Dangerous Skills Got Certified: Measuring the Trustworthiness of Skill Certification in Voice Personal Assistant Platforms

Long Cheng, Christin Wilson, Song Liao, Jeffrey Young, Daniel Dong, Hongxin Hu
School of Computing, Clemson University, SC, USA

ABSTRACT

With the emergence of the voice personal assistant (VPA) ecosystem, third-party developers are allowed to build new voice-apps[1] and publish them to the skills store, which greatly extends the functionalities of VPAs. 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 skill certification is of significant importance to platform providers, developers, and end users. Yet, little is known about how difficult it is for a policy-violating skill to get certified and published in VPA platforms. In this work, we study the trustworthiness of the skill certification in Amazon Alexa and Google Assistant platforms to answer three 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 published in their skills stores. 3) What are VPA users' perspectives on the skill certification and their vulnerable usage behavior when interacting with VPA devices? Over a span of 15 months, we crafted and submitted for certification 234 Amazon Alexa skills and 381 Google Assistant actions that intentionally violate content and privacy policies specified by VPA platforms. Surprisingly, we successfully got 234 (100%) policy-violating Alexa skills certified and 148 (39%) policy-violating Google actions certified. Our analysis demonstrates that policy-violating skills exist in the current skills stores, and thus users (children, in particular) are at risk when using VPA services. We conducted a user study with 203 participants to understand users' misplaced trust on VPA platforms. Unfortunately, user expectations are not being met by the skill certification in leading VPA platforms.

CCS CONCEPTS

• **Security and privacy → Privacy protections**.

KEYWORDS

Trustworthiness; Skill Certification; Voice Personal Assistants

[1]Voice-apps are called skills in the Amazon Alexa platform and actions in the Google Assistant platform, respectively. For the sake of brevity, we use the term *skills* to describe voice-apps including Amazon skills and Google actions, unless we need to distinguish them for different VPA platforms.

CCS '20, November 9–13, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-7089-9/20/11…$15.00
https://doi.org/10.1145/3372297.3423339

ACM Reference Format:
Long Cheng, Christin Wilson, Song Liao, Jeffrey Young, Daniel Dong, Hongxin Hu. 2020. Dangerous Skills Got Certified: Measuring the Trustworthiness of Skill Certification in Voice Personal Assistant Platforms. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security (CCS '20), November 9–13, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3372297.3423339

1 INTRODUCTION

Voice personal assistants (VPAs) 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 VPA devices as of 2019 [7]. In particular, according to Edison Research's report, 73% of surveyed owners reported that their children actively interact with at least one VPA device at home [15]. The estimated number of VPA users worldwide will reach 1.8 billion by 2021 [17]. 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. However, this convenience comes with an increasing concern about users' privacy and security. Several recent incidents highlighted the risks inherent when using VPA 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 [13]. In another case, a toddler asked Alexa to play songs but received inappropriate adult jokes instead [18].

The emergence of the VPA ecosystem allows third-party developers to build new skills. In an effort to thwart unscrupulous developers, VPA platforms have implemented a set of policy requirements [3–5, 12] to be adhered to by third-party developers. Both Amazon and Google state that a skill will be rejected or suspended if it violates any of these policies (See Appendix D). 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. A trustworthy skill certification process is of significant importance for a number of reasons to platform providers, developers, and end users. When interacting with VPA devices, users trust the VPA platform to fulfill their requests without compromising their privacy. Benign third-party developers trust the VPA platform to provide a reliable marketplace to publish skills 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 [35]) 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 the skill certification eventually undermines the VPA platform provider's competitiveness in the market.

We identify several challenges to achieving a trustworthy skill certification in VPA systems. First, the distributed architecture of VPA platforms, where a skill runs on the developer's server, poses

a challenge for trustworthy skill certification. Since the skill's code is hosted externally and out of reach of the certification process, using static code analysis to thoroughly explore a skill's behavior is not an option for current VPA systems. Second, due to the diverse nature of policy requirements defined by different VPA platforms, there is a lack of efficient certification tools to detect malicious or problematic skills. Third, both the Amazon Alexa and Google Assistant platforms allow third-party developers to update a skill's code at any time and do not require a re-certification. This can lead to code update attacks where malicious developers exploit this feature to add policy-violating content or privacy-invasive questions into the code even after a skill is certified.

In this work, we perform a set of "adversarial" measurements to understand the extent to which VPA platforms implement policy enforcement during the skill certification to prevent policy-violating skills from being published. A lenient certification would allow malicious or careless developers to publish dangerous skills in skills stores. We are curious whether policy-violating skills have existed in the current stores. We examine how external factors (*e.g.*, user review and rating) may affect the outcome of skill discovery if there are conflicts/ambiguities with names for skill invocation. This may allow adversarial developers to increase the chance of a malicious skill reaching end users. Through a user study, we aim to understand the usage habits of people using VPA services, and their concerns and expectations related to VPA platforms.

Findings. Our experimental findings reveal that the current VPA platforms (in particular the Amazon Alexa platform) have not strictly enforced policy requirements, despite claimed to the contrary. The lack of trustworthiness of the skill certification poses challenges to a VPA platform's long-term success.

- We crafted and submitted for certification 234 Alexa skills that intentionally violate 55 content and privacy policies defined by VPA platforms. We were able to get *all* of them certified. As a comparative study, we also submitted 381 policy-violating Google actions, out of which 148 actions were certified and 233 actions never passed the certification. While Google did a better job in the certification process based on our measurement, it also has potentially exploitable flaws that could lead to malicious skills being available in the skills store[2].
- To identify existing policy-violating skills in the current stores (as of July 2020), we manually tested 755 Alexa skills under the kids category, and identified 31 problematic skills with policy violations and 34 broken skills. There were only 114 actions under the families/kids category in the Google Assistant platform. We tested all of them and did find one policy-violating action.
- We analyzed post-certification vulnerabilities, and investigated how to increase the chance of dangerous skills reaching end users after they are certificated and published in the skills store. We tested Alexa's skill discovery process in an automated manner, and revealed that an adversary may potentially manipulate the skill discovery mechanism by posting fake reviews and ratings to increase the chance that a malicious skill reaches end users. This combined with the lenient certification process puts daily VPA users at a high risk.
- To understand the risks and consequences due to a lack of trustworthiness of the skill certification, we conducted a user study with 203 participants using Amazon Mechanical Turk (MTurk) crowdsourcing platform. Our survey results suggest that the users' trust on VPA platforms is misplaced, and the expectations of users are not met by leading VPA platforms.

Ethical consideration. Consideration of ethical issues is one of the most important parts in this work. Our study applies the ethical principles described in the Menlo Report [11]. Although our objective is to measure the trustworthiness of skill certification systems in VPA platforms rather than any human behavior, it is undisclosed whether the skill certification is performed by automated vetting tools or a combination of human and machine intelligence. We sought a waiver of informed consent from potential human testers (or content moderators) involved in the skill certification for the following reasons. 1) We don't study the behavior of a particular human tester, and we don't even know their identities or any contact information. 2) This research could not practicably be carried out without the waiver. Obviously, the knowledge of our measurements could influence the results. 3) Our study benefits mass customers and the community with minimal risk to subjects. 4) We have taken active steps to minimize potential risks to human testers.

We have obtained approval from our university's IRB office to conduct our experiments, and an online user study using the MTurk platform. We took the following strategies to minimize any risk to end users as well as the certification team.

- We consider the possible risk of human reviewers being exposed to inappropriate content (*e.g.*, mature content or hate speech). We classified 29 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 7 of Appendix A and Table 8 of Appendix B. For high-risk content guideline policies, we added a disclaimer "This skill/action contains policy-violating content for testing, please say Alexa/Google Stop to exit" before the malicious response, informing the human tester or user about the content to be delivered and giving an instruction on how to stop the skill.
- When a skill gets certified, we removed the policy-violating content but kept 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 developer consoles) 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 few of our skills. 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.

Responsible disclosure. We have reported all our findings about certification issues to both Amazon and Google. We have received acknowledgments from both vendors. We also shared our results to Federal Trade Commission (FTC) researchers and received recognition from them. Amazon replied that they had put additional

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

certification checks in place to further protect customers, and appreciated our work, which helps bring potential issues to their attention. Google replied that they would continually enhance the processes and technologies. It is worth mentioning that Google had immediately removed the problematic action (details in Section 5.1) and a back-end vulnerability (details in Section 5.2) we reported, and awarded us a bug bounty for reporting these issues.

2 BACKGROUND & THREAT MODEL

2.1 Alexa Platform and Third-Party Skills

We mainly focus on two mainstream VPA platforms, Amazon Alexa and Google Assistant, which are similar to each other in their structures, as illustrated in Figure 1. In addition to maintaining the directory of skills, the skills store also hosts skill metadata, such as descriptions, sample utterances, ratings, and user reviews.

