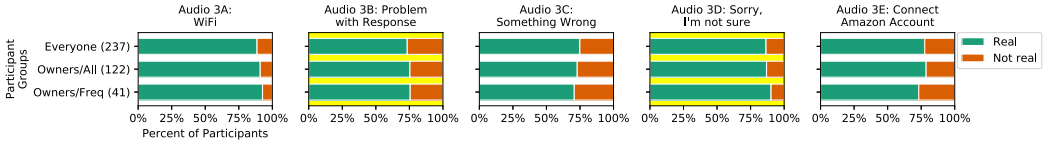

Fig. 6. Differentiating real system message from fake ones. Only the messages from Audios 3B and 3D are real (highlighted); the rest are fake.

5.2 Finding 2: Some Participants Cannot Differentiate between Native and Third-Party Skills and Messages

Even if users have an accurate conceptual model of Alexa with regard to its third-party skills, it is still crucial that they receive clear feedback [16] during conversations with Alexa that suggest whether that exchange was with a third party.

In this section, we show that the majority of our participants were unable to differentiate between native and third-party skills. A consequence of these results is that Alexa users, even if they have an accurate conceptual model of Alexa and its skills, might not get clear feedback from Alexa with regard to whether they have interacted with a third party. In fact, users might mistake third-party skills for native functionality, which can have serious ramifications for their privacy and security.

5.2.1 Differentiating between Native and Third-Party Skills. In Survey Section 3, we asked participants to watch Videos 3A through 3E of a person interacting with an Alexa device. After each video, we asked participants whether the person in the video interacted with a native or third-party Alexa skill. Only Video 3A referred to a native skill.

Figure 5 presents the participant responses. Although the majority of participants could correctly identify Videos 3A and 3B (intended as easier examples), the accuracy was much lower for the remaining videos. In the worst case, only 3.0% of "Everyone" and 4.9% of "Owners/Freq" could correctly identify "Parental control" (Video 3E) as a third-party skill. This result shows that a user could potentially confuse a third-party skill—whether malicious or not—with what appears to be native functionality; the user may accidentally leak sensitive information to the unintended third party.

Additionally, experience and familiarity with Alexa did not always correlate with more correct responses. In fact, whereas 33.3% of "Everyone" could correctly identify "Baseball Scores" (Video 3C) as a third-party skill, only 22.0% of "Owners/Freq" could do so. This result is in line with our previous findings for Video 1B (Figure 3).

5.2.2 Differentiating between Real (native) and Fake Messages (Which We Built). In Survey Section 3, we asked each participant to listen to Audios 3A through 3E. As shown in Figure 6, a majority of participants were unable to differentiate between real (i.e., as a result of native skills) and fake (i.e., as a result of third-party skills) Alexa system messages. For example, 88.6% of

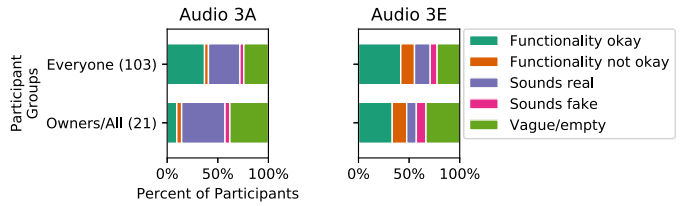

Fig. 7. User explanation (coded) for their responses to Audios 3A and 3E.

"Everyone" and 92.7% of "Owners/Freq" thought that the "Wi-Fi" message was real (Audio 3A). Again, familiarity of Alexa may not be correlated with a higher rate of correct responses.

Participant responses after Audios 3A and 3C are particularly troubling. In Audio 3A, following the message's instructions (holding down the circle button on an Echo device) would restart the system. In Audio 3C, the fake message prompts users to go to a website (in this case, just the Amazon website), creating potential for a phishing attack if the website is not Amazon.com. In both cases, we do not have enough evidence that a participant would actually perform some of the tasks suggested by the fake skills. However, the possibility that participants might accept system information verbally gives potentially malicious skills significant leeway in the types of attacks they might perform. For example, a fake skill could ask users for their Wi-Fi password or tell them their Alexa device is malfunctioning. A user's potential inability to differentiate between real and fake system messages helps enable voice squatting and masquerading attacks [33, 34], and such attacks could be expanded to better incorporate system messages (e.g., a voice masquerading skill could respond with an error message and then stay open).

