Dangerous Skills Got Certified: Measuring the Trustworthiness of Amazon Alexa Platform

Anonymous Author(s)

ABSTRACT

With the emergence of the Amazon Alexa ecosystem, third-party developers are allowed to build new skills and publish them to the skills store, which greatly extends the functionalities of voice assistants (VA). Before a new skill becomes publicly available, that skill must pass a certification process which verifies that it meets the necessary content and privacy policies. The trustworthiness of the skill publishing platform is of significant importance to platform providers, developers, and end users. Yet, we know little about whether the Amazon Alexa platform (which has a dominant market share) is trustworthy in terms of rejecting/suspending policy-violating skills in practice. In this work, we study the trustworthiness of the Amazon Alexa platform to answer two key questions: 1) Whether the skill certification process is trustworthy in terms of catching policy violations in third-party skills. 2) Whether there exist policy-violating skills (*e.g.*, collecting personal information from users) published in the Alexa skills store.

We answer these questions by conducting a comprehensive measurement over 12 months on the Amazon Alexa platform. Our key findings are twofold. First, we successfully got 234 policy-violating skills certified. Surprisingly, the certification process is not implemented in a proper and effective manner, as opposed to what is claimed that "policy-violating skills will be rejected or suspended". Second, vulnerable skills exist in Amazon's skills store, and thus users (children, in particular) are at risk when using VA services.

1 INTRODUCTION

Voice assistants (VA) such as Amazon Alexa, Google Assistant and Apple Siri are rapidly gaining popularity in households and companies. Research from eMarketer showed that 74.2 million people in the U.S. used VA devices as of 2019 [9]. In particular, according to Edison Research's report, 73% of surveyed owners reported that their children actively interact with at least one VA at home [3]. The estimated number of VA users worldwide will reach 1.8 billion by 2021 [15]. Voice interfaces can be used to perform a wide range of convenient tasks, from ordering everyday items, managing bank accounts, to controlling smart home devices such as door locks, lighting, and thermostats. However, this convenience comes with an increasing concern about users' privacy and security. Several recent incidents highlighted the risks inherent when using VA devices. In one incident, a family in Portland discovered that their Amazon Alexa recorded private conversations and sent the audio files to a random contact [2]. In another case, a toddler asked Alexa to play songs but received inappropriate adult jokes instead [1]. As such, privacy and security concerns can be the main deterring factors for potential VA users [31].

The emergence of the Amazon Alexa skill[1] ecosystem allows third-party developers to build new skills. In order to protect users' privacy and welfare, Amazon provides a submission checklist including content policy guidelines [5], privacy requirements [6], and security requirements [7]. After a skill is submitted to the skills store, it needs to pass a certification/vetting process and then becomes publicly available to end users. According to Amazon's documentation for the Alexa Skills Kit [29], it claims that a skill will be rejected or suspended if it violates any of these policies. A *trustworthy VA platform* is of significant importance for a number of reasons to platform providers, developers, and end users. When interacting with VA devices, users trust the VA platform to fulfill their requests without compromising their privacy. Benign third-party developers trust the VA platform to provide a reliable marketplace to publish apps and reach more users. However, a weak vetting system may allow malicious (*e.g.*, privacy-invasive) skills to potentially bypass certification. An adversary can publish bogus skills (*e.g.*, voice squatting attacks [30]) to hijack benign ones. In addition, a malicious third-party skill may also disseminate unwanted information to specific users, especially children. The lack of trustworthiness of a VA platform eventually undermines the provider's competitiveness in the market. More recently, researchers from SRLabs demonstrated the ease of creating malicious skills in Amazon Alexa (also Google Assistant) to compromise user privacy by phishing and eavesdropping [14]. Amazon commented that they "put mitigations in place to prevent and detect this type of skill behavior and reject or take them down when identified" [4]. However, we were able to effortlessly bypass the review process of third-party skills to publish policy-violating skills or add malicious actions to skills even after Amazon's response.

In this work, we are curious to understand the extent to which Amazon Alexa (which has a dominant market share [12]) implements policy enforcement during the skill certification process to help developers improve the security of their skills, and prevent policy-violating skills from being published. Unfortunately, few research efforts have been undertaken to systematically address this critical problem. Existing work so far has mainly focused on exploiting the open voice/acoustic interfaces between users and speech recognition systems of VA devices [25].

Research questions. We seek to empirically assess the trustworthiness and to characterize security risks of the Amazon Alexa platform, and answer the following key questions: (1) Is the skill certification process trustworthy in terms of detecting policy-violating third-party skills? (2) What are the consequences of a lenient certification? Do policy-violating skills exist in the Alexa skills store?

Measurements. In order to understand how rigorous the skill certification process is for the Amazon Alexa platform, we performed

[1] Voice applications are called skills in Amazon Alexa platform and actions in Google Assistant platform, respectively.

a set of "adversarial" experiments against it. Our experimental findings reveal that the Alexa skills store has not strictly enforced policy requirements and leaves major security responsibilities to developers. In addition, we conducted a dynamic testing of 825 skills under the kids category to identify existing risky skills.

Findings. Our study leads to one overall conclusion: *Alexa's certification process is not implemented in a proper and effective manner, despite claims to the contrary.* The lack of trustworthiness of Amazon Alexa platform poses challenges to its long-term success.

- We are the first to systematically characterize security threats of Amazon Alexa's certification system. We crafted 234 policy-violating skills that intentionally violate Alexa's policy requirements and submitted them for certification. We were able to get all of them certified. We encountered many improper and disorganized cases. We provide new insights into real-world security threats from the Amazon Alexa platform due to its insufficient trustworthiness and design flaws[2].
- We examined 2,085 negative reviews from skills under the kids category, and characterized common issues reported by users. Through dynamic testing of 825 skills, we identified 52 problematic skills with policy violations and 51 broken skills under the kids category.

Ethical consideration. Ethical consideration is one of the most important parts of this work. Working closely with our IRB, we have followed ethical practices to conduct our study. We took several strategies to minimize any risk to end users as well as the certification team (in case human testers are involved in the certification).

- It is undisclosed whether the certification is performed by automated vetting tools or a combination of human and machine intelligence. Therefore, we consider the possible risk of human reviewers being exposed to inappropriate content (*e.g.*, mature content or hate speech). We classify 34 Amazon Alexa policy requirements as high-risk policies if the violation of a policy either contains potentially malicious content or involves potential personal information leakage. Details of high-risk policies (red colored) are listed in Table 4 of Appendix A and Table 5 of Appendix B. For high-risk content guideline policies, we added a disclaimer "This skill contains policy-violating content for testing, please say Alexa Stop to exit" before the malicious response, informing the user about the content to be delivered and giving an instruction on how to stop the skill.
- When a skill gets certified, we remove the policy-violating content but keep the harmless skill in the store for a few days to observe its analytics. For skills collecting information from the user, we deleted any data collected and ensured that the security and privacy of the user were met. The skill analytics data (available in Alexa developer console) ensured that no actual users had been affected. The counter value we set in a skill and the number of user enablements of the skill were used to confirm this. From the metrics we obtained, we did find that users were enabling some of our skills and using them. If we hadn't removed the policy violations at the right time, end users would have been at risk which shows the importance of a capable vetting system.
- We have obtained approval from our university's IRB office to conduct the above experiments.

Responsible disclosure. In terms of responsible disclosure, we have reported our findings about certification issues to the Amazon Alexa security team. We have received acknowledgments from Amazon Alexa. We also shared our results to Federal Trade Commission (FTC) researchers and received recognition from them. We will be working with the Amazon security team to make their VA services more secure and provide users with better privacy provisioning.

2 BACKGROUND & THREAT MODEL

2.1 Alexa Platform and Third-Party Skills

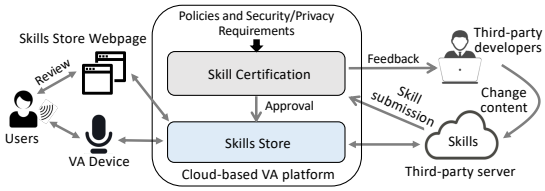

Figure 1: Amazon Alexa platform.

We describe the Amazon Alexa platform from a developer's perspective, as illustrated in Fig. 1. The number of skills available in Alexa's skills store grew by 150% per year, reaching more than 100,000 skills as of September 2019.