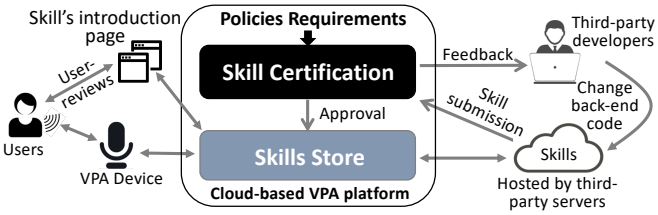


Figure 1: The distributed architecture of VPA platforms.

Front-end and back-end. A skill is composed of a front-end interaction model and a back-end cloud service (code) that processes requests and tells a VPA device what to respond. 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). 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 [34]. 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-party skills 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 code. 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). With this information, the certification team will be able to determine if a skill needs to provide a privacy policy or whether it is a kids skill.

Skill certification. To be publicly available in the skills store, each skill needs to pass a certification process, which verifies that the skill meets the Amazon Alexa's policy guidelines [3], privacy requirements [4], security requirements [5], and policies for actions on Google [12][3]. In particular, VPA platforms define strict data collection and usage policies for child-directed skills. The distributed architecture of VPA platforms gives developers more flexibility especially for those who want to protect their proprietary code and make frequent updates to the code. However, the code update after the certification does not require a re-certification. Unscrupulous developers may exploit this feature to inject malicious activities into a previously certified skill after the certification process. 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. There is a difference between Amazon Alexa and Google Assistant in terms of the skill invocation. 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 includes details such as the skill's description, the privacy policy and terms of use provided by the developer, and the user 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 invoke 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 skill candidates based on multiple factors such as the popularity of skills [8] (details in Section 5.3). Google does not require users to enable an action before using it, where users can directly say "Talk to {action name}" to invoke an action. The problem with using this method for both VPA platforms is that users do not see the details of the skill being enabled.

2.2 Threat Model

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 VPA device. This is a realistic threat model, as our empirical experiments in Section 4 show the ease of policy-violating skills being certified by both the Amazon Alexa and Google Assistant platforms, and studies in Section 5 reveal the existence of risky skills in the skills store. Our study focuses on content and privacy policy violations in skills, and we seek to understand the security threats caused by poor implementation or flawed design of VPA platforms. We assume VPA devices are not compromised. Security vulnerabilities in software, hardware and network protocols of VPA devices are out of the scope of this work.

3 RELATED WORK

There has been a number of studies showing that users are concerned about the security/privacy of VPA devices [20, 21, 23, 28, 31, 32, 39, 40, 49, 53]. Lau *et al.* revealed that privacy concerns can be the main deterring factor for new users [36]. Edu *et al.* [29] categorized

[3]To be concise, we use *policy requirements* to refer to both content policy guidelines and privacy requirements specified by both Amazon and Google. Amazon Alexa's security requirements [5] 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.

common attack vectors (*e.g.*, weak authentication, weak authorization, data inference) and their countermeasures in VPA ecosystems. Due to a lack of proper authentication from users to VPA devices, an adversary can generate hidden voice commands that are either not understandable or inaudible by humans [24, 25, 42, 43, 45, 46, 48, 50] to compromise speech recognition systems. The corresponding defense mechanisms include continuous authentication [30], canceling unwanted baseband signals [50], correlating magnetic changes with voice commands [26], and user presence-based access control [37].

On the other hand, the openness of VPA ecosystems brings with it new authentication challenges from the VPA to users: a malicious third-party skill may impersonate a legitimate one. Kumar *et al.* [35] 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 [51]. This is because the longest string match was used to find the requested skill in VPA platforms. Zhang *et al.* [51] 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 [52] is a black-box mutation-based fuzzing tool to systematically discover misinterpretation-prone voice commands in existing VPA platforms. Mitev *et al.* [41] presented a man-in-the-middle attack between users and benign skills, where an adversary can modify arbitrary responses of benign skills. Shezan *et al.* [44] developed a natural language processing tool to analyze sensitive voice commands for their security and privacy implications. Hu *et al.* [33] performed a preliminary case study to examine whether the 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. Liao *et al.* [38] conducted data analytics to measure the effectiveness of privacy policies provided by developers in VPA platforms. SkillExplorer [19] is a dynamic testing tool to explore skills' behaviors and detect privacy violations in skills. The authors tested 28,904 Amazon skills and 1,897 Google actions, and identified over 1,000 skills request users to provide personal information without following developer specifications.

Despite the recent progress, existing research has mainly focused on addressing challenges in open voice/acoustic interfaces between users and speech recognition systems of VPA devices. While dangerous skills (*e.g.*, voice squatting or masquerading attacks) have been reported by existing research [2, 16, 35, 51], these dangerous skills do not necessarily violate VPA's policy requirements at the certification phase. Little is known about how difficult it is for a policy-violating skill (*e.g.*, with malicious content) to get certified and published by VPA platforms. In this paper, our focus and methodology are *different from existing research efforts*. We aim at comprehensively assessing the trustworthiness of skill certification in leading VPA platforms, and characterizing threats between third-party skill developers and VPA platforms.

4 MEASURING THE TRUSTWORTHINESS OF SKILL CERTIFICATION

In this section, we conduct "adversarial" experiments to measure the trustworthiness of skill certification processes of the Amazon Alexa and Google Assistant platforms, *i.e.*, verifying whether a policy-violating (dangerous) skill can pass the verification. Detailed ethical discussions are presented in Section 1.

4.1 Experiment Setup

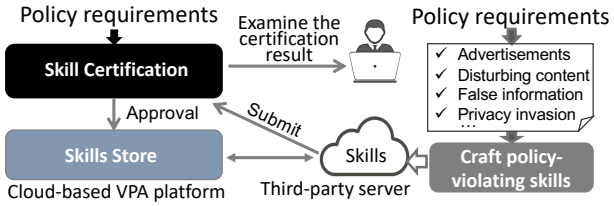


Figure 2: Experiment setup for measuring the skill certification process of VPA platforms.

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 VPA platforms, and examine if it gets certified and published to the store or not. Figure 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 VPA platforms. Amazon has defined content policy guidelines [3] which are categorized into 14 main sections and 7 specific privacy requirements [4]. We use the Amazon's policy classification as a reference, and compile a list of 55 policies with consideration of Google's policy requirements [12], as shown in Table 7 and Table 8 in the appendices.

Skill Development and Submission. In terms of skill development, both Amazon and Google are similar to each other in their structures. We list three notable differences between them. 1) Amazon requires developers to provide a privacy policy only when the skill explicitly states that it collects personal data. For other skills, it is not mandatory to provide one. While Google requires every action to provide a privacy policy on submission. 2) Duplicate skill names (*i.e.*, multiple skills have the same or phonetically similar name) are allowed by Amazon. However, Google restricts that all actions must have a phonetically unique name. 3) Amazon has no limits on the number of skills that each developer account can deploy. While Google only allows each developer account to deploy up to 12 action projects.

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.*, toxic content or advertisement). Our experiments were conducted from April 2019 to July 2020.

For the Amazon Alexa platform, we crafted 234 skills for certification, including 115 skills in general categories and 119 kids skills. 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. 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 these skills. For the Google Assistant platform, we submitted 381 policy-violating actions (201 actions in the general category and 180 kids actions) to observe if they can pass the certification.

Policy Violations in Our Experiments. Each testing skill violated one of the 55 policies listed in Table 7 and Table 8. We briefly summarize the policy violations in our crafted skills.

- The content policies as shown in Table 7 of Appendix A mostly focus on the content in a skill being delivered to users. These involve disturbing content, false information, profanity, etc. It also restricts the collection of health related information. Our crafted skills had advertisements, alcohol and tobacco usage promotions, etc. Some skills also include offered shopping services for physical products with payments accepted through a bank transfer rather than the Alexa in-skill purchasing. We used trademarked logos as the icons for a skill to violate the guideline regarding trademarks (*i.e.*, policy 1 in Table 7).
- Table 8 of Appendix A lists the privacy requirements defined by VPA platforms, which actually have overlaps with the content policy guidelines in Table 7. These policies mostly focus on the collection of data, the method of collection and the information being provided to the users about the data collection. We built skills that request particular information from users through voice and the collected information was also read back to them to confirm their input. For example, a travel skill that would collect the users' passport number to check if he/she requires a visa or not. These skills were also capable of storing the information collected in a DynamoDB database. Note that all the collected data were immediately deleted to safeguard the users privacy during the skill certification in our experiments.
- In particular, Table 7 contains 6 policies for child-directed skills (*i.e.*, policies 2.a to 2.f in Table 7). We built interactive story skills to collect personal information from children. We mentioned about personalizing the story based on names in the skill description. We did not specify that we were collecting personal information in the Privacy & Compliance form, and did not provide a privacy policy for these skills either. Since Google requires each action providing a privacy policy, we provided a privacy policy with the content "We don't collect any information" for each action. Skills were submitted to violate the other policies in Table 7 as well. In addition, we re-submitted all the skills that we developed for violating the general content guidelines to the kids category in order to measure whether the policy enforcement for kids-directed skills is stricter than skills in other categories.