After Audios 3A and 3E, we asked participants to briefly explain their answers in the University survey.[1] We grouped the responses into four categories: (i) functionality appearing to make sense ("Functionality okay"), (ii) functionality not making sense ("Functionality not okay"), (iii) audio sounding real or participant having heard it before ("Sounds real"), and (iv) audio sounding fake or participant never having heard it before ("Sounds fake"). We coded vague or empty responses as "Vague/empty."

We present the distribution of the codes in Figure 7. In both cases, "Functionality okay" and "Sounds real" dominate the reasons (ignoring vague/empty responses). In particular, 33.3% of "Everyone" thought Audio 3E's functionality made sense. For example, *S1R23* said the response was real because *"WiFi is necessary for Alexa function,"* and *S1R42* responded it was fake because *"I am not sure that Alexa has anything to do with Wifi."* These responses suggest that some participants made judgments on the authenticity of a message based on whether the message was consistent with Alexa functionality. The fact that participants made judgments based on functionality implies that a fake skill masquerading as the native system performing a reasonable task might seem believable; as we present later, participants did not have clear conceptions of what is reasonable on Alexa.

Furthermore, 9.6% of "Everyone" felt the clip sounded real or claimed to have heard it before. For example, *S1R8* responded: *"I've heard this one before,"* whereas *S1R2* said the exact opposite: *"I have not heard this previously."* Similarly, *S1R11* explained the response that the video was real based on Alexa's voice: *"It sounds official?"* These explanations suggest that participants made judgments based on the sound of a message. Although these judgments are reasonable for users of a VUI, they can confuse users when Alexa uses the same voice for all functionality. This can be seen in the examples of participants who insisted they had heard Audio 3A before, which is impossible given

[1]We did not ask for free-text response in the MTurk survey to reduce the survey burden.

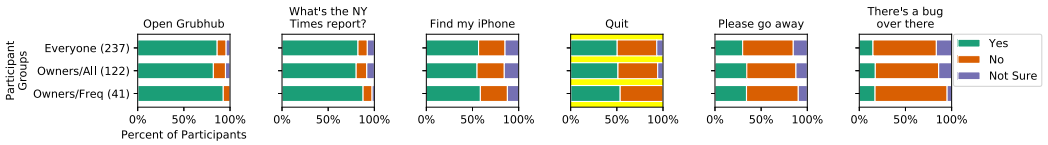

Fig. 8. Can each of the phrases above invoke an Alexa skill? In reality, all phrases, except "Quit" (highlighted), can invoke an actual skill.

that we faked the message. It is worth noting that no responses mentioned the light on top of the Echo device (even though it is visible in the videos), and rather focused entirely on Alexa's voice and the functionality. At least one Alexa error message displays a red light. These results suggest that participants may have not been paying close attention to the light color when determining if a message was real, and thus that lights may not be an effective means to differentiate between native functionality and third-party skills.

5.3 Finding 3: Some Participants Do Not Know What Voice Commands Can Invoke Skills

Given previous findings that many users cannot differentiate between native and third-party skills, it is crucial that discoverability be well incorporated into Alexa's design [16]. If users are unable to differentiate third-party skills from native functionality, they need a clear understanding of Alexa standards with regard to third-party skills so that a third-party skill cannot mimic native functionality. In this section, we present results that suggest that users do not have a clear understanding of what phrases can invoke third-party Alexa skills and what verbal functionality the Alexa system does and does not provide.

5.3.1 Users Do Not Understand What Phrases Can Invoke Third-Party Skills on Alexa. Many participants held incorrect assumptions regarding what phrases can invoke third-party Alexa skills. Although many participants believed that there were logical limits to what phrases can invoke an Alexa skill, in reality, nearly any phrase is enough (as long as it begins with the wake word "Alexa"). Although Amazon encourages developers to design skills with a few recommended invocation phrases (e.g., "Open <invocation name>" and "Ask <invocation name> <some action>"), Alexa is designed to, at a minimum, open skills by just their name [20]. Since this name can be arbitrary, the invocation phrase is unbounded, thus creating challenges for discoverability.