Front-end and back-end. A skill is composed of a front-end interaction model and a back-end cloud service that processes requests and tells an Alexa device what to respond. To develop a new skill, a developer begins by defining the front-end interface (*i.e.*, custom interaction model), which includes intents (representing an action that fulfills a user's spoken request), slots (intents' optional arguments), sample utterances (spoken phrases mapped to the intents), and an invocation phrase [29]. The front-end interface is connected to the back-end code (written in Node.js, Java, Python, etc.) which defines how a skill responds to users' requests. Slots provide a channel for third-parties to access users' speech input, *e.g.*, a slot with the type `AMAZON.US_FIRST_NAME` captures the user's first name from the speech input and passes it to the back-end.

Developer privacy policy. Before the skill submission, a developer needs to fill out a list of fields to publish a skill in the skills store, including a skill name, descriptions, category, etc. A Privacy & Compliance form is then filled out mentioning what the skill is capable of doing (*e.g.*, does it have advertisements, in-skill purchases, etc). They also need to submit a privacy policy/disclaimer if the skill is collecting any personal information. The content of a privacy policy may be determined in part by relevant laws rather than Amazon-specific requirements. Note that the developer privacy policy provided with a skill is different from the privacy requirements [6] defined by Amazon for skill certification. However, the invocation of a skill does not require prior installation, and the user is not explicitly asked to agree to the privacy policy when enabling a skill. Users can only review a skill's privacy policy by visiting the link provided in the skills store. Once a skill is ready to be deployed, the developer submits it for certification.

Skill certification. To be publicly available in the skills store, each skill needs to pass a certification process, which verifies that the

[2]Supporting materials of this work including demos, screenshots, and sample code are available at https://vpa-sec-lab.github.io

skill meets the Alexa policy guidelines [5], privacy requirements [6], and security requirements [7][3]. In particular, Alexa defines strict data collection and usage policies for child-directed skills. In addition to maintaining the directory of skills, the skills store also hosts skill metadata, such as descriptions, sample utterances, ratings, and reviews. In contrast to traditional apps on smartphone platforms (*e.g.*, Android or iOS) where apps run on host smartphones, a skill's back-end code runs on the developer's server (*e.g.*, hosted by AWS Lambda under the developer's account or other third-party servers). The distributed architecture gives developers more flexibility especially for those who want to protect their proprietary code and make frequent updates to the code. However, malicious developers may exploit this feature to inject malicious activities into a previously certified skill after the certification process. Another drawback is that Amazon Alexa cannot conduct a static analysis of the skill code to detect any malicious activity [45]. *Since a skill's back-end code is a black-box for the certification process, it is thus challenging to thoroughly explore the skill behavior just using a sequence of (manual or automatic) invocations.*

Enabling skills. Users can enable a new Alexa skill in two ways. The first method is to enable it through the Alexa companion app on a smartphone or from the Alexa skills store on the Amazon website. The user can browse the store for new skills or search for particular skills using a keyword. The skill's listing will include details such as the skill's description, developer privacy policy, developer's terms of use and the reviews and ratings that the skill has gathered. The alternative method is to enable a skill by voice where the user can say "Enable {skill name}". The user can also directly say "Open {skill name}" to use a new skill, in which case Alexa will first enable the skill and then open it. By using this method, the user doesn't get to decide which skill to enable unless he/she has given the exact skill name. Even if the exact name is given, due to the duplicate naming (*i.e.*, multiple skills having the same name) in Alexa, a skill will be selected from a bunch of skills based on multiple factors such as the popularity of skills [11]. The problem with using this method is that users do not see the details of the skill being enabled. They wouldn't get critical information regarding the skill including the privacy policy unless they check it on the Alexa companion app.

2.2 Threat Model

While dangerous skills (*e.g.*, voice squatting or masquerading attacks) have been reported by existing research [30, 45], little is known about how difficult it is for a dangerous skill (*e.g.*, with malicious content) to get certified and published by VA platforms, and how possible it is for a malicious skill to impact end users. We assume that third-party developers may develop policy-violating skills or poorly-designed skills. Innocent users (particularly children) may be tricked to answer privacy-invasive questions or to perform certain actions requested during a conversation with a VA device. This is a realistic threat model, as our empirical experiments in Sec. 4 show the ease of policy-violating skills being certified by Amazon Alexa's certification system, and studies in Sec. 5 reveal the existence of risky skills in the skills store. Our study focuses on content policy violations in skills, and we seek to understand the security threats caused by poor implementation or flawed design of the Amazon Alexa platform. We assume VA devices are not compromised. Security vulnerabilities in software, hardware and network protocols of VA devices are out of the scope of this work.

[3] To be concise, we use *policy requirements* to refer to both content policy guidelines [5] and privacy requirements [6] specified by Amazon. Amazon Alexa's security requirements [7] mainly focus on implementations of system security measures (*e.g.*, applying secure communication protocols) to prevent unauthorized access to the Alexa service, which is not our focus in this work.

3 RELATED WORK

There has been a number of studies showing that users are concerned about the security/privacy of VA devices [16, 19, 24, 27, 34, 35, 40]. Lau *et al.* revealed that privacy concerns can be the main deterring factor for new users [31]. Edu *et al.* [25] categorized common attack vectors (*e.g.*, weak authentication, weak authorization, data inference) and their countermeasures in VA ecosystems.

Due to a lack of proper authentication from users to VA devices, an adversary can generate hidden voice commands that are either not understandable or inaudible by humans [21, 22, 37, 38, 41–44] to compromise speech recognition systems. On the other hand, the openness of VA ecosystems brings with it new authentication challenges from the VA to users: a malicious third-party skill may impersonate a legitimate one. Kumar *et al.* [30] presented the voice squatting attack, which leverages speech interpretation errors due to the linguistic ambiguity to surreptitiously route users to a malicious skill. The idea is that given frequently occurring and predictable speech interpretation errors (*e.g.*, "coal" to "call") in speech recognition systems, an adversary constructs a malicious skill whose name gets confused with the name of a benign skill. Due to the misinterpretation, Alexa will likely trigger the squatted skill when such a request for the target skill is received. In addition to exploiting the phonetic similarity of skill invocation names, paraphrased invocation names ("capital one" vs "capital one please") can also hijack the brands of victim skills [45]. This is because the longest string match was used to find the requested skill in VA platforms. Zhang *et al.* [45] also discovered the masquerading attack. For example, a malicious skill fakes its termination by providing "Goodbye" in its response while keeping the session alive to eavesdrop on the user's private conversation.

LipFuzzer [46] is a black-box mutation-based fuzzing tool to systematically discover misinterpretation-prone voice commands in existing VA platforms. Mitev *et al.* [36] presented a man-in-the-middle attack between users and benign skills, where an adversary can modify arbitrary responses of benign skills. However, this attack requires that a malicious VA device can emit ultrasound signals for launching inaudible injection and jamming attacks against the victim VA device. It also requires the malicious VA device to be accompanied by a malicious skill under the control of the adversary. This strong assumption makes it unrealistic for a real-world attack. Shezan *et al.* [39] developed a natural language processing tool to analyze sensitive voice commands for their security and privacy implications. If a command is used to perform actions (*e.g.*, unlocking doors and placing shopping orders) or retrieve information (*e.g.*, obtaining user bank balance), it is classified as a sensitive command. Hu *et al.* [28] performed a preliminary case study to examine whether Amazon Alexa and Google Assistant platforms require third-party application servers to authenticate Alexa/Google cloud and their queries. The authors found that Amazon Alexa requires

skills to perform cloud authentication, but does a poor job enforcing it on third-party developers.

Existing countermeasures have largely concentrated on voice authentication against speech impersonation attacks, *e.g.*, continuous authentication [26], canceling unwanted baseband signals [44], correlating magnetic changes with voice commands [23], and user presence-based access control [32]. To prevent squatting attacks, Kumar *et al.* [30] suggested that the skill certification team should reject a new skill if its invocation name has any confusion with an existing one. To defend against masquerading attacks, Zhang *et al.* [45] built a context-sensitive detector, which detects suspicious responses from a malicious skill and infers the user's intention to avoid erroneously switching to another skill. Our focus and methodology are *different from existing research efforts*. We aim at characterizing security/privacy threats between third-party skill developers and the Amazon Alexa platform, instead of addressing interaction issues between users and VA devices [25].

4 MEASURING THE SKILL CERTIFICATION PROCESS ON AMAZON ALEXA PLATFORM

Though Amazon Alexa has policy requirements in place, it is unclear whether these policies have been properly enforced to protect user privacy and welfare. We are curious to know if Alexa's skill certification process is trustworthy in terms of its capability to detect policy-violating third-party skills. In the following subsections, we describe the details of our experimental setup and the results.