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 received for our skill submissions demonstrated the existence of policy violations in our testing skills.

4.2 Overview of the Measurement Results

	Amazon Alexa Platform		Google Assistant Platform	
	General skills	Kids skills	General actions	Kids actions
Submitted	115	119	201	180
Certified	115 (100%)	119 (100%)	104 (52%)	44 (24%)

Table 1: Skill certification results.

Table 1 shows the overview of our skill certification results. For the Amazon Alexa platform, 115 skills in general categories and 119 kids skills were submitted. Surprisingly, we successfully certified 193 policy-violating skills on their first submissions, and 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 detailed experiment results 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. To work around these rejections, we modified the back-end code by creating a session counter so that the policy-breaking response/question would be selected only when the counter reached a certain threshold, *e.g.*, after the 5th 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 skills were certified after three or more submissions. Using this simple method, we managed to get a total of 234 skills with policy violations certified.

For the Google Assistant platform, we submitted 381 policy-violating actions in total, out of which 148 actions were certified and 233 actions eventually failed to pass the certification, even after we applied the same bypassing technique used for the Amazon Alexa platform. There were 180 actions that violate policy guidelines specific to actions intended for children, out of which 44 actions did pass the certification. While most of these kids actions were rejected, we were still able to certify actions that violated privacy requirements and had advertisements. From our experiment results, we could infer that Google has a strict certification procedure for actions for families but it is still not completely capable of rejecting

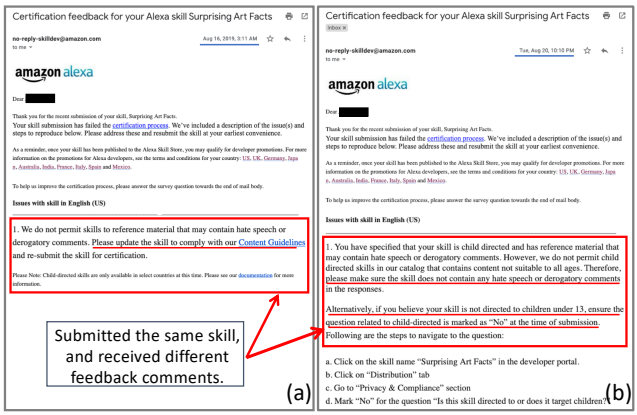


Figure 3: Inconsistent certification feedback emails from the Amazon Alexa platform.

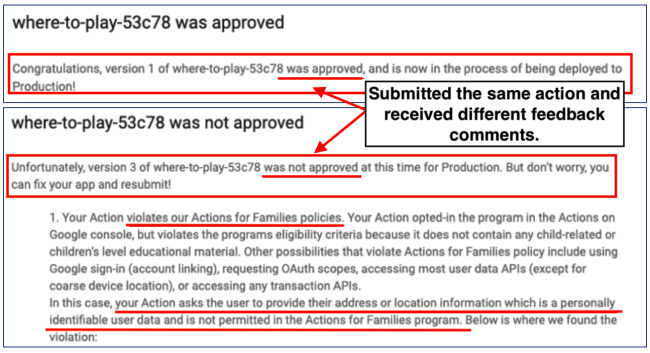


Figure 4: Inconsistent certification feedback emails from the Google Assistant platform.

all malicious actions. Finally, we got 104 actions in the general category certified, and 97 actions could never pass the certification.

Overall, our measurement results show that the current VPA platforms (in particular the Amazon Alexa platform) have not strictly enforced policy requirements in their skill certification processes, despite claimed to the contrary. Google relatively did a better job based on our measurement (especially for the kids actions), but it is still not perfect and it does have potentially exploitable flaws that allow developers to publish policy-violating skills in the store.

4.3 Key Observations

We summarize our key observations that lead to the untrustworthiness of skill certification in two mainstream VPA platforms.

4.3.1 Inconsistency in checking. We have received varied feedback from the certification team after submitting the same skill multiple times, which happened to both VPA platforms. In multiple 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 particular, for the Amazon Alexa platform, the largest amount of bulk certifications we were able to achieve was 20 Amazon 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. 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's certification team about the developer's suspicious intentions. 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 skill certification is not a well-organized systematic process.

Figure 3 shows two certification feedback emails from the Amazon Alexa platform. 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 7 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 content in all submissions, feedback received was different for each submission. The first rejection (see Figure 3(a)) stated that no skills are allowed to have such content. On the second submission, the rejection feedback (shown in Figure 3(b)) stated that a kids skill cannot have such content. On the third submission, the skill was certified. These feedback comments show an inconsistency in the certification process. An example of inconsistency in feedback from the Google Assistant platform is shown in Figure 4. The same action was submitted twice, but we received different certification results (one got rejected, and the other one got certified).

4.3.2 Limited voice checking. This is the main reason that we could easily bypass the skill certification. We observed that the certification process tested the skill only for a limited number of times (normally less than three voice responses). As a result, policy-violating content could be easily concealed from the certification team's testing range. There were multiple cases where the skill that provided a response with a policy violation in the first session itself was accepted. One live example of the certified Amazon skill with policy violation *on its first response* is shown in Figure 8 of Appendix E. Some rejections of Alexa skills were based on the information provided in the distribution section of a 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.

For the Amazon Alexa platform, 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 certification process neglects this or it did not draw an 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. 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. And the skill testing was done from a

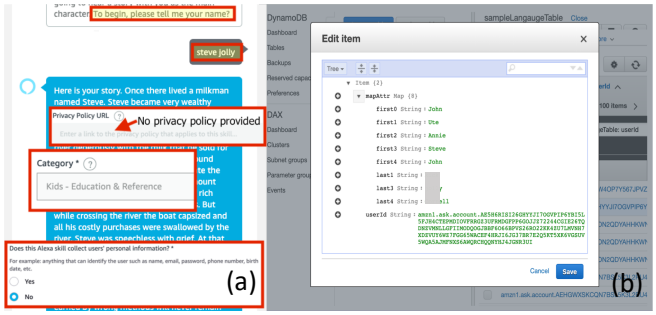


Figure 5: (a) A certified kids skill collecting personal information, and it has the policy violation on its first response. (b) Data that the skill stored in DynamoDB database.

user's perspective with checks conducted based on the information and access of the skill available to the users. Although Google could identify spam actions and reject them if an action is a duplicate of another action in the directory, the limited voice checking is not a unique problem to Amazon Alexa, and we also observed the similar issue for the Google Assistant platform.

4.3.3 Overtrust placed on developers and negligence during certification. This observation is mainly for the Amazon Alexa platform. From our experiments, we understood that the Amazon Alexa platform 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 skill 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.

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}` `{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. To ensure the research ethics, we have deleted any collected information after the skill certification. Figure 5 shows a certified kids skill collecting user names. The skill had no privacy policy and the Privacy & Compliance form filled by the developer denied collecting personal information from users. Note that the names shown in Figure 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.

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 in the Amazon Alexa platform. On the contrary, for some Google actions that contain a disclaimer "This action contains policy-violating content for testing, please say Stop to exit", we received feedback "Your action violates our user experience policy. Your action appears to be a test action based on your action's simulation. Please redeploy your Action once it is in a more complete state".

In addition, Amazon Alexa and Google Assistant platforms exhibited different patterns in terms of the certification feedback time, where details can be found in Appendix F. From the timestamps when we received certification feedback emails, as shown in Figure 9 of Appendix F, the Amazon Alexa's certification team possibly works based out of the US timezone.

5 ANALYZING POST-CERTIFICATION RISKS

The results in Section 4 reveal that VPA platforms have not strictly enforced the security policies in the certification process. It provides opportunities for malicious developers to trick the certification process, and thus placing end users in a vulnerable position. We next ask the question, "whether there exist policy-violating (*e.g.*, collecting personal information from users) or problematic skills in the current skills stores because of the lenient skill certification processes". However, it is non-trivial to test all (more than 100,000) published skills in the two VPA platforms to find policy violations, due to the lack of an automated testing tool. Instead, we focus on kids' skills in our analysis because VPA platforms specify more stringent policies in the kids category than the other categories. We also discuss other consequences of a lenient skill certification, including possible COPPA violations, post-certification vulnerabilities, and the risks that unscrupulous developers increase the chance of malicious skills reaching end users.