In Survey Section 4, we asked participants whether the six invocation phrases could open skills on Alexa. As shown in Figure 8, most participants understood that more conventional (based on Amazon's standards [20]) phrases like "Open Grubhub" and "What's the NY Times report?" can invoke skills on Alexa. In contrast, a majority of participants (54.8% and 68.6%, respectively) incorrectly responded that "Please go away" and "There's a bug over there" cannot invoke skills on Alexa. Even though, at the time of this writing, no skills on the Amazon Store respond to such invocation phrases, we successfully developed two private skills that could respond as such.

These results highlight a problem, especially given that users often cannot differentiate native and third-party skills (as shown in Finding 2). The fact that many users may not understand which phrases can successfully invoke third-party skills makes it even more likely they can accidentally invoke some skill and not realize it has been built by a third party. It may also increase the likelihood of invoking a malicious skill that can try to imitate the system or mimic another skill [33, 34]. A salient example of an attack that could leverage this result is the fake parental controls skill presented in Finding 2, which most participants believed was real and native. Even if a malicious actor is not involved, users could still accidentally invoke a third-party skill without realizing so and transmit sensitive information to unintended third parties.

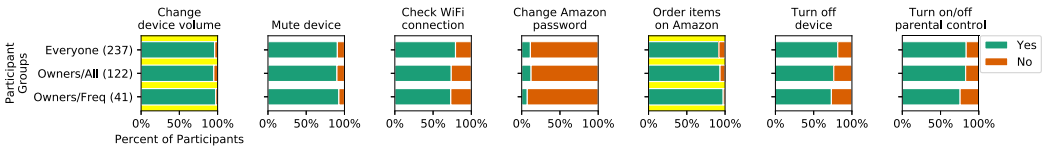

Fig. 9. Can you verbally do this with Alexa? In reality, users can only change the device volume and order items on Amazon (both highlighted) through verbal commands.

5.3.2 Some Participants Did Not Know What Can and Cannot Be Done with Alexa Verbally. Participants often did not have clear intuitions regarding what can and cannot be done with Alexa verbally (rather than through the app or with physical buttons on the device).

As shown in Figure 9, participants believed that most of the given tasks—except changing the Amazon password—could be done verbally with Alexa, further expanding the potential attack space for malicious skills. In reality, only changing volume and ordering Amazon goods are feasible through Alexa's voice interface, although 90.7% of "Everyone" thought they could verbally mute Alexa and 79.3% of "Everyone" believed they could check the status of Wi-Fi verbally. If a skill were to exist (whether malicious or benign) that responded to any of these invalid invocation phrases, a user may believe that he or she was interacting with the native system (especially given Finding 2) and potentially leak sensitive information.

It is worth pointing out that most (88.6%) of "Everyone" did not believe one can verbally change their Amazon.com password with Alexa, presumably because changing a password on non-verbal interfaces (e.g., on the Web) could be the conventional practice and doing so over the VUI may deviate from this standard. As such, there is potentially hope of raising awareness for users to understand what can (e.g., changing volume) and cannot (e.g., changing passwords) be achieved natively on Alexa; this awareness would likely help users distinguish some third-party skills and native skills and protect their privacy.

6 RECOMMENDATIONS FOR VUI DESIGN

Some of Alexa's design decisions are inconsistent with the design principles of Norman [16]: conceptual model, feedback, and discoverability. These inconsistencies likely led to the observations in our survey results. In this section, we propose design recommendations for Alexa—and VUIs in general—based on these principles and our findings.

6.1 Recommendation 1: Having Clear Indications of Contexts

Our results show that many participants were unable to distinguish between native and third-party skills (Finding 2), and this problem was compounded by the lack of awareness of third-party skills in the first place (Finding 1). These findings suggest that Alexa's design is inconsistent with the conceptual model and feedback principles.