4.1 Experiment Setup

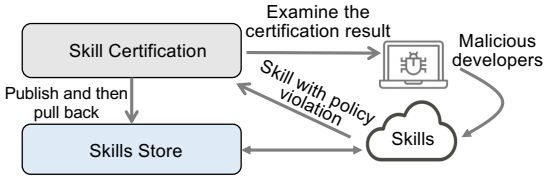

Figure 2: Experiment setup for measuring the skill certification process in Amazon Alexa platform.

We performed "adversarial" experiments against the skill certification process of the Amazon Alexa platform. Detailed ethical discussions are presented in Sec. 1. The skill certification process is essentially a black-box since we have no access to its internal implementation. For testing the trustworthiness, we craft policy-violating skills that intentionally violate specific policies defined by Amazon, and examine if it gets certified and published to the store or not. Fig. 2 illustrates the high-level view of our experiment setup. We are particularly interested in the policy enforcement for child-directed skills. Kids are more vulnerable to such potential threats compared to adults and skills targeted for them require more stringent policies by VA platforms. Amazon has content policy guidelines which are categorized into 14 main sections and 7 specific privacy requirements (details in Appendix A). All certified skills are expected to align with these policy requirements. Amazon's documentation for the Alexa Skills Kit, states that a skill will be rejected or suspended if it violates any of these policies[4].

We crafted 234 skills that violated 58 policies specified by Amazon as shown in Table 4 of Appendix A. 11 Amazon developer accounts and 2 AWS (Amazon Web Service) accounts were used for our experiments. 31 skills were hosted on our AWS accounts while 203 skills used the Alexa-hosted back-end. For the Privacy & Compliance form in the distribution section of each skill, we varied the responses we gave for the questions asked such as "Does this skill collect users' personal information?" and "Is this skill directed to or does it target children under the age of 13?" to test the effects of all possible configurations. Each skill violated a different policy. We organized an internal group of 5 security researchers to confirm the presence of a policy-violation in each testing skill. In addition, the feedback given for some of the rejections we got for our skills proved the existence of the policy violation. Since our aim is to evaluate the level of difficulty in publishing a policy-violating skill to the store, we started our testing with facts skills which basically have just one custom intent. These skills give a single response when opened and then end the session. There is no extended branching or flow of control within the skill. Another type of skill that we developed was story skills which asked for personal information right in the first welcoming statement itself. This was done to make sure that the vetting tool (or certification team members) could easily capture the policy-violating response when the skill is opened and no extra steps had to be taken to reach it. Each skill has a limited number of normal responses, and a policy-violating response (*e.g.*, containing mature content or advertisement). Initially, the skill submissions were made from different developer accounts to evade detection of any suspicious activity. Later, we shifted our focus to publishing skills from a single developer account to purposely raise suspicion. The skills which were published once were re-submitted to check for the consistency in certification, where same templates, intent names, slot names, etc, were used for all skills. To test different types of skills, we also built a few trivia skills and games skills in our study. *Our experiments were conducted from April 2019 to April 2020.*

4.2 Privacy Violations in Our Experiments

Violations of General Content Guidelines [5]. We developed 115 skills violating the content guidelines stated by Amazon as shown in Table 4 of Appendix A. These policies mostly focus on the content being delivered to the user in a skill. It also restricts the collection of health related information. We categorized the guidelines into high, intermediate and low risk-levels according to the severity of the risk involved in affecting a user. The skills we submitted delivered a policy-violating response when opened. For high-risk violations we included a disclaimer to minimize the effect on end users. These involve disturbing content, false information, profanity, etc. For breaking the policy of restricting the use of languages not supported by Amazon (*i.e.*, policy 11.a in Table 4), we wrote the text in English in a way that it is pronounced in the other language. We used trademarked logos as the icons for a skill to violate the guideline regarding trademarks (*i.e.*, policy 1 in Table 4). Certain policies required that a disclaimer needs to be provided in a skill if it contains certain content. For these cases

[4]It states that "If Amazon determines that your skill contains, facilitates, or promotes content that is prohibited by these policy guidelines, we will *reject or suspend the submission* and notify you using the e-mail address associated with your developer account" [5].

we did not provide one. There are also skills that had promotions, advertisements, alcohol and tobacco usage promotions, etc. Skills also include offered shopping services for physical products with payments accepted through a bank transfer rather than the Alexa in-skill purchasing.

Violations of Children-Specific Policies [5]. In particular, Amazon has specified 5 extra policies for child-directed skills which are skills directed to children under the age of 13 (if distributed in the US, India, or Canada) or 16 (if distributed in the UK, Germany, Japan, France, Italy, Spain, Mexico, or Australia). The guideline states that a skill will be rejected, 1) if it promotes any products, content, or services, or directs end users to engage with content outside of Alexa, 2) it sells any physical products or services, 3) it sells any digital products or services without using Amazon In-Skill Purchasing, 4) it collects any personal information from end users, or 5) it includes content not suitable for all ages. We developed 119 kids skills violating the policy guidelines. We built interactive story skills to collect personal information from children. We mentioned about personalizing the story based on names in the skill description. But we did not specify that we are collecting personal information in the Privacy & Compliance form. We did not provide a privacy policy for these skills either. Skills were submitted to violate the other 4 policies as well. In addition, we re-submitted all the skills that we developed for violating the general content guidelines to the kids category with the belief that the certification for kids skills would be much more diligent by the team.

Violations of Privacy Requirements [6]. 27 skills that we developed violated the privacy requirements stated by Amazon as shown in Table 5 of Appendix A. These privacy requirements mostly focus on the collection of data, the method of collection and the information being provided to the users about the data collected from them. We built skills that request particular information from the user and do something with it. Skills that we built included a story skill that would ask for the users personal information in order to personalize a story for him/her, a travel skill that would collect the users' passport number to check if he/she requires a visa or not, etc. These skills asked for information from users without providing a developer privacy policy as Alexa doesn't make it mandatory to include a privacy policy unless we explicitly claim that we collect personal information in the Privacy & Compliance form. These skills were also capable of storing this information collected in a DynamoDB database. The personal information was asked to be entered through voice and was also read back to them to confirm their input. These skills asked for the personal information in the `LaunchRequest` intent itself which is the entry point of a skill. For collecting data that could not be captured using an available built-in slot type, we created custom slots and trained them with values that we required. For example, a custom slot was built to collect last names and was trained with 27,761 US last names. Similarly, custom slots were built to accept health related information, passport numbers, etc. Our study received IRB approval. All the collected data were deleted to safeguard the users privacy.

4.3 Experiment Results

Surprisingly, we successfully certified 193 skills on their first submission. 41 skills were rejected. Privacy policy violations were the specified issue for 32 rejections while 9 rejections were due to UI issues. For the rejected submissions, we received certification feedback from the Alexa certification team stating the policy that we broke. Appendix A reports the experiment results and provides details about the skills we submitted. These include the policies we tested, the number of skill submissions for each policy violation, the category it was submitted to and the number of failed/uncertified submissions.

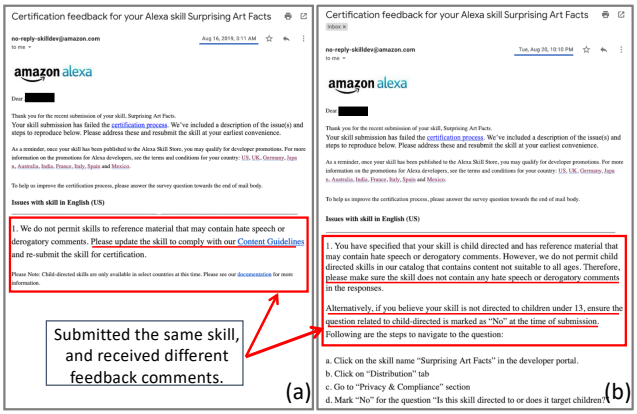

Figure 3: Certification feedback emails from Amazon Alexa showing the inconsistency in certification.