5.1 Can We Find Policy-Violating Kids Skills?

In Table 2, we provide the high-level statistics of kids' skills. For the Amazon Alexa platform, as of July 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 privacy policies, with 37 of these having either broken links or links to webpages that do not contain a privacy policy. Only 114 actions were available in the families/kids category of the Google Assistant platform, and 80 of these actions have at least one review or rating. All the actions provide a privacy policy. The number of available Google actions is very less compared to that of the skills available in Alexa's skills store. This low number of actions and no broken privacy policy can be a result of Google's stricter certification system. Note that the skills/actions we submitted are not counted in Table 2. In Appendix G, we also list some representative critical user reviews from the Amazon Alexa's skills store.

Experiment Setup. We are curious about how existing skills conform to the policy requirements in the skills stores. Due to the fact that skills' code is not readily available to users (and also out

Platform	Total skills	Skills w/ reviews	Skills w/ privacy policy	Skills w/ broken privacy policy
Amazon	3,401	880	461	37
Google	114	80	114	0

Table 2: Statistics of kids skills in two VPA platforms.

of reach of the certification process), using static code analysis to explore a skill's behavior is not an option. To identify existing policy-violating skills, we need to interact with them to collect their responses (*e.g.*, statements or questions), and then check if these responses violate any policy requirement. Considering manually testing skills is time-consuming (we discuss the dynamic analysis of skills in Section 7), we used the privacy policy and user ratings as filters to narrow down the skills that we tested. This is because Amazon Alexa requires a mandatory privacy policy only for skills that collect personal information. Therefore, we assumed that skills with a privacy policy provided might be collecting information. On the other hand, low-rating user reviews may reflect potential issues of policy violation in existing skills. We examined 755 Alexa skills which either had a privacy policy (461 skills) or a low rating (*i.e.*, 294 skills with star ratings below 3-star) using the Alexa developer console simulator and recorded all the responses. For the Google Assistant platform, we manually tested all the 114 actions listed in the families/kids category. We developed a simple natural language processing (NLP)-based policy violation detector which analyzes skills' outputs to detect policy-violating skill behaviour. We defined different sets of keywords to detect violations of different policies. We focused on detecting personal data collection, illegal and violent content, advertisements and promotions that direct users to engage with content outside of a skill. We used the SpaCy library [9] to obtain nouns and verbs in skills' responses. The detection is based on the similarity comparison of phrases between a skill's response and the policy-specific keywords.

Experiment Results. For the Amazon Alexa platform, our detector reported 33 problematic skills with policy violations. We leveraged our security expertise to assess the results, and found only 2 false positives (6%). Due to the lack of the ground truth number of policy-violating skills, we did not measure the false negative rate. We mainly focused on the existence of policy-violating skills rather than the exact number of them. Table 3 shows the list of these skills with different types of policy violations. We found 18 skills asking sensitive information such as user name, zip code, and age. For example, "Let's Read" and "Mrs. Howard Classroom" asked "please say your name" and "would you tell me your name", respectively. There were 9 skills containing advertisements and promoting users to engage with content outside of Alexa. For instance, the skill "Wizard of Oz" redirected users to the website daysfly.com. 4 skills explicitly asked users to give 5 stars ratings. In addition, we found 34 broken skills (details in Table 9 of Appendix C).

For the Google Assistant platform, we only found one problematic action called "Smallfoot" that asked the user's name. In addition, it's privacy policy URL does not lead to a webpage that contains a privacy policy but rather to a page advertising the company's products. After we reported our findings to Google, this action has been removed from the store. Our result shows that the lack of trustworthiness of skill certification leads to a realistic threat to VPA users, especially children.

Policy violation (# of skills)	Skill names
Possible collection of personal data from kids (18)	**Guess the animal sound**, **Say Please**, **Get infor about the actual date**, **Let's be Friends**, **Ready Freddy**, Wake Up Clock, Let's Read, Mrs.Howard Classroom, Intone riddle, Loud Bird: Story Theater Enabled, Three Wishers Cairo Book One, Three Wishes Cairo Book 2: Story Theater Enabled, Dragon Palm: Story Theater Enabled, Stamp My Passport, Who Stole The Cookie From The Cookie Jar, Guess My Age, Community Helpers, Animal Noises Quiz
Skill recommendations, advertisements and promotes end users to engage with content outside of Alexa (9)	Animal Game for Kids - Play and Learn, Wizard of Oz, Red Riding Hood - Interactive story for kids, Divine Office Magnificat, Ask My Kid - Student Learning - AskMyClass at Home, My Chandigarh University, Campfire Stories, Naughty Or Nice Quiz, Dinofun Dinosaur Facts
Asks for positive rating (4)	Kids Animal Sounds, Lemonade Stand, What if...?, Beewhiz

Table 3: List of skills with policy violations under the kids category in Alexa's skills store (as of July 2020).

Possible COPPA Violations. For the kids skills that can collect personal information in Table 3, it is possible these Alexa skills suffer the risk of violating the Children's Online Privacy Protection Act (COPPA) rules [22], a federal legal framework to protect the online privacy of children under the age of 13. 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. However, we found 5 Alexa skills ("Guess the animal sound", "Say Please", "Get info about the actual date", "Let's Be Friends", and "Ready Freddy" highlighted in Table 3) collecting personal data without providing a privacy policy. 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 skill 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.

5.2 Code Update Vulnerabilities

Both the Amazon Alexa and Google Assistant platforms do not require a re-certification when a change is made in the back-end code of a skill. Malicious developers may exploit this feature to arbitrarily change the content of responses (*e.g.*, making users exposed to inappropriate content) or questions (*e.g.*, asking users' personal information) in a certified skill, which we call the code

update vulnerability. Even if policy requirements were strictly enforced, users are vulnerable against code update attacks after the skill certification. While earlier research has mentioned about the code update vulnerabilities [16, 51], exploiting such vulnerabilities to craft a privacy-invasive skill which can *successfully capture and store the sensitive information in the back-end* is not that straightforward. This is because, even attackers can modify a skill's back-end code to fool users to provide sensitive data, the skill may not be able to capture the sensitive data in the back-end, which can be filtered out by the natural language understanding module [52] in VPA systems. For a skill to collect a particular type of data, it must have already had the capability for capturing the specific type of data 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. Therefore, the malicious developer has to carefully model a custom slot with suitable training data (*i.e.*, custom slot values provided by developers) in order to capture a particular type of data which is not declared before the certification process. For example, collecting passwords requires the training data of a custom slot including all sorts of alphabets, numerals and symbols combinations in order to accept user responses perfectly. However, a custom slot with diverse types of slot values is suspicious for stealthy data collection.

In our experiment, due to the lenient skill certification, we were able to publish an Alexa kids skill 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 VPA's privacy requirements even if the certification process were to properly enforce policy requirements. We aim to measure whether the certification team could detect suspicious custom slots that accept multiple types of sensitive values (*e.g.*, first/last names and city/street names) in a skill. We were able to effortlessly get this skill certified. After the certification, we changed the question to ask for several other types of personal information that could potentially build a complete profile of a user. We were able to request and receive the full name of a user, and save the personal information to a database. To ensure the research ethics, we immediately removed all the data collected after our internal testing. The skill analytics data in the Alexa developer console confirmed no actual users had been affected.

Similar to Alexa skills, code update attacks can be performed on Google actions as well. The back-end of an action can be updated after the certification and it does not require a re-certification. The Google Assistant platform has entities that denote the type of data that can be collected by a placeholder which works similar to slots and slot types in the Amazon Alexa platform. The entity `sys.any` can be used to store all types of data. By changing the content of the question in the back-end after certification, an action that uses this entity can collect any type of information including personal information from the voice response provided by users. Thus, different from Alexa skills, the developers do not have to create a custom entity/slot to collect personal information for Google actions. We have reported the code update vulnerabilities to both vendors. We noticed that the entity `sys.any` has been removed from the Google action developer console.

The code update vulnerability can also be exploited by developers to publish a seemingly benign 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 may open 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 VPA platform provider being notified about the change.

5.3 Is It Possible To Increase the Chance of Dangerous Skills Reaching End Users?

We have demonstrated that malicious third-party skills can easily bypass the certification process and become available in the skills store (especially for the Amazon Alex platform). However, publishing policy-violating skills in the store only provides a possibility for malicious developers to launch attacks. The success of such attacks essentially depends on the skill discovery/recommendation algorithm implemented on VPA platforms. For example, a voice-squatting skill [35, 51] can only impersonate a legitimate one if it is recommended/invoked by VPA platforms to fulfill a user's request. We examine how different factors (*e.g.*, user review and rating) may affect the outcome of skill discovery if there are conflicts/ambiguities with names for skill invocation. Since the Google Assistant platform does not allow the duplicate naming, we only investigate Amazon Alexa's skill discovery process. If external factors and parameters can influence the skill discovery process, adversarial developers may manipulate these factors (*e.g.*, posting fake reviews) to increase the chance of a malicious skill reaching end users.