Our recommendation is for Alexa to clearly indicate the context to its users through means other than those Amazon already takes (colored lights, reserved phrases, and other methods outlined Section 2.4). This approach would provide users with the correct conceptual model that there are differences between native and third-party skills and among the third-party skills. Moreover, the approach would offer feedback to users as to what context the interactions are in.

Past research in this realm has already yielded useful insights. To protect against voice masquerading attacks, for instance, Zhang et al. [15] proposed a "Skill Response Checker" that checks VUI responses for phrases that can be used to mimic the system. Amazon already implements a version of this approach by blacklisting some phrases (see Section 2.4). Although such features could be effective deterrents in some cases, our research suggests that users might believe a wide

array of messages (e.g., Audio 3A, 3C, and 3E) to be native system messages that would be difficult to blacklist individually. Furthermore, our research suggests that privacy concerns can arise even when skills are not trying to be malicious. Because users cannot always differentiate between native and third-party skills, it is possible that a third-party skill might request information that, although not inherently malicious, a user may not want to give.

One recommendation is for Alexa devices to play audio cues in place of limited visual cues that already exist. This could include using different voices for native and third-party skills or playing a chime as a user switches from one skill to another (similar to how Google Home plays an audio cue as it begins third-party functionality). The fact that 79.7% of participants responded that Jeopardy is a non-native skill (Figure 5) suggests that the change from Alexa to Alex Trebek's voice may have tipped off users. Participants also significantly relied on the familiarity of Alexa's voice when determining whether error messages were real or fake (see Section 5.2.2).

Although these recommendations may help a user distinguish between native and third-party skills, our research demonstrates an inherent tradeoff between usability and transparency about the origin of a skill. The audio cues, although potentially effective, may be a distraction to users, as Amazon attempts to build a seamless voice conversation experience where users are not expected to notice the switch in the skill context [40]. Additional voices would similarly be a simple way to differentiate native functionality from third-party skills; however, such a change would likely make the Alexa experience less seamless.

6.2 Recommendation 2: Following Consistent Alexa Design Standards

Finding 3 shows that some participants do not know what commands Alexa can understand to invoke skills. This observation highlights a design of Alexa that is inconsistent with the discoverability principle.

A comprehensive education of all available commands is unrealistic, as it places unnecessary cognitive burden on the user. According to one guide [41], there are more than 200 commands to invoke various native skills. Furthermore, for every new third-party skill invoked, a user would have to remember the new commands associated with the skill.

Given that there are at least 47,000 skills available, Alexa could learn from the discoverability design principle [16] and follow common standards on what functions are and are not available on Alexa natively. For instance, it is possible for an Alexa user to change the volume but not mute the device, set an alarm but not change the time zone, and buy groceries but not music. One simple solution is for all hardware-related commands to be strictly non-verbal. Whenever the Alexa system detects a command for a hardware-related feature such as changing the volume, it should clearly respond that such kinds of commands cannot be done; currently, if a user asks Alexa to mute the device or turn off, Alexa just ignores the command. Again, the exact design is not as important as Amazon setting a consistent standard that it clearly shares with developers and users.

Additionally, Alexa could impose strict standards on how to invoke skills. The fact that "Please go away" could actually invoke a skill (Finding 3) potentially threatens users' privacy. Although Amazon recommends certain common phrases for invoking skills such "Ask <invocation name> <some action>," "Tell <invocation name> <some action>," and "Open <invocation name>," any phrase (other than some reserved for system functionality) can be used to open a third-party skill on Alexa [20]. This design creates a potentially confusing situation for users. Although many skills conform to common naming standards, Alexa's design leaves a backdoor for malicious skills to trick users or for one skill to accidentally obtain sensitive user information instead of the intended one. We recommend that Alexa follow a strict standard for invocation—for instance, "Open <invocation name>," but not any other phrases. Another recommendation is for Alexa to announce

information about the skill, such as the developer's name, before running the skill for the first time; this approach could provide users with more transparency on the third party. However, these recommendations are, again, associated with usability tradeoffs, because they make Alexa's VUI less flexible and more cumbersome to interact with and may go against Amazon's attempts to build a seamless voice conversation experience [40].