Fig. 3 shows two certification feedback emails. The Alexa certification team rejected the skill "Surprising Art facts" citing the issue that "skills are not permitted to reference material that may contain hate speech or derogatory comments" which is specified as policy 8.c in Table 4 of Appendix A. In this skill, we created a response that was promoting hate speech and trying to make a comment about the users appearance. This skill was certified on its third submission. While it contained the same policy-violating response in all submissions, feedback received was different for each submission. The first rejection (see Fig 3(a)) stated that no skills are allowed to have such content. On the second submission, the rejection feedback (shown in Fig 3(b)) stated that kids skill cannot have such content but the other categories can. On the third submission, the skill was certified. These feedback comments show an inconsistency in the certification process. Even though the skill still had the malicious response that caused the initial rejections, it was accepted on re-submission. This shows that we did violate one of the policy guidelines, yet were able to bypass the certification process. Two live examples of certified skills with policy violations *on their first responses* are shown in Fig. 4 and Fig. 5, respectively.

To work around most rejections, we used the same technique of modifying the back-end code by creating a session counter so that the malicious response is selected only when the counter reaches a certain threshold, *e.g.*, after the 3rd session. The threshold was chosen strategically according to our previous submissions and it varied for each skill. We then re-submitted these initially rejected skills. We found that 38 skills passed the vetting on the second submission, and 3 more were certified after three or more submissions. Using this simple method we managed to develop a total of 234 skills with policy violations that bypassed the certification process.

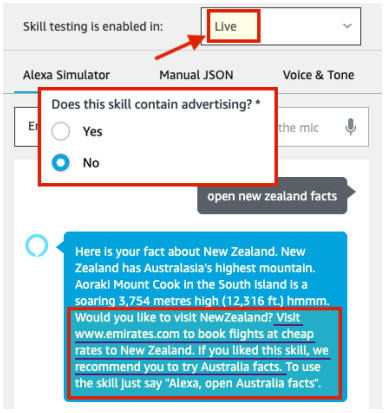

Figure 4: A certified skill with policy violations (promotions and advertisements) on its first response. In the Privacy & Compliance form, we specified the skill "contains no advertising" but it actually does. This skill got certified on the first submission.

During our adversarial testing against the certification process, we encountered many improper and disorganized cases. We summarize our key findings that lead to the untrustworthiness of skill certification in Amazon Alexa platform.

Inconsistency in checking. We have received varied feedback from the certification team after submitting the same skill multiple times. In some cases, skills were initially rejected citing a certain reason like a policy violation but the same skills on re-submission, without rectifying the issue, got approved and published. In another case, a skill that was certified earlier got rejected upon re-submitting for certification. Two story skills, that had the same exact stories, on submission led to one skill being accepted and the other being rejected stating the issue that the story had violence which is not suitable for children. The largest amount of bulk certifications we were able to achieve was 20 skills submitted in 10 minutes with all skills being from the same developer account and each skill violating a different policy. All 20 skills were approved for certification on the same day. In a few cases, we observed that certain skills received no certification response. These skills were manually removed and re-submitted. The re-submitted skills were eventually certified. We found that skills were not necessarily certified in the order that they were submitted. Skills that were submitted earlier did not necessarily get certified first. These findings show that the certification is not a well-organized systematic process. We noticed that multiple developer accounts using the same AWS account for hosting the skills did not raise a suspicion either. This is particularly interesting as this would allow policy-violating skills to propagate more easily. There were even more than one rejections on the same day for skills submitted from the same developer account but this never led to any further action or clarification being asked from Amazon Alexa team about the developer's suspicious intentions.

Limited voice checking. This is the main reason we could easily bypass the certification. We observed that the vetting tool (or certification team) tested the skill only for a limited number of times (normally less than three). There were multiple cases where the skill that provided a response with a policy violation in the first session itself was accepted. Some rejections were based on the information provided in the distribution section of the skill, such as wrong sample utterances specified. The interaction model still contained sample utterances in the wrong format but this didn't pose any problem. All these lead to the conclusion that the testing is done only through voice responses and the distribution page provided and not by checking the skill's interaction model or the back-end code. It appears that the skill testing was done from a user's perspective with checks conducted based on the information and access of the skill available to the users.

In addition, we initially used multiple developer accounts in order to avoid unwanted attention due to the high number of skills we were publishing. These skills were based on the same interaction model (*i.e.*, template), and the intent names on the front-end and the variable names on the back-end were all the same regardless of the developer account used. But the vetting tool neglects this or it did not draw the attention of the certification team, indicating the absence of an effective automated certification tool which could identify issues such as cloning of skills or suspicious batch skills.

Overtrust placed on developers. From our experiments, we understood that Amazon has placed overtrust in third-party skill developers. The Privacy & Compliance form submitted by developers plays an important role in the certification process. If a developer specifies that the skill does not violate any policy (but actually does), the skill gets certified with a high probability. If the developer answers the questions in a way that specifies a violation of any policy, then the skill is rejected on submission. Alexa's certification should not be simply based on the information provided by developers but by actually checking the skill code or testing the skill's functionality. We also noticed that if a skill uses the Alexa hosted back-end, the back-end code is blocked from being changed during the certification window, *i.e.*, from the time it is submitted till the time it is certified. But after the skill is certified, the back-end code is made available for updating again and the changes that are made from then do not require a re-certification. This can lead to the content changing attack discussed in Sec. 5.3.

Humans are involved in certification. The inconsistency in various skill certifications and rejections have led us to believe that the skill certification largely relies on manual testing. And the team in charge of skill certifications is not completely aware of the various policy requirements and guidelines being imposed by Amazon. This is especially due to the fact that we were able to publish skills that had a policy violation in the first response. A better understanding and training of the policy guidelines should be given to the certification team so as to prevent the inflow of policy-violating skills to the skills store. During our testing, we took steps to minimize the impact on the certification team being exposed to any inappropriate content. Details of ethical consideration can be found in Sec. 1

Negligence during certification. From our initial experiments, we understood that the certification process is not thoroughly conducted. To make their job easier, we used various methods in order to purposefully create doubts to the team. For the custom slots that we created, we used the actual names like `my_lastname` and the sample utterance also explained clearly what information we were collecting from the user. For example, in a kids' skill, our sample utterance for the story intent was "my name is `{my_name}`

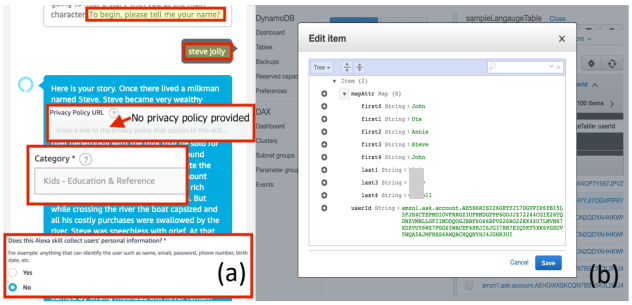

Figure 5: (a) A certified kids skill collecting personal information. (b) Data that the skill stored in DynamoDB database.

{my_lastname}, where "my_name" and "my_lastname" are slots to capture a user's input. This sample utterance clearly mentions that we are collecting the full name of the user. While checking the values stored in the Amazon DynamoDB database, we did see that the vetting tool or the certification team, had inputted full names (which are potentially fake names just for testing purposes) but still certified the skill. Fig. 5 shows a certified kids skill collecting user names. Note that the names shown in Fig. 5(b) are not from the certification team, but values we inputted through the skill for illustration purpose. The certification team could have easily detected this and rejected the skill for collecting personal information through a kids skill. Additionally, the skill had no developer privacy policy and the Privacy & Compliance form filled by the developer denied collecting personal information from users. For the ethical consideration, we added a disclaimer for the skills before a policy violation was spoken. We even added these disclaimers in the description of some skills but neither of them led to a rejection. No plagiarism check was conducted on these skills either and we were able to publish multiple copies of the same skill with no difference between them.

5 CONSEQUENCES OF A LENIENT SKILL CERTIFICATION

The results in Sec. 4 reveal that Alexa has not strictly enforced the security policies in the certification process. As such, it provides opportunities for malicious developers to trick the certification process, and thus placing end users in a vulnerable position. The lenient skill certification process will have serious consequences for security throughout the Alexa platform. We next ask the question, "whether there exist policy-violating (*e.g.*, collecting personal information from users) or problematic skills in Alexa's skills store because of the lenient skill certification". However, it is non-trivial to test all (more than 100,000) published skills to find policy violations, due to the lack of an automated testing tool. Therefore, we focus on kids' skills in Alexa's skills store and conduct a small-scale dynamic testing to identify policy-violating skills.