Experiment Setup. We first selected 10 groups of third-party Alexa skills that have identical invocation names in each group, and obtained their skill IDs, user reviews, star ratings, and developer information. We aimed to observe whether a particular skill is constantly discovered by Amazon Alexa when invoking a skill that has a duplicate name with other skills. We developed a tool using the Selenium WebDriver (SWD) [14] to automate the skill discovery testing in the Alexa developer console. For each group testing, we started off with a newly formatted Alexa device, where no skills have been enabled. SWD was set to enter the Alexa developer console by automatically inputting our developer credentials, and then enable the device logging. Next, SWD was used to enter the skill invocation into the Alexa simulator. The device log captured all communication between the Alexa simulator and the invoked skill. Then, we retrieved the skill ID from the device log. Interestingly, from the results we encountered, the skill that was enabled was the one with the highest number of reviews/ratings in the store. Then, SWD was set to disable that skill. On the next attempt at opening a skill with the invocation name, a new skill with the same invocation name was enabled. This was the skill that was second in terms of ranking based on the number of ratings/reviews. For invocation names that have more than 2 skills published on the store, we did the above process repeatedly. We logged into two other Amazon accounts and repeated the same experiment. We found that the

skills that were enabled and their ordering were the exact same on every account.

Skill Name	Developer	Rating	Total Reviews	Order
Zipcode lookup	GPDL Industries	3.9	15	1
	Bachelor Pad Development	2.9	10	2
Yoga music	Zen apps	3.8	21	1
	App316.com	4.4	6	2
Yo mama jokes	Franks & Beans	4.2	2024	1
	Per4mnce Software	4.1	184	2
Yes man	Dawson Foushee	4.0	19	1
	Naitian Zhou	3	1	2
World war two trivia	Mustang D	4.2	146	1
	Jobo's Rum	1.7	3	2
World history	Appbly.com	3.5	91	1
	Mudit Surana	1.5	3	2
Work time tracker	Alex	2.6	11	1
	estherleah	5.0	1	2
Who goes first	Christopher Pierce	3.3	61	1
	Novavux	3.7	5	2
	Nirav	5.0	1	3
	ProfDeCube	0	0	4
Weird facts	Venkateswara Rao	4.0	220	1
	Baujaco	3.3	10	2
	Supermann	4.5	5	3
Simon Says	Ed Lea	3.5	7255	1
	David Suarez	3	953	2
	Drew Firment	2.5	257	3
	Dokka Inc.	3.5	181	4
	clara-jr	3.3	13	5

Table 4: List of skills in our measurements.

Experiment Results. Table 4 shows the list of skills in our measurements to understand the skill discovery process. Our measurements reveal that: i) the priority first goes to the skill that is already enabled on the device; and ii) the skill discovery is likely a function of user reviews and ratings (but it is hard to demystify the exact skill ranking algorithm, which takes many factors into account including the quality and usage of skills [8]). The second priority goes with the skill that has the highest number of reviews/ratings. The user's opinion about which skill to enable is neither asked for nor is the user properly informed about the particular skill other than to just state the "public name" of the skill when it is enabled. After a malicious skill (*e.g.*, a voice-squatting skill) is certified and published in the skills store, an attacker may post fake positive reviews and ratings to increase the success rate of skill hijacking. Since there are very few skills that have a sizable number of reviews, in many cases the number of required fake reviews that an attacker needs to post can be considerably less. Also, the users would not know which skill they are using unless they check their Alexa companion apps on their mobile devices. We recognize that Amazon's fraud detection mechanisms may catch fake reviews. However, the longstanding problem of preventing the manipulation/inflation of usage statistics makes it probable for the adversarial developers to manipulate the skill discovery [47]. For example, Amazon stated that there exists fake reviews designed to evade detection, and review abuse continues, despite the company's efforts [10]. The poor certification combined with potentially manipulable skill discovery puts VPA users at high risk.

6 USER STUDY

Different from existing user studies that focus on understanding users' security and privacy attitudes towards VPA [21, 23, 27, 32, 39], our study investigates user behaviour about exploring new skills, when encountering something inappropriate, and their trust to VPA platforms. We are also interested in how parents manage their kids' usage of VPA devices. Answers to these questions will help us understand the risks and serious consequences due to a lack of trustworthiness of skill certification in VPA platforms.

6.1 Methodology

We conducted an online study using the Amazon Mechanical Turk (MTurk) crowdsourcing platform. We had two groups of users: Amazon Alexa users and Google Assistant users. There were 101 valid participants who took part in our survey for Alexa users and 102 valid participants who took part in the survey for Google Assistant users. Out of these, 36 of the Alexa users and 31 of the Google Assistant users were parents who have kids under the age of 13. All the participants in our survey were MTurk masters residing in USA and have an acceptance rate of at least 98%. These filters were added to reduce the amount of junk data that we may have collected. All participants were initially presented with a consent form approved by our university's IRB office. Participants who did not consent to the form were denied to proceed with the study. The participants were paid $1.00 for completing the study.

6.2 Response Analysis

6.2.1 Usage habits questions. Table 5 shows the survey responses from Amazon Alexa and Google Assistant users. We wanted to know which method users chose to explore and enable new skills on their devices. 42% of the users said that they enable skills by voice while the other 58% browse the store. This was asked only to the Amazon users because Google does not require users to enable an action before using it. When something inappropriate is encountered, 69% of Amazon users and 72% of Google users claimed to provide a review for the skill on the skills store. This result suggests that users' negative feedback may help us in identifying the existing policy-violating skills (In Appendix G, we understand users' common concerns through a preliminary user review analysis). When asked if they trusted Amazon and Google to detect and filter out the inappropriate content, *71% of Amazon users and 90% of Google users* responded with definitely yes or probably yes. About *95% of participants* from both the groups have the opinion that all skills and actions must be thoroughly tested and checked by Amazon and Google. *91% of Alexa users and 94% of Google users* expect this to be done before the skill/action is published to the store rather than after the skill/action is certified and available to the public. These statistics imply that *the trust on VPA platforms are actually misplaced, and the expectations of the users are not met by Google and Amazon in this regard.* When asked if the participants knew what personal data the skills and actions they currently use are capable of collecting from them, only around 10% of users responded with definitely yes. This partially implies that the skills are not doing a good job in informing the users about their data collection and storage practices. About 50% of users from both groups believed that the privacy policies were only focused on preventing lawsuits and are not aimed at providing a clear picture and useful information to the users. When asked if the Alexa users check their activities on the Alexa smartphone app, 45% of participants claimed to check

it at least half the time. While this is a good method to monitor the conversations happening between the VPA device and the family members, it can be pretty tiresome to do it on a regular basis. Also, we noticed an inconsistency in the responses that we received from the device and the ones logged in the activities tab on the Alexa smartphone app. This can be a bug that needs to be rectified (We have reported this bug to the Amazon security team). It does defeat the purpose of the activities tab at this stage.

Question	Response	Amazon users	Google users
How do you enable new skills?	Browse the store	58.42%	-
	Enable through voice	41.58%	-
Would you add a negative review for a skill/action if you encounter something inappropriate?	Definitely Yes	23.76%	26.47%
	Probably Yes	45.54%	45.10%
	Probably not	27.72%	23.53%
	Definitely not	2.97%	4.90%
Do you trust Amazon/Google to detect inappropriate content?	Definitely Yes	14.85%	33.33%
	Probably Yes	56.44%	56.86%
	Probably not	22.77%	9.80%
	Definitely not	5.94%	0
Do you think that the skills/actions should be thoroughly tested and checked by Amazon/Google ?	Definitely Yes	66.34%	58.82%
	Probably Yes	29.70%	35.29%
	Probably not	2.97%	4.90%
	Definitely not	0.99%	0.98%
When should malicious skills/actions that violate policies be rejected by Amazon/-Google?	After Publishing to store	8.91%	5.88%
	Before Publishing to store	91.09%	94.12%
Do you know what personal data the skills/actions you use are capable of collecting from you?	Definitely Yes	10.89%	8.82%
	Probably Yes	28.71%	43.14%
	Might or might not	23.76%	32.35%
	Probably not	24.75%	15.69%
	Definitely not	11.88%	0
What do you think the privacy policies currently being provided with services are actually for?	Inform users about data collection and storage practices	52.48%	46.08%
	Prevent lawsuits	47.52%	53.92%
How often do you check the activity on your Alexa app?	Always	10.89%	-
	Most of the time	18.81%	-
	About half the time	14.85%	-
	Sometimes	32.67%	-
	Never	22.77%	-

Table 5: Survey responses: general questions

6.2.2 Parents specific questions. We asked a few questions specifically to parents who have kids under the age of 13. Table 6 lists the survey responses. 61% of Amazon users in our group have provided permission to use kids skills and 25% were not sure if they did, and Google users show similar proportions. When asked if they use the Amazon Freetime subscription or the Google Family Link, only around 50% said yes from both the groups. While such mechanisms provide better control to the parents over how their children interact with VPA devices, the number of users making use of the paid subscription is less. Only 33% of Alexa users and 48% of Google users extensively test the kids skills at least half the time before enabling it for them. This shows the trust that users have on Amazon and Google. Also testing every skill extensively before enabling is not feasible from a user's perspective and is something that should be done by the skill certification team before publishing it. 33% of Alexa users and 58% of the Google users in our survey had a dedicated VPA device for their kids. This shows that the parents might not be around with the children to monitor the interactions when the kids use the device. We noticed that the majority of users feel uncomfortable with promotions and advertisements targeted at children. In addition, when asked to review the comfort level of 4 of the actual responses we had added in the skills we submitted (which are not shown in Table 6), only a maximum of 20% of users from both groups were at least somewhat comfortable with each of the responses being delivered to their children. This demonstrates that the responses in the skills that were certified by Amazon and Google were indeed policy-violating and disturbing to children.