7 LIMITATIONS AND FUTURE WORK

Scaling to more skills. Although most of our participants believed that the "malicious" skills we had developed were native skills, it is unclear how often similar skills could be deployed on the Amazon Skill Store. If these malicious skills are prevalent, a user could confuse a malicious skill for a native skill or another benign skill, or a user could confuse one benign skill for another benign skill; in either case, the user could be revealing sensitive information to an unintended third-party skill developer, which poses a privacy risk.

To identify such skills in the wild, one of the challenges is scalability. In particular, more than 50,000 skills are available on the Amazon Skill Store at the time of the writing. We could leverage existing tecniques [42] to programmatically invoke each skill and, based on the skill's response, determine if the skill resembles another native or third-party skill whether intentionally or unintentionally, as such resemblance may cause confusion among users. This automatic technique is difficult, because skills are executed remotely and each verbal interaction with a skill is associated with an HTTP request [18].

Furthermore, we plan to include more skills that require additional account linkage (e.g., linking to a user's third-party shopping account) and/or which asks for additional permission—which we currently did not include. These additional configurations are likely to require additional interactions from users; participants in the new user study may potentially exhibit different behaviors.

Finally, we plan to test vishing and eavesdropping attacks on Alexa [43]. These attacks take advantage of what appears to be long, silent pauses spoken by skills while listening for the user's input and/or convincing the user that Alexa has terminated the current skill. Although this attack has shown to work in the lab, we plan to evaluate, through a similar user study like this current one, whether users would fall prey to such attacks.

Expanding recruitment. For the survey, we recruited 103 university students and 134 MTurk workers. Although we were able to show that the participants responses may differ across demographic groups (e.g., "Everyone," "Owners/All," and "Owners/Freq"), we could demonstrate that such differences had statistical significance for only one survey question (Finding 1 in Section 5.1). For all other questions, our chi-squared tests failed to reject the null hypotheses that the responses of the survey questions were independent of the demographic groups (i.e., whether the participant was recruited in the university or MTurk group, whether the participant owns an Alexa device, and whether the participant indicated being a frequent user of the device).

As a result of this limitation, we plan to recruit more participants in our future work. An increased sample size could give us more statistical power in chi-squared tests. Additionally, we plan to collect more demographic information, such as technology literacy and education level, which could potentially highlight differences in behaviors across specific demographic groups.

8 CONCLUSION

In this work, we surveyed 237 new and existing users of Alexa devices. We found that some participants were unaware that skills could be developed by third parties, that most participants failed to distinguish native and third-party skills and voice messages, and that they often did not understand what functions or voice commands could be understood by Alexa. Surprisingly, participants

with more familiarity and experience with Alexa tended to show signs of confusion. These findings suggest that a user may accidentally invoke an unintended skill without being aware of this mistake; regardless of whether the skill is malicious or benign, the unintended third party may obtain sensitive user information, thus giving rise to privacy risks. Our recommendations include developing audio and visual indicators of native and third-party contexts, as well as following a consistent design standard to help users learn what functions are and are not possible on Alexa.

REFERENCES

[1] Sarah Perez. 2019. Over a quarter of US adults now own a smart speaker, typically an Amazon Echo. *Tech Crunch*. Retrieved February 2, 2020 from https://techcrunch.com/2019/03/08/over-a-quarter-of-u-s-adults-now-own-a-smart-speaker-typically-an-amazon-echo/.

[2] Amazon. n.d. Number of English Skills on Amazon Alexa (Internet Archive). Retrieved July 29, 2021 from https://bit.ly/366Z70G.

[3] Amazon. 2019. Alexa Skills Store. Retrieved September 11, 2019 from https://www.amazon.com/alexa-skills/b?ie=UTF8&node=13727921011.

[4] Nathaniel Fruchter and Ilaria Liccardi. 2018. Consumer attitudes towards privacy and security in home assistants. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–6.

[5] 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 (Nov. 2018), Article 102, 31 pages. DOI: http://dx.doi.org/10.1145/3274371

[6] Nathan Malkin, Joe Deatrick, Allen Tong, Primal Wijesekera, Serge Egelman, and David Wagner. 2019. Privacy attitudes of smart speaker users. *Proceedings on Privacy Enhancing Technologies* 2019, 4 (2019), 250–271.