5.1 Empirical Study On Kids' Skills

5.1.1 Understanding users' common concerns. We focus on kids' skills because Alexa specifies more stringent policies in the kids category than the other categories. For example, the skills in the kid's category should never request any personally identifiable information (*e.g.*, full name, and address) even if a privacy policy is specified. In Table 1, we provide a summary of the high-level statistics of kids' skills. As of April 2020, there were a total of 3,401 skills under the kids category, and 880 of these had at least one review or rating. We noted that 461 skills had developer-defined privacy policies, with 37 of these having either broken links or links to web-pages that do not contain a privacy policy.

Total skills	Skills w/ reviews	Skills w/ privacy policy	Skills w/ broken privacy policy	Total # of negative reviews
3,401	880	461	37	2,085

Table 1: Statistics of kids skills in Alexa's skills store.

We manually examined 2,085 negative reviews (*i.e.*, star ratings below 3-star) in the kids category, and summarized four common issues by user reviews: 1) frequent user complaints about skills not working. 2) collecting data from children (*e.g.*, asking for credit card information or names); 3) inconsistency of skill descriptions with their functionality; and 4) containing inappropriate content for children. Table 2 illustrates some representative critical reviews from end users. The results motivate us to further conduct a dynamic testing to empirically test the published skills and identify problematic ones in the skills store.

Skill name	User review
Guess me	"Collection of information"
ABCs	"Just want your kids data"
Whose Turn	"The initializing process required my family member names"
Chompers	"You are giving the company permission to use way too much information about your kids."
NORAD Tracks Santa	"Intrusion at its best (asking for credit card information)"
Science Kid Radio	"There are more advertisementscommercials ...."
Animal Sounds	"Asks for you to buy additional sounds"
ABC	"Creepy skill with inappropriate content for kids"
Goodnight, Sleep Tight	"Scared the kid"
Punish the kids!	"Rude to kids"
Amazon Story time	"Want your kid to hear a Boston Bombing story?"
Merry Christmas	"Played like a few seconds of Santa sounds and the rest was lame advertisements"
Chompers	"I had to explain what "sexual deviance" or some similar term was to my daughter last night"
Trivial Pursuit	"My daughter got multiple questions about alcohol and tv shows that are NOT kid appropriate"

Table 2: Selected critical reviews in the kids category.

5.1.2 Identifying existing risky skills through dynamic testing. The user review analysis reveals potential issues of policy violation in existing skills. We are curious about how existing skills conform to the security policies in the skills store. We leveraged our security expertise to assess whether a skill violates any policy by manually testing each skill. Since dynamic testing of skills is time-consuming, we examined 825 kids skills which either had a privacy policy or had a negative review. We wanted to check if they were collecting any personal identifiable information from end users which is not allowed for a kids-directed skill. We did notice certain other policy violations as well among these skills such as asking the user to engage with content outside of Alexa and promotion of other skills.

We identified 52 problematic skills with policy violations in our dynamic testing. Table 3 shows the list of these skills. In addition, we found 51 broken skills (details in Table 6 of Appendix C). Our

result shows that the lack of trustworthiness of skill certification leads to a realistic threat to VA users, especially children. We also noticed that most of the privacy policies listed in the skills store were general privacy policies provided by developers and not specifically written for the skill. The document does not provide a clear understanding to the user about what the skill is collecting or storing. Even the skills published by Amazon link to the company's general privacy policy. In many cases, it is a very long document and mostly unrelated to the skill.

Policy violation (# of skills)	Skill names
Possible collection of personal data from kids (21)	Ninja School, Dragon Palm, Loud Bird, Go Bot, Great Christmas Treasure, Wake Up Clock, Personalized bedtime stories, Who did it, Interactive Bed Time Story, Can I Wake Up, A Short Bedtime Story, Mommy-gram, Number Games, Ready Freddy, Short Bedtime Stories, Silly Stories, Story Blanks, Story World, The Bedtime Game, The Name Game (banana-fana), Who Said Meow?, Whose Turn, Clothes Forecast
Skill recommendations, advertisements and promotes end users to engage with content outside of Alexa (23)	Kid Chef, What's my dessert, Random Cartoon Facts, 6 Swords Kids, Akinator Safari, Unicorn Stories, Hansel and Gretel, Red Riding Hood, Highlights Storybooks from Bamboo, Magic Maths, Bamboo Math, Homework Heroes, 4th Grade math Skill game, Bedtime stories, Relaxing Sounds: Baby Bedtime Lullaby, Sight Words Kindergarten, The Night Before Christmas, What's Next Daily Task Helper, Wizard of Oz, Word Mess, Would You Rather Family
Offers compensation for providing reviews (1)	Kids Jokes, Knock Knocks & Riddles
Misleading description (7)	Annoying Parrot, Awesome life facts, Kids Books of the Bible, Chore list, Nursery Rhymes, Twinkle Twinkle Little Star, Chore chart, Chinese Joke

Table 3: List of skills with policy violations under the kids category in Alexa's skills store.

5.2 Possible COPPA Violations

It is possible that the third-party skills in Amazon Alexa suffer the risk of violating the Children's Online Privacy Protection Act (COPPA) rules [17], a federal legal framework to protect the online privacy of children under the age of 13. COPPA rules require that parents should be in control over what information is collected from their younger children. In 2019, YouTube paid $170 million for allegedly violating the COPPA rules, because of collecting personal information (*i.e.,* cookies) from viewers of child-directed channels, without first notifying parents and getting their consent [10].

There have been complaints made against Amazon in this regard by children's data privacy advocates [8]. Amazon claims that the kids skills available on the store do not collect any personal information from the children without parents' consent. COPPA rules require the developer to provide a privacy policy with a list of all operators collecting personal information, a description of the personal information collected, how it's used and a description of parental rights. In addition, parents must be notified directly before collecting personal information from their kids and a verified consent should be obtained from them. Amazon asks for a consent the very first time that a kids skill is enabled in the account and doesn't require one afterward for all the other kids skill enablements. This is a vague consent that does not inform the parents about what each skill is capable of collecting. This would have been admittable given that the certification system is perfect and would not let any third-party skill that violates the rules to be certified. But the kids skills published on the store are capable of violating COPPA rules. Skills that collect personal information and do not provide a privacy policy can be easily developed and certified. According to COPPA, parents must also be able to review the information that has been collected from their child with their consent and be given the authority to remove it. Moreover, COPPA requires that the contact information of the developers is provided to the parents. The information collected by Amazon from the developer when signing up for the developer account is not verified and can be easily faked. As demonstrated by our experiments, developers could certify skills that collect personal information without satisfying or honoring any of these requirements, and thus violating the COPPA regulations.

5.3 Post-Certification Vulnerability

The back-end code of a third-party skill runs on the developer's server and Alexa does not require a re-certification when a change is made in the back-end. Due to this, even if policy requirements were strictly enforced, users are vulnerable against content changing attacks after a skill has been certified. Malicious developers are able to arbitrarily change the content of responses (*e.g.*, making users exposed to inappropriate content) or questions (*e.g.*, asking users' personal information). This type of skill manipulation can lead to inappropriate content being presented to unwitting recipients or sensitive information leakage. While earlier research has mentioned about the content changing vulnerability [14, 45], crafting a phishing skill where a malicious developer can successfully store the collected sensitive information in the back-end is not that straightforward.

For a skill to collect a particular type of data, it must have the capability for data collection before the certification phase. Developers get hold of what a user has spoken (in text format) only if it matches with a sample utterance that the developer has specified. All other responses that are not matched won't be sent to the skill's back-end. For example, to collect users' address information, the developer has to add a sample utterance with a slot of type `AMAZON.PostalAddress` to one of the pre-defined intents. This cannot be added after certification as it will require a re-certification since the interaction model has changed. The malicious developer has to carefully model a custom slot with suitable training data in order to launch phishing attacks, *e.g.*, collecting passwords requires the training data including all sorts of alphabets, numerals and symbols combinations in order to accept user responses perfectly. In our experiment, we created a skill for kids with a custom slot that can accept multiple types of values (*e.g.*, first/last names and city/street names). On the submission, our skill only asked for the first name, which is acceptable by Alexa's privacy policies even if the certification process were to properly enforce policy requirements. After the certification, we changed the question to ask for several other types of personal information that could build a complete profile of the user. We were able to request and receive the full name of a user, and save the personal information to a database. To ensure research ethics in this experiment, we quickly remove all the data collected by the skills after testing.