Question	Response	Amazon users	Google users
Have you provided the permission to enable kid's skills on your devices?	Yes	61.11 %	54.84%
	No	25%	35.48%
	Not sure	13.89%	9.68%
Do you use Amazon Freetime /Google Family Link?	Yes	50%	48.39%
	No	44.44%	35.48%
	Not sure	5.56%	16.13%
Do you extensively test kid's skills/actions before enabling it for them?	Always	8.33%	9.67%
	Most of the Time	16.67%	22.58%
	About Half the Time	8.33%	16.13%
	Sometimes	44.44%	35.48%
	Never	22.22%	16.12%
Do you have a dedicated smart speaker device for your kids?	Yes	33.33%	58.06%
	No	66.66%	41.94%
How comfortable are you with Alexa/the Assistant promoting any kind of products or websites to your child?	Extremely Comfortable	5.55%	16.12%
	Somewhat Comfortable	22.22%	29%
	Neither Comfortable nor Uncomfortable	25%	19.35%
	Somewhat Uncomfortable	27.77%	22.58%
	Extremely Uncomfortable	19.44%	12.9%

Table 6: Survey responses: parents specific questions

7 DISCUSSION

7.1 Why Lenient Skill Certification?

There are a number of potential reasons for the leniency of the skill certification processes (in particular for the Amazon Alexa platform). There are over 100,000 Alexa skills on its skills store, but closer inspection reveals that the vast majority of these skills go unused. Being lenient with the skill 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 Assistant developer console. Google limits developers to a maximum of 12 action projects, 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 skill certification process that insufficiently checks the submitted skills for policy violations.

7.2 Mitigation Suggestions

Based on our measurements and findings, we provide recommendations and actionable mitigation strategies to help VPA platform providers to enhance the trustworthiness of skill certification.

Training for the certification team. The inconsistency in various skill certifications and rejections have led us to believe that the skill certification largely relies on manual testing. And the skill certification team may not be completely aware of the various policy requirements and guidelines being imposed by VPA platforms. This is especially due to the fact that we were able to publish skills that had a policy violation in the first response. From the certification

feedback time we received for Alexa skills as shown in Figure 9 of Appendix F, it appears that the certification was done by a team outside the U.S. In this case, it is very important that the team is made aware of related policies and regulations for different places where the VPA services target at, *e.g.*, COPPA [22] and General Data Protection Regulation (EU GDPR) [6]. A better understanding and proper training of the policy guidelines in place should be given to the certification team to prevent the inflow of policy-violating skills to the skills stores.

In-depth checking during the skill certification. Given the fact that skills' back-end code is not available, the certification team should conduct an in-depth checking during the skill certification. Our measurement results (in particular for the Amazon Alexa platform) indicate that the verification of a skill was mainly done in a manual manner and through very limited voice response based testing, without checking the skill's interaction model (*i.e.*, the front-end of a skill). In addition to increasing the number of voice interactions with a testing skill, the certification team should not simply trust the information provided by developers, and need to pay attention to 1) batch skill submissions from the same developer account; 2) inconsistency between description and privacy policy; 3) custom slots that can accept multiple types of sensitive data; and 4) plagiarism on skill's interaction models. VPA platforms may also build a dataset of front-end interaction models from existing skill submissions (both certified and rejected ones), extract features and train a machine learning model to automatically identify a suspicious skill submission.

Deploying automated skill testing tools for policy violation detection. It is necessary to deploy dynamic analysis tools which can automate the interaction with skills to explore functionalities actually undergoing implementation in skills. However, building a reliable and scalable voice-based testing tool is non-trivial. More recently, SkillExplorer [19] has been proposed, which is driven by a set of grammar-based rules to explore interaction behaviors of skills. However, it mainly focuses on identifying skills that collect private information without evaluating skills' conformity to various policies in a broader context (*e.g.*, the content policies listed in Table 7). As skills become more intelligent, the hard-coded grammar-based rules in SkillExplorer may not be scalable in handling diverse responses from skills. In Section 5.1, we have used our preliminary NLP-based detector to identify policy-violating kids skills in an automated manner. As our future work, we plan to develop a data-driven dynamic analysis tool based on publicly available dialogue datasets [1] to understand diverse questions from skills and generate corresponding responses, and extend our basic design of the policy violation detector.

Enforcing skill behavior integrity throughout the skill lifecycle. Our experiment shows that developers can arbitrarily change a skill's functionality after the certification on both VPA platforms. 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, 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 skill publishing latency. We also came across many broken skills during our analysis of the Alexa kids skills (details in Table 9 of Appendix C). Skills should be periodically checked and removed from the skills store if they are broken or violating any policy.

To fundamentally prevent policy-violating skills upon submission, VPA 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 when a skill is submitted, which could greatly enhance the trustworthiness of the skill certification.

7.3 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 processes in both the Amazon Alexa and Google Assistant platforms, 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, due to the diverse nature of content policies from different VPA platforms, we manually crafted and submitted policy-violating skills in our measurements. As our future work, we plan to employ natural language processing (NLP) techniques to automatically generate policy-violating content which can be filled in skill templates. Given a crafted skill, it is possible to automate the skill deployment and submission process by using the Selenium WebDriver [14]. Third, although we developed a policy violation detector, we manually explored skills' interaction behavior to identify existing problematic skills. Future work is needed to design dynamic analysis tools to automate the interaction with skills in different VPA platforms. Nevertheless, we have collected strong evidence in revealing the untrustworthiness of skill certification in leading VPA platforms, and empirically characterize potential security risks due to a lenient skill certification process. Our work has the potential to greatly evolve the current view of the trustworthiness of VPA platforms (*i.e.*, users blindly trust that VPA platforms prevent malicious skills from being published and properly protect their privacy).

8 CONCLUSION

In this work, we have conducted a comprehensive measurement to evaluate the trustworthiness of skill certification in two leading VPA platforms. We have crafted and submitted for certification 234 Alexa skills and 381 Google actions that intentionally violate 55 content and privacy policies defined by VPA platforms. Surprisingly, all Alexa skills and 39% of Google actions could pass the certification. We have further conducted an empirical study focusing on kids' skills and identified 31 (out of 755) problematic skills with policy violations and 34 broken skills in the Alexa platform, and one Google action (out of 114) collecting user names. Our user study with 203 participants demonstrated users' misplaced trust on VPA platforms. We have discussed actionable mitigation strategies to help VPA platform providers enhance the trustworthiness of their skill certification processes.

ACKNOWLEDGMENT

We are grateful to anonymous reviewers for their constructive feedback. This work is supported in part by National Science Foundation (NSF) under the Grant No. 2031002, 1846291, and 1642143, and the NSF/VMware Partnership on SDI-CSCS program under the Grant No. 1700499.

REFERENCES

[1] A Survey of Available Corpora for Building Data-Driven Dialogue Systems. https://breakend.github.io/DialogDatasets/. [Accessed 08-25-2020].

[2] 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/. [Accessed 08-25-2020].

[3] Alexa Skills Policy Testing. https://developer.amazon.com/fr/docs/custom-skills/policy-testing-for-an-alexa-skill.html. [Accessed 08-25-2020].

[4] Alexa Skills Privacy Requirements. https://developer.amazon.com/fr/docs/custom-skills/security-testing-for-an-alexa-skill.html#25-privacy-requirements. [Accessed 08-25-2020].

[5] Alexa Skills Security Requirements. https://developer.amazon.com/fr/docs/alexa-voice-service/security-best-practices.html. [Accessed 08-25-2020].