[7] 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*.

[8] Y. Gao, Z. Pan, H. Wang, and G. Chen. 2018. Alexa, my love: Analyzing reviews of Amazon Echo. In *Proceedings of the 2018 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People, and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI'18)*. 372–380. DOI: http://dx.doi.org/10.1109/SmartWorld.2018.00094

[9] Irene Lopatovska and Harriet Williams. 2018. Personification of the Amazon Alexa: BFF or a mindless companion. In *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval (CHIIR'18)*. ACM, New York, NY, 265–268. DOI: http://dx.doi.org/10.1145/3176349.3176868

[10] Amanda Purington, Jessie G. Taft, Shruti Sannon, Natalya N. Bazarova, and Samuel Hardman Taylor. 2017. "Alexa Is My New BFF": Social roles, user satisfaction, and personification of the Amazon Echo. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA'17)*. ACM, New York, NY, 2853–2859. DOI: http://dx.doi.org/10.1145/3027063.3053246

[11] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice interfaces in everyday life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18)*. ACM, New York, NY, Article 640, 12 pages. DOI: http://dx.doi.org/10.1145/3173574.3174214

[12] 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 (DIS'18)*. ACM, New York, NY, 857–868. DOI: http://dx.doi.org/10.1145/3196709.3196772

[13] Aarthi Easwara Moorthy and Kim-Phuong L. Vu. 2015. Privacy concerns for use of voice activated personal assistant in the public space. *International Journal of Human–Computer Interaction* 31, 4 (2015), 307–335. DOI: http://dx.doi.org/10.1080/10447318.2014.986642

[14] Deepak Kumar, Riccardo Paccagnella, Paul Murley, Eric Hennenfent, Joshua Mason, Adam Bates, and Michael Bailey. 2018. Skill squatting attacks on Amazon Alexa. In *Proceedings of the 27th USENIX Conference on Security Symposium (SEC'18)*. 33–47. http://dl.acm.org/citation.cfm?id=3277203.3277207.

[15] Nan Zhang, Xianghang Mi, Xuan Feng, XiaoFeng Wang, Yuan Tian, and Feng Qian. 2018. Understanding and mitigating the security risks of voice-controlled third-party skills on Amazon Alexa and Google Home. arXiv:1805.01525

[16] Don Norman. 2013. The psychology of everyday actions. In *The Design of Everyday Things* (revised, expanded ed.). Basic Books, 37–122.

[17] XXX. 2018. Amazon Echo Has 23% Share of Smart Speakers in Use: Strategy Analytics. Retrieved May 3, 2019 from https://news.strategyanalytics.com/press-release/intelligent-home/amazon-echo-has-23-share-smart-speakers-use-strategy-analytics.

[18] Alexa. 2019. Alexa Voice Service. Retrieved May 3, 2019 from https://developer.amazon.com/alexa-voice-service.