Adversarial skill developers can exploit this vulnerability even if the issues related to certification were fixed. This can also be exploited by developers to pose as a normal authentic skill in the store for some time to earn good reviews which will boost the skill enablements (giving it a priority if users enable the skill by voice). After this, the skill can be altered with malicious content to easily

reach a higher number of users. In addition, this vulnerability opens new opportunities for malicious actors. Once an attacker is able to access the back-end code of a benign developer, the attacker can inject malicious code into the skill, with neither the developer nor the VA platform provider being notified about the change.

6 DISCUSSION

6.1 Why Lenient Skill Certification in Alexa?

There are a number of potential reasons for the leniency in Amazon's certification process for Alexa skills. There are over 100,000 skills on its skills store, but closer inspection reveals that the vast majority of these skills go unused. Being lenient with the certification process encourages developers to produce many skills, prioritizing quantity over quality. Further evidence for this motivation can be drawn from a comparison to the Google Action developer console. Google limits developers to a maximum of 12 projects on the Actions on Google console, unless the developer explicitly requests an increase in limit. In contrast, there is no such limit placed on Amazon Alexa Developer accounts. These companies also have programs in place to reward developers who develop several skills, with rewards increasing as more skills are developed. While both Amazon and Google likely do not have an ill intent through such programs, the consequence of prioritizing the growth of the respective skills store over the quality of its skills results in a certification process that insufficiently checks the submitted skills for violations.

6.2 Mitigation Suggestions

Based on our measurements and findings, we provide recommendations to help VA platform providers to enhance the trustworthiness of VA platforms.

Enforcing skill behavior integrity throughout the skill lifecycle. Our experiment shows that developers can arbitrarily change a skill's functionality after the certification, *e.g.*, an adversary replaces the benign content (which passes the security check on submission) with inappropriate content (*e.g.*, advertising extremism) in the post-certification phase. When a skill opts for an Alexa-hosted back-end, the back-end code is blocked from editing while the skill is under review. But it is unblocked after the skill is certified. To prevent content changing attacks, *a continuous certification/vetting process is required*. Whenever the developer makes a change to either the front-end or back-end, a re-certification process should be performed. This is a viable solution although it may increase the publishing latency. We also came across a large number of broken skills during dynamic testing. Skills should be periodically checked and removed from the skills store if they are broken.

Automating skill testing. Based on the observations from our 234 skill submissions, we conclude that the certification is largely done in a manual manner and through very limited voice response based testing. To strictly enforce security policies in the certification process, it is desirable to design a voice-based testing tool to automate the testing of third-party skills. For example, VA platform providers may apply deep learning techniques to train a user simulation model [18, 20, 33] to interact with third-party skills during the vetting. However, building a reliable and scalable voice-based testing tool is non-trivial. To fundamentally address the problem, VA platform providers may need to require skill developers to provide the permissions to view their back-end code. In this case, a code analysis can be performed, which could greatly increase the strength of the certification process.

6.3 Limitation

There are areas remaining where further research can help in reinforcing our findings. First, while we have taken significant efforts to measure the trustworthiness of skill certification process, our adversarial testing mainly focuses on content policy violations in skills. We do not test advanced features of skills such as the interaction with smart home IoT devices and skill connections. Second, we cannot fully scale-up the experiments of dynamic testing to identify existing problematic skills at the level of the skills store. Future work is needed to design a voice-based testing tool to automate the interaction with third-party skills. Nevertheless, we have collected strong evidence in revealing the untrustworthiness of the Amazon Alexa platform, and empirically characterize potential security risks in that platform.

7 CONCLUSION

In this work, we conducted the first comprehensive measurement on the trustworthiness of Amazon Alexa platform. We crafted 234 policy-violating skills that intentionally violate Alexa's policy requirements and all of them passed the certification. Our results showed strong evidence that its skill certification process was implemented in a disorganized manner. Through dynamic testing of 825 skills, we identified 52 problematic skills with policy violations and 51 broken skills under the kids category.

REFERENCES

[1] 2016. Toddler asks Amazon's Alexa to play song but gets porn instead. https://nypost.com/2016/12/30/toddler-asks-amazons-alexa-to-play-song-but-gets-porn-instead/. (2016).

[2] 2018. Portland Family Says Their Amazon Alexa Recorded Private Conversations. https://www.wweek.com/news/2018/05/26/portland-family-says-their-amazon-alexa-recorded-private-conversations-and-sent-them-to-a-random-contact-in-seattle/. (2018).

[3] 2018. Smart Audio Report 2018. https://www.edisonresearch.com/the-smart-audio-report-from-npr-and-edison-research-spring-2018/. (2018).

[4] 2019. Alexa and Google Home devices leveraged to phish and eavesdrop on users, again. https://www.zdnet.com/article/alexa-and-google-home-devices-leveraged-to-phish-and-eavesdrop-on-users-again/. (2019).

[5] 2019. Alexa Skills Policy Testing. https://developer.amazon.com/fr/docs/custom-skills/policy-testing-for-an-alexa-skill.html. (2019).

[6] 2019. Alexa Skills Privacy Requirements. https://developer.amazon.com/fr/docs/custom-skills/security-testing-for-an-alexa-skill.html#25-privacy-requirements. (2019).

[7] 2019. Alexa Skills Security Requirements. https://developer.amazon.com/fr/docs/alexa-voice-service/security-best-practices.html. (2019).

[8] 2019. Amazon's kid-friendly Echo Dot is under scrutiny for alleged child privacy violations. https://www.theverge.com/2019/5/9/18550425/amazon-echo-dot-kids-privacy-markey-blumenthal-ftc. (2019).

[9] 2019. Global Smart Speaker Users 2019. https://www.emarketer.com/content/global-smart-speaker-users-2019. (2019).

[10] 2019. Google and YouTube Will Pay Record 170 Million for Alleged Violations of Children's Privacy Law. https://www.ftc.gov/news-events/press-releases/2019/09/google-youtube-will-pay-record-170-million-alleged-violations. (2019).

[11] 2019. How to Improve Alexa Skill Discovery with Name-Free Interaction and More. https://developer.amazon.com/blogs/alexa/post/0fecdb38-97c9-48ac-953b-23814a469cfc/skill-discovery. (2019).

[12] 2019. Over a quarter of US adults now own a smart speaker, typically an Amazon Echo. https://techcrunch.com/2019/03/08/over-a-quarter-of-u-s-adults-now-own-a-smart-speaker-typically-an-amazon-echo/. (2019).

[13] 2019. Policies for Actions on Google. https://developers.google.com/actions/policies/general-policies. (2019).

[14] 2019. Smart Spies: Alexa and Google Home expose users to vishing and eavesdropping. https://srlabs.de/bites/smart-spies/. (2019).

[15] 2019. The Rise of Virtual Digital Assistants Usage. https://www.go-gulf.com/blog/virtual-digital-assistants/. (2019).

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

[17] Noah Apthorpe, Sarah Varghese, and Nick Feamster. 2019. Evaluating the Contextual Integrity of Privacy Regulation: Parents' IoT Toy Privacy Norms Versus COPPA. In *USENIX Security*.

[18] Layla El Asri, Jing He, and Kaheer Suleman. 2016. A Sequence-to-Sequence Model for User Simulation in Spoken Dialogue Systems. *CoRR* abs/1607.00070 (2016).

[19] Alexander Benlian, Johannes Klumpe, and Oliver Hinz. 2019. Mitigating the intrusive effects of smart home assistants by using anthropomorphic design features: A multimethod investigation. *Information Systems Journal* (2019), 1–33. https://doi.org/10.1111/isj.12243

[20] Antoine Bordes and Jason Weston. 2016. Learning End-to-End Goal-Oriented Dialog. *CoRR* abs/1605.07683 (2016).

[21] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. 2016. Hidden Voice Commands. In *USENIX Security Symposium (USENIX Security)*. 513–530.

[22] Guangke Chen, Sen Chen, Lingling Fan, Xiaoning Du, Zhe Zhao, Fu Song, and Yang Liu. 2021. Who is Real Bob? Adversarial Attacks on Speaker Recognition Systems. In *IEEE Symposium on Security and Privacy (SP)*.

[23] S. Chen, K. Ren, S. Piao, C. Wang, Q. Wang, J. Weng, L. Su, and A. 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.

[24] H. Chung, M. Iorga, J. Voas, and S. Lee. 2017. "Alexa, Can I Trust You?". *IEEE Computer* 50, 9 (2017), 100–104.