[6] General Data Protection Regulation. https://gdpr-info.eu. [Accessed 08-25-2020].

[7] Global Smart Speaker Users 2019. https://www.emarketer.com/content/global-smart-speaker-users-2019. [Accessed 08-25-2020].

[8] 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. [Accessed 08-25-2020].

[9] Industrial-Strength Natural Language Processing. https://spacy.io. [Accessed 08-25-2020].

[10] Inside Amazon's Fake Review Economy. https://www.yotpo.com/blog/amazon-fake-reviews/. [Accessed 08-25-2020].

[11] Menlo Report: Ethical Principles Guiding Information and Communication Technology Research. https://www.dhs.gov/publication/st-menlo-report. [Accessed 08-25-2020].

[12] Policies for Actions on Google. https://developers.google.com/actions/policies/general-policies. [Accessed 08-25-2020].

[13] 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/. [Accessed 08-25-2020].

[14] Selenium WebDriver. https://www.selenium.dev. [Accessed 08-25-2020].

[15] Smart Audio Report 2018. https://www.edisonresearch.com/the-smart-audio-report-from-npr-and-edison-research-spring-2018/. [Accessed 08-25-2020].

[16] Smart Spies: Alexa and Google Home expose users to vishing and eavesdropping. https://srlabs.de/bites/smart-spies/. [Accessed 08-25-2020].

[17] The Rise of Virtual Digital Assistants Usage. https://www.go-gulf.com/blog/virtual-digital-assistants/. [Accessed 08-25-2020].

[18] 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/. [Accessed 08-25-2020].

[19] Skillexplorer: Understanding the behavior of skills in large scale. In *29th USENIX Security Symposium (USENIX Security 20)*, 2020.

[20] Noura Abdi, Kopo M. Ramokapane, and Jose M. Such. More than smart speakers: Security and privacy perceptions of smart home personal assistants. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, Santa Clara, CA, 2019. USENIX Association.

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

[22] Noah Apthorpe, Sarah Varghese, and Nick Feamster. Evaluating the contextual integrity of privacy regulation: Parents' iot toy privacy norms versus coppa. In *USENIX Security*, 2019.

[23] Alexander Benlian, Johannes Klumpe, and Oliver Hinz. Mitigating the intrusive effects of smart home assistants by using anthropomorphic design features: A multimethod investigation. *Information Systems Journal*, pages 1–33, 2019.

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

[25] Guangke Chen, Sen Chen, Lingling Fan, Xiaoning Du, Zhe Zhao, Fu Song, and Yang Liu. Who is real bob? adversarial attacks on speaker recognition systems. In *IEEE Symposium on Security and Privacy (SP)*, 2021.

[26] S. Chen, K. Ren, S. Piao, C. Wang, Q. Wang, J. Weng, L. Su, and A. Mohaisen. 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)*, pages 183–195, 2017.

[27] Eugene Cho, S. Shyam Sundar, Saeed Abdullah, and Nasim Motalebi. Will deleting history make alexa more trustworthy? effects of privacy and content customization on user experience of smart speakers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

[28] H. Chung, M. Iorga, J. Voas, and S. Lee. "alexa, can i trust you?". *IEEE Computer*, 50(9):100–104, 2017.

[29] Jide S. Edu, Jose M. Such, and Guillermo Suarez-Tangil. Smart home personal assistants: A security and privacy review. *CoRR*, abs/1903.05593, 2019.

[30] Huan Feng, Kassem Fawaz, and Kang G. Shin. Continuous authentication for voice assistants. In *Annual International Conference on Mobile Computing and Networking (MobiCom)*, pages 343–355, 2017.

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

[32] Christine Geeng and Franziska Roesner. Who's in control?: Interactions in multi-user smart homes. In *Conference on Human Factors in Computing Systems (CHI)*, 2019.

[33] Hang Hu, Limin Yang, Shihan Lin, and Gang Wang. 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)*, 2020.

[34] Anjishnu Kumar, Arpit Gupta, Julian Chan, Sam Tucker, Björn Hoffmeister, and Markus Dreyer. Just ASK: building an architecture for extensible self-service spoken language understanding. In *Workshop on Conversational AI at NIPS'17*, 2017.

[35] Deepak Kumar, Riccardo Paccagnella, Paul Murley, Eric Hennenfent, Joshua Mason, Adam Bates, and Michael Bailey. Skill Squatting Attacks on Amazon Alexa. In *27th USENIX Security Symposium (USENIX Security)*, pages 33–47, 2018.

[36] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. Alexa, are you listening?: Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW):1–31, 2018.

[37] X. Lei, G. Tu, A. X. Liu, C. Li, and T. Xie. The insecurity of home digital voice assistants - vulnerabilities, attacks and countermeasures. In *2018 IEEE Conference on Communications and Network Security (CNS)*, pages 1–9, 2018.

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

[39] Nathan Malkin, Joe Deatrick, Allen Tong, Primal Wijesekera, Serge Egelman, and David Wagner. Privacy attitudes of smart speaker users. In *19th Privacy Enhancing Technologies Symposium (PETS)*, 2019.

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

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

[42] Nirupam Roy, Sheng Shen, Haitham Hassanieh, and Romit Roy Choudhury. Inaudible voice commands: The long-range attack and defense. In *15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18)*, pages 547–560, 2018.

[43] Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. In *Network and Distributed System Security Symposium*, 2019.

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

[45] Tavish Vaidya, Yuankai Zhang, Micah Sherr, and Clay Shields. Cocaine noodles: Exploiting the gap between human and machine speech recognition. In *9th USENIX Workshop on Offensive Technologies (WOOT 15)*, 2015.

[46] Qiben Yan, Kehai Liu, Qin Zhou, Hanqing Guo, and Ning Zhang. Surfingattack: Interactive hidden attack on voice assistants using ultrasonic guided wave. In *Network and Distributed Systems Security (NDSS) Symposium*, 2020.

[47] Yuanshun Yao, Bimal Viswanath, Jenna Cryan, Haitao Zheng, and Ben Y. Zhao. Automated crowdturfing attacks and defenses in online review systems. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1143–1158, 2017.

[48] Xuejing Yuan, Yuxuan Chen, Yue Zhao, Yunhui Long, Xiaokang Liu, Kai Chen, Shengzhi Zhang, Heqing Huang, XiaoFeng Wang, and Carl A. Gunter. Commandersong: A systematic approach for practical adversarial voice recognition. In *USENIX Conference on Security Symposium (USENIX Security)*, pages 49–64, 2018.

[49] Eric Zeng, Shrirang Mare, and Franziska Roesner. End user security and privacy concerns with smart homes. In *Proceedings of the Thirteenth USENIX Conference on Usable Privacy and Security*, SOUPS '17, page 65–80, USA, 2017. USENIX Association.

[50] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. Dolphinattack: Inaudible voice commands. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 103–117, 2017.

[51] Nan Zhang, Xianghang Mi, Xuan Feng, XiaoFeng Wang, Yuan Tian, and Feng Qian. 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)*, 2019.

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

[53] Serena Zheng, Noah Apthorpe, Marshini Chetty, and Nick Feamster. User perceptions of smart home iot privacy. *Proc. ACM Hum.-Comput. Interact.*, 2018.

Appendix A CONTENT POLICIES OF VPA 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	2/1/1	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	2/0/2	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	2/2/0	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	2/2/0	11/10/1
3.d	Provides information about black market sale of prescription drugs.	A	1/1/0	1/1/0	2/0/2	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	2/2/0	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	2/2/0	3/3/0
4.b	Recommends skills in Alexa's voice.	A	2/2/0	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	2/1/1	5/4/1
4.d	Solicits donations from end users.	A	3/3/0	2/2/0	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	2/2/0	5/4/1
6	Sexually Explicit content: Pornography and sex	A/G	3/3/0	6/6/0	2/0/2	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	2/0/2	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	2/0/2	5/3/2
7.c	Self-harm, including instructions to carry out self-harm.	G	2/2/0	2/2/0	2/1/1	3/2/1
7.d	Bullying and harassment	G	3/3/0	2/2/0	2/0/2	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	2/0/2	3/2/1
8.b	Purports to be able to predict gender.	A	1/1/0	2/2/0	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	2/0/2	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	2/1/1	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	2/0/1	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	2/0/2	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			2/0/2	2/0/2
10	Content					
10.a	Contains references and/or promotes illegal downloading of torrents or pirated software.	A/G	2/2/0	4/4/0	2/1/1	5/2/3
10.b	Contains specific advice on how to join an illegal organization.	A/G	3/3/0	4/4/0	2/0/2	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	3/0/3	4/3/1
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	3/0/3	6/1/5
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	3/0/3	6/3/3
10.f	Promotes use, sale, or distribution of recreational or illegal drugs.	A/G	3/3/0	5/5/0	3/0/3	6/3/3
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	3/0/3	6/3/3
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	3/0/3	6/1/5
10.i	Contains excessive profanity.	A/G	3/3/0	4/4/0	3/0/3	6/3/3
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	3/0/3	6/2/3
11.b	Contains profanity aimed at children.	A	3/3/0		5/0/5	1/0/1
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	3/0/3	6/4/2
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	3/0/3	6/3/3
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	4/2/2	6/2/4
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	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	2/2/0	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	2/0/2
14.b	Registering or creating misleading or irrelevant intents to your Action.	G	Most of our skills violated it		2/2/0	4/3/1
15	Spam					
15.a.	Submitting multiple duplicative Actions to the Actions directory.	G	4/4/0	2/2/0	2/2/0	4/0/4
Overall Summary			**119/119/0**	**112/112/0**	**180/44/136**	**201/104/97**