[19] Alexa. 2019. Host a Custom Skill as a Web Service. Retrieved May 3, 2019 from https://developer.amazon.com/docs/custom-skills/host-a-custom-skill-as-a-web-service.html.
[20] Alexa. 2019. Understanding How Users Invoke Custom Skills. Retrieved May 6, 2019 from https://developer.amazon.com/docs/custom-skills/understanding-how-users-invoke-custom-skills.html.
[21] Amazon.com help: What do the lights on your echo device mean? [Online]. Retrieved from https://www.amazon.com/gp/help/customer/display.html?nodeId=GKLDRFT7FP4FZE56.
[22] Choose the invocation name for a custom skill | alexa skills kit. [Online]. Retrieved from https://developer.amazon.com/en-US/docs/alexa/customskills/choose-the-invocation-name-for-a-custom-skill.html.
[23] Amazon.com: Home wifi: Alexa skills. [Online]. Retrieved from https://voiceapp.store/listing/home-wifi/.
[24] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2017. DolphinAttack: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, New York, NY, 103–117.
[25] Noah Apthorpe, Danny Yuxing Huang, Dillon Reisman, Arvind Narayanan, and Nick Feamster. 2019. Keeping the smart home private with smart(er) IoT traffic shaping. *Proceedings on Privacy Enhancing Technologies* 2019, 3 (2019), 128–148.
[26] Robert Nyman. 2012. Using the Fullscreen API in web browsers. *Mozilla Hacks*. Retrieved July 29, 2021 from https://hacks.mozilla.org/2012/01/using-the-fullscreen-api-in-web-browsers.
[27] Apple Insider Staff. 2017. Proof of concept phishing attack mimics iOS popups to steal user passwords. *AI*. Retrieved July 29, 2021 from https://appleinsider.com/articles/17/10/10/proof-of-concept-phishing-attack-mimics-ios-popups-to-steal-user-passwords.
[28] Joseph Weizenbaum. 1966. ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9, 1 (Jan. 1966), 36–45. DOI: http://dx.doi.org/10.1145/365153.365168
[29] Brenda Laurel and S. Joy Mountford (Eds.). 1990. *The Art of Human-Computer Interface Design*. Addison-Wesley-Longman, Boston, MA.
[30] Cathy Pearl. 2016. *Designing Voice User Interfaces*. O'Reilly Media.
[31] Gustavo López, Luis Quesada, and Luis A. Guerrero. 2018. Alexa vs. Siri vs. Cortana vs. Google Assistant: A comparison of speech-based natural user interfaces. In *Advances in Human Factors and Systems Interaction*, Isabel L. Nunes (Ed.). Springer International Publishing, Cham, Switzerland, 241–250.
[32] H. Chung, M. Iorga, J. Voas, and S. Lee. 2017. "Alexa, Can I Trust You?" *Computer* 50, 9 (2017), 100–104. DOI: http://dx.doi.org/10.1109/MC.2017.3571053
[33] Deepak Kumar, Riccardo Paccagnella, Paul Murley, Eric Hennenfent, Joshua Mason, Adam Bates, and Michael Bailey. 2018. Skill squatting attacks on Amazon Alexa. In *Proceedings of the 27th USENIX Security Symposium (USENIX Security'18)*. 33–47.
[34] 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 *Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP'19)*. IEEE, Los Alamitos, CA.
[35] Madiha Tabassum, Tomasz Kosiński, Alisa Frik, Nathan Malkin, Primal Wijesekera, Serge Egelman, and Heather Richter Lipford. 2019. Investigating users' preferences and expectations for always-listening voice assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–23.
[36] Jide S. Edu, Jose M. Such, and Guillermo Suarez-Tangil. 2020. Smart home personal assistants: A security and privacy review. *ACM Computing Surveys* 53, 6 (2020), 116.
[37] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. 2019. How well do my results generalize? Comparing security and privacy survey results from MTurk, web, and telephone samples. In *Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP'19)*. IEEE, Los Alamitos, CA, 227–244.
[38] Johnny Saldañna. 2013. *The Coding Manual for Qualitative Researchers* (2nd ed.). SAGE, Los Angeles, CA.
[39] Amazon Alexa. 2019. Save Data Between Sessions. Retrieved July 29, 2021 from https://developer.amazon.com/docs/custom-skills/manage-skill-session-and-session-attributes.html#save-data-between-sessions.
[40] Alexa. 2019. Alexa Conversations: Creating Natural Voice Experiences Faster. Retrieved September 14, 2019 from https://developer.amazon.com/en-US/alexa/alexa-skills-kit/alexa-conversations.
[41] Taylor Martin. 2019. The Complete List of Alexa Commands So Far. Retrieved September 14, 2019 from https://www.cnet.com/how-to/amazon-echo-the-complete-list-of-alexa-commands/.
[42] Zhixiu Guo, Zijin Lin, Pan Li, and Kai Chen. 2020. SkillExplorer: Understanding the behavior of skills in large scale. In *Proceedings of the 29th USENIX Security Symposium (USENIX Security'20)*. 2649–2666.
[43] Security Research Labs. n.d. Smart Spies: Alexa and Google Home Expose Users to Vishing and Eavesdropping. Retrieved July 29, 2021 from https://srlabs.de/bites/smart-spies/.

Received June 2020; revised November 2020; accepted December 2020