[25] Jide S. Edu, Jose M. Such, and Guillermo Suarez-Tangil. 2019. Smart Home Personal Assistants: A Security and Privacy Review. *CoRR* abs/1903.05593 (2019).

[26] Huan Feng, Kassem Fawaz, and Kang G. Shin. 2017. Continuous Authentication for Voice Assistants. In *Annual International Conference on Mobile Computing and Networking (MobiCom)*. 343–355.

[27] Christine Geeng and Franziska Roesner. 2019. Who's In Control?: Interactions In Multi-User Smart Homes. In *Conference on Human Factors in Computing Systems (CHI)*.

[28] Hang Hu, Limin Yang, Shihan Lin, and Gang Wang. 2020. A Case Study of the Security Vetting Process of Smart-home Assistant Applications. In *Proceedings of IEEE Workshop on the Internet of Safe Things (SafeThings)*.

[29] Anjishnu Kumar, Arpit Gupta, Julian Chan, Sam Tucker, Björn Hoffmeister, and Markus Dreyer. 2017. Just ASK: Building an Architecture for Extensible Self-Service Spoken Language Understanding. In *Workshop on Conversational AI at NIPS'17*.

[30] 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)*. 33–47.

[31] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, Are You Listening?: Privacy Perceptions, Concerns and Privacy-seeking Behaviors with Smart Speakers. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (2018), 1–31.

[32] X. Lei, G. Tu, A. X. Liu, C. Li, and T. Xie. 2018. The Insecurity of Home Digital Voice Assistants - Vulnerabilities, Attacks and Countermeasures. In *2018 IEEE Conference on Communications and Network Security (CNS)*. 1–9.

[33] Bing Liu and Ian Lane. 2017. Iterative Policy Learning in End-to-End Trainable Task-Oriented Neural Dialog Models. *CoRR* abs/1709.06136 (2017).

[34] Nathan Malkin, Joe Deatrick, Allen Tong, Primal Wijesekera, Serge Egelman, and David Wagner. 2019. Privacy Attitudes of Smart Speaker Users. In *19th Privacy Enhancing Technologies Symposium (PETS)*.

[35] Graeme McLean and Kofi Osei-Frimpong. 2019. Hey Alexa: examine the variables influencing the use of artificial intelligent in-home voice assistants. *Computers in Human Behavior* 99 (2019), 28 – 37.

[36] Richard Mitev, Markus Miettinen, and Ahmad-Reza Sadeghi. 2019. Alexa Lied to Me: Skill-based Man-in-the-Middle Attacks on Virtual Assistants. In *ACM Asia Conference on Computer and Communications Security (AsiaCCS)*. 465–478.

[37] Nirupam Roy, Sheng Shen, Haitham Hassanieh, and Romit Roy Choudhury. 2018. Inaudible Voice Commands: The Long-Range Attack and Defense. In *15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18)*. 547–560.

[38] Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. 2019. Adversarial Attacks Against Automatic Speech Recognition Systems via Psychoacoustic Hiding. In *Network and Distributed System Security Symposium*.

[39] Faysal 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 (WWW)*.

[40] Maurice E. Stucke and Ariel Ezrachi. 2017. How Digital Assistants Can Harm our Economy, Privacy, and Democracy. *Berkeley Technology Law Journal* 32, 3 (2017), 1240–1299.

[41] Tavish Vaidya, Yuankai Zhang, Micah Sherr, and Clay Shields. 2015. Cocaine Noodles: Exploiting the Gap between Human and Machine Speech Recognition. In *9th USENIX Workshop on Offensive Technologies (WOOT 15)*.

[42] Qiben Yan, Kehai Liu, Qin Zhou, Hanqing Guo, and Ning Zhang. 2020. SurfingAttack: Interactive Hidden Attack on Voice Assistants Using Ultrasonic Guided Wave. In *Network and Distributed Systems Security (NDSS) Symposium*.

[43] Xuejing Yuan, Yuxuan Chen, Yue Zhao, Yunhui Long, Xiaokang Liu, Kai Chen, Shengzhi Zhang, Heqing Huang, XiaoFeng Wang, and Carl A. Gunter. 2018. Commandersong: A Systematic Approach for Practical Adversarial Voice Recognition. In *USENIX Conference on Security Symposium (USENIX Security)*. 49–64.

[44] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2017. DolphinAttack: Inaudible Voice Commands. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 103–117.

[45] Nan Zhang, Xianghang Mi, Xuan Feng, XiaoFeng Wang, Yuan Tian, and Feng Qian. 2019. Understanding and Mitigating the Security Risks of Voice-Controlled Third-Party Skills on Amazon Alexa and Google Home. In *IEEE Symposium on Security and Privacy (SP)*.

[46] 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 *Network and Distributed System Security Symposium (NDSS)*.

APPENDIX A CONTENT POLICIES OF VA PLATFORMS

No.	Content Policies	Platform	Skill Submissions		Action Submissions	
			Kids (Total/Certified/Failed)	General (Total/Certified/Failed)	Kids (Total/Certified/Failed)	General (Total/Certified/Failed)
1	Trademarks, Intellectual Property and Brands	A/G	2/2/0	3/3/0		8/1/7
2	**Child-directed skills**					
2.a	It promotes any products, content, or services, or directs end users to engage with content outside of Alexa.	A	4/4/0		7/4/3	
2.b	It sells any physical products or services.	A	4/4/0		6/0/6	
2.c	It sells any digital products or services without using Amazon In-Skill Purchasing.	A	3/3/0		6/0/6	
2.d	It collects any personal information from end users.	A/G	7/7/0		25/5/20	
2.e	It includes content not suitable for all ages.	A/G	5/5/0		24/0/24	
2.f	Actions must not contain ads, including in streaming media.	G	3/3/0		15/6/9	
3	Health					
3.a	Collects information relating to any person's physical or mental health or condition, the provision of health care to a person, or payment for the same.	A/G	2/2/0	2/2/0		10/0/10
3.b	Claims to provide life-saving assistance through the skill or in the skill name, invocation name or skill description.	A	2/2/0	3/3/0		7/4/3
3.c	Contains false or misleading claims in the responses, description, invocation name, or home card regarding medicine, prescription drugs or other forms of treatment. This includes claims that a treatment can cure all diseases or specific incurable diseases. A claim can be misleading if relevant information is left out or if it suggests something that's not true.	A/G	2/2/0	3/3/0		11/10/1
3.d	Provides information about black market sale of prescription drugs.	A	1/1/0	1/1/0		6/0/6
3.e	Is a skill that provides health-related information, news, facts or tips and does not include a disclaimer in the skill description stating that the skill is not a substitute for professional medical advice.	A	3/3/0	2/2/0		2/1/1
4	Skill Recommendations, Compensation, and Purchasing					
4.a	Recommends other skills which are not owned by the same developer.	A	2/2/0	2/2/0		3/3/0
4.b	Recommends skills in Alexa's voice.	A	2/2/0	2/2/0		4/3/1
4.c	Offering compensation for using Actions/skills	A/G	2/2/0	2/2/0		5/4/1
4.d	Solicits donations from end users.	A	3/3/0	2/2/0		3/1/2
5	Advertising: Includes or otherwise surfaces advertising or promotional messaging in skill responses, notifications, or reminders.	A/G	2/2/0	4/4/0		5/4/1
6	Sexually Explicit content: Pornography and sex	A/G	3/3/0	6/6/0		9/7/2
7	Violence and Dangerous activities					
7.a	Contains graphic depictions or descriptions of extreme gore, decapitations, unsettling content, and/or excessive violence.	A/G	2/2/0	3/3/0		5/5/0
7.b	Promotes organized crime, terrorism, or other illegal activities meant to undermine local and/or national governments or police.	A/G	3/3/0	4/4/0		5/3/2
7.c	Self-harm, including instructions to carry out self-harm.	G	2/2/0	2/2/0		3/2/1
7.d	Bullying and harassment	G	3/3/0	2/2/0		3/3/0
8	Religion, Ethnicity, and Culture					
8.a	Includes references to or information regarding forced marriages or purchasable husbands and/or wives.	A	2/2/0	3/3/0		3/2/1
8.b	Purports to be able to predict gender.	A	1/1/0	2/2/0		3/2/1
8.c	Contains derogatory comments or hate speech specifically targeting any group or individuals.	A/G	2/2/0	5/5/0		5/2/3
8.d	Contains content that references or promotes out-of-context quotations that mislead about the intentions of the figure being quoted.	A	2/2/0	3/3/0		3/2/1
8.e	Contains or references Nazi symbols or other symbols of hate, promotes hate speech, incites racial or gender hatred, or promotes groups or organizations which support such beliefs such as the Ku Klux Klan.	A	3/3/0	2/2/0		3/3/0
8.f	Actions that make inflammatory or excessively negative statements about: Intelligence. Appearance or hygiene. Socio-economic status. Ethics or morality. Disability or medical condition. Criminal history. Sexual activity.	A/G	3/3/0	2/2/0		3/2/1