Table 7: Content policies [3, 12] we tested in our experiments against the skill certification process of VPA 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 381 policy-violating Google actions in total out of which 148 actions were certified and 233 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 VPA 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 8: Privacy requirements [4] defined by VPA 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 7. 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 7 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

Ask Santa's Elves, Be a Poet, Bit Pal, children calculation game, Count sheeps, fact or fake news, Fast Math Game, Good manners, Hillbrook School, June's Vocab & Quiz, Kid Power, Kids Booklet, Kidz Riddles, math tester, Maths Game, Medico Help educator, Mr. Murallo Casting, My yellow name, Party Trivia, Properly Brush My Teeth, Rylee's Sleep Sounds, Santa Cam, Santa's Letter, Santa's Log, Sentence Inventor, SH Arcade, Skyla's Unicorn, Snuggle Sounds, Talking Parrot, The Head Elf Hotline, Today's special, wildlife sounds, World History Quiz, Young Picasso .

Table 9: List of broken skills under the kids category in Alexa's skills store (as of July 2020).

Appendix D VPA PLATFORMS' CLAIMS ON POLICY VIOLATIONS

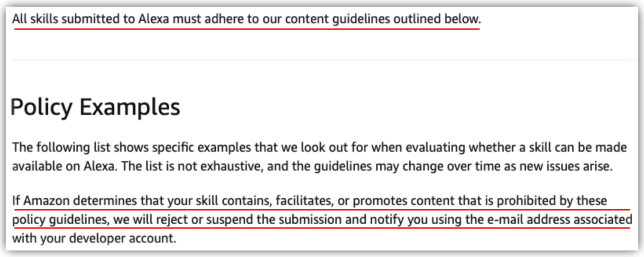


Figure 6: Amazon Alexa's claims on policy violations [3].

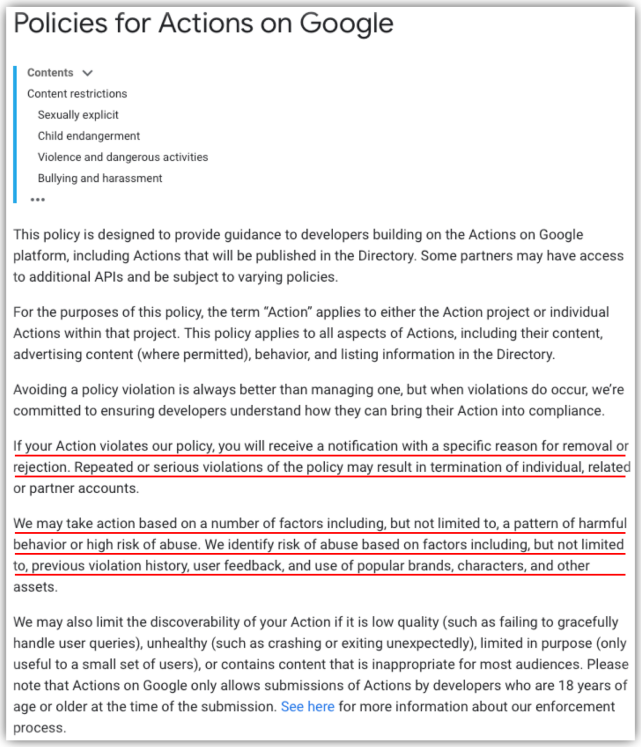


Figure 7: Google Assistant's claims on policy violations [12].

Appendix E AN EXAMPLE OF CERTIFIED AMAZON SKILL WITH POLICY VIOLATION

Figure 8 shows a live example of a certified skill with a policy violation on its first response. This skill also got certified on the first submission. It violates policy 2.a in Table 7 "promoting any products, content, or services, or directs end users to engage with content outside of Alexa".

Appendix F FEEDBACK TIME FOR CERTIFICATION

It is interesting that Amazon Alexa and Google Assistant platforms exhibited different patterns in terms of the certification feedback

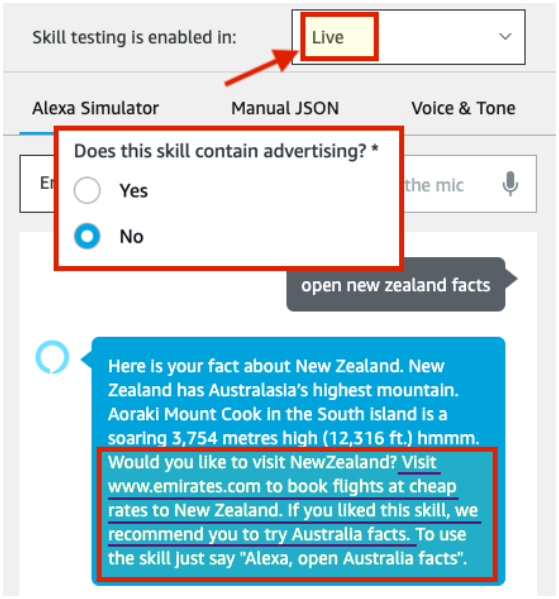
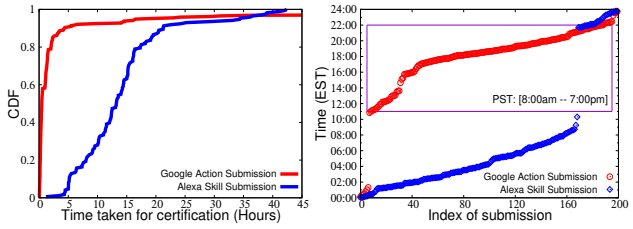


Figure 8: A certified Amazon skill with a policy violation (promotions and advertisements) on its first response. In the Privacy & Compliance form, we specified the skill "contains no advertising" but it actually does. This skill got certified on the first submission.

time. However, we did not observe a temporal pattern in terms of the certification results. Figure 9(a) shows the CDF (cumulative distribution function) of time taken to receive a feedback for skill/action certifications. For skill submissions, in most cases, we received the certification feedback within 24 hours. The certification process of Google actions took varying amounts of time. For our first 10 submissions in 2019, most certification results were received within 1-2 days (one case took 3 days and another case took 7 days to get the feedback). However, for our recent submissions from February to July 2020, the certification feedback was received much faster. About 50% of the actions took less than 1 hour to get a feedback.



(a) Time taken to receive feedback (b) Time when getting feedback

Figure 9: Feedback time for certification.

Figure 9(b) plots randomly selected 200 timestamps when we received certification feedback emails from Amazon Alexa and Google Assistant platforms, respectively. These timestamps are sorted by EST time. In most cases, feedback emails we received from Amazon had timestamps between 11:40pm and 8:37am EST. Since there are no certifications happening at any time other than this specific

time range, it is likely the feedback emails were sent either by an automated process (but evidence shows that the skill certification was largely performed in a manual fashion) or sent manually by an outsourced certification team. Actually, we were contacted by an Amazon employee based out of the US to help us in publishing a skill that was rejected multiple times on submission. For Google actions, the feedback emails for all the actions we submitted this year were received between 10:51am and 1:19am EST (in most cases, it was between 8:00am and 7:00pm PST, as shown in Figure 9(b)). This indicates that the Google certification team possibly works based on the US timezone, as opposed to the Alexa's certification team.

Appendix G NEGATIVE REVIEWS FOR KIDS SKILLS FROM THE AMAZON ALEXA PLATFORM

Our user study results in Section 6 show that when encountering something inappropriate during the interaction with VPA devices, 70% of users would provide a review for the skill on the skills store. This provides an opportunity to collect user attitudes and complaints about using VPA services, and may help us identify existing policy-violating skills. It also motivates us to conduct a preliminary user review analysis. To this end, we manually examined 2,085 negative reviews (*i.e.*, star ratings below 3-star) in the kids category from the Amazon Alexa platform, 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 10 illustrates some representative critical reviews from end users.

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 10: Selected critical reviews in the kids category.