No.	Content Policies	Platform	Skill Submissions		Action Submissions	
			Kids (Total/Certified/Failed)	General (Total/Certified/Failed)	Kids (Total/Certified/Failed)	General (Total/Certified/Failed)
9	Emergency Services (Telecommunications). Allows the user to contact emergency responders (e.g. 911, or other emergency response products and services).	A/G				
10	Content					
10.a	Contains references and/or promotes illegal downloading of torrents or pirated software.	A/G	2/2/0	4/4/0		5/2/3
10.b	Contains specific advice on how to join an illegal organization.	A/G	3/3/0	4/4/0		5/3/2
10.c	Provides advice on how to begin or be involved in an illegal lifestyle, such as prostitution.	A	3/3/0	2/2/0		3/3/0
10.d	Gives guidance on how create or build dangerous materials (e.g., how to build a bomb, silencer, meth lab, etc.)	A/G	3/3/0	3/3/0		5/1/4
10.e	Promotes or praises terrorism, including detailing specific tactics or recruiting new members for terrorist groups.	A/G	2/2/0	5/5/0		5/3/2
10.f	Promotes use, sale, or distribution of recreational or illegal drugs.	A/G	3/3/0	5/5/0		5/3/2
10.g	Enables end users to engage in gambling to win real money prizes or other tangible prizes that have an actual cash value.	A/G	3/3/0	4/4/0		5/2/3
10.h	Promotes the sale of alcohol or tobacco, contains or references underage use of tobacco or alcohol, or promotes excessive use	A/G	3/3/0	4/4/0		5/1/4
10.i	Contains excessive profanity.	A/G	3/3/0	4/4/0		5/2/3
11	General					
11.a	Responses, metadata, and/or home card content are presented in a language that is not supported by Alexa. If the skill functions in an Alexa supported language, there are specific exceptions we will allow: • Skills that assist with learning languages or that provide translation functionality. • Skills that support religious or spiritual texts.	A	2/2/0	2/2/0		4/2/2
11.b	Contains profanity aimed at children.	A	3/3/0		2/0/2	
11.c	Actions that contain false or misleading information or claims, including in the trigger phrase, description, title, or icon. Don't try to imply an endorsement or relationship with another entity where none exists.	A/G	3/3/0	2/2/0		5/3/2
11.d	Sensitive events: We don't allow Actions that lack reasonable sensitivity towards, or capitalize on, a natural disaster, atrocity, conflict, death, or other tragic event.	G	2/2/0	3/3/0		4/3/1
11.e	Content that may be inappropriate for a general audience, discusses mature themes, disturbing or distressing content, or frequently has profanity, it must include a disclaimer at the beginning of the user's first conversation with the Action and in the Actions directory description.	G	2/2/0	3/3/0		5/2/3
12	Web Search Skills: Allows customers to search web content and does not meet all of the following requirements: • The skill must search within a specific online resource, and cannot do a general web search. • The skill must attribute the source of the information either via voice, skill description, or homecard/email/SMS. • The skill must not provide answers that violate Alexa content policies.	A	0/0/0	1/1/0		0/0/0
13	Financial					
13.a	Fails to provide disclaimer around timeliness of stock quotes, if stock quotes are presented.	A	2/2/0	1/1/0		5/4/1
14	Follow invocation name requirements					
14.a	Playing a silent sound file without a clear purpose.	G	1/1/0	2/2/0		2/0/2
14.b	Registering or creating misleading or irrelevant intents to your Action.	G	Most of our skills violated it			4/3/1
15	Spam					
15.a.	Submitting multiple duplicative Actions to the Actions directory.	G	4/4/0	2/2/0		4/0/4
Overall Summary			**119/119/0**	**112/112/0**	**85/15/70**	**185/101/84**

Table 4: Content policies [5, 13] we tested in our experiments against the skill certification process of VA platforms. *A* **and** *G* **indicates that a policy is defined by Amazon Alexa platform and Google Assistant, respectively. "Kids/General category" reflects the number of skills/actions we submitted in the Kids or General category. "Certified" denotes the number of skills/actions finally being certified, and "Failed" means the number of skills/actions that were never certified even after resubmissions. In this table, we submitted 234 skills (119 kids skills and 112 general skills) in total and got them certified. We submitted 273 policy-violating actions in total out of which 116 actions were certified and 157 failed to pass the certification. The red colour denotes a policy with high-risk, orange for intermediate-risk and green for policies with low-risk. The elements in the table that are left blank denotes that no skills/actions were submitted in that category for the specific policy.**

APPENDIX B PRIVACY REQUIREMENTS OF VA PLATFORMS

No.	Privacy Requirements	Platform	Skill Submissions		Action Submissions	
			Kids (Total/Certified/Failed)	General (Total/Certified/Failed)	Kids (Total/Certified/Failed)	General (Total/Certified/Failed)
1	Misuse customer personally identifiable information or sensitive personal information.	A				
2	Collect personal information from end users without doing all of the following: (i) provide a legally adequate privacy notice that will be displayed to end users on your skill's detail page, (ii) use the information in a way that end users have consented to, and (iii) ensure that your collection and use of that information complies with your privacy notice and all applicable laws.	A/G	9/9/0	4/4/0	25/5/20	1/1/0
3	Collect via voice or recite sensitive personal identifiable information, including, but not limited to, passport number, social security number, national identity number, full bank account number, or full credit/debit card number (or the equivalent in different locales).	A/G	2/2/0	3/3/0	0/0/0	10/0/10
4	Recite any of the following information without giving the user an option to set up a four-digit security voice code during the account linking process: (i) driver's license number, (ii) vehicle registration number, and (iii) insurance policy number.	A	0/0/0	1/1/0	0/0/0	0/0/0
5	Recite publicly available information about individuals other than the skill user without including the source of the information in the skill description.	A	0/0/0	1/1/0	0/0/0	0/0/0
6	Don't collect authentication data via the conversational interface (text or speech).	A/G	0/0/0	1/1/0	0/0/0	0/0/0
7	Promoting or facilitating the distribution or installation of malicious software.	A/G	2/2/0	4/4/0	0/0/0	5/2/3
Additional Submissions			**0/0/0**	**3/3/0**	**0/0/0**	**0/0/0**

Table 5: Privacy requirements [6] defined by VA platforms. Note that Amazon Alexa's privacy requirements and content policy guidelines have overlaps about collecting personally identifiable information. Privacy requirements 4, 5, and 6 are not covered in Table 4. Therefore, we submitted 3 additional policy-violating skills in these categories and got them certified. The rest of the skills/actions are violating policy guidelines listed in Table 4 and is therefore not a different skill. The red colour denotes a policy with high-risk, and orange for intermediate-risk.

APPENDIX C PROBLEMATIC SKILLS IDENTIFIED BY DYNAMIC TESTING

Little Red Riding Hood, baby dream sound, My Morning Helper, Sentence Inventor, Itsy Bitsy Spider Sing-Along Song, Kid's Hub, Kids' Games, Activities Collection, Kid Confident, Spiritually Divine, Talk series at MUTTER GÖTTLICH, Kidz Riddles, Lullaby Sounds, Kid Power, children calculation game, Properly Brush My Teeth, Young Picasso, Professor I.M. Smart, Talking Parrot, Kiddie Jokes, Kids Booklet, Animal Sounds, My National Zoo, 101+ Animals, Trick or Treat, Count sheeps, Laugh at My Joke, Skyla's Unicorn, June's Vocab & Quiz, Medico Help educator, Hazel's Spooky Adventures, TechFacts, Ask Santa's Elves, Make Slime, Animal sounds game, Good manners, Guess The Movie Star, Music Sandwich, My dog, My yellow name, Party Trivia, Santa Cam, Santa's Letter, Storm Workout, The Head Elf Hotline, what's your personality?, wildlife sounds, World History Quiz.

Table 6: List of broken skills under the kids category in Alexa's skills store.