Measuring Alexa Skill Privacy Practices across Three Years

Jide Edu
King's College London
London, UK

Xavier Ferrer-Aran
King's College London
London, UK

Jose Such
King's College London
London, UK

Guillermo Suarez-Tangil
IMDEA Networks institute
Madrid, Spain

ABSTRACT

Smart Voice Assistants are transforming the way users interact with technology. This transformation is mostly fostered by the proliferation of voice-driven applications (called skills) offered by third-party developers through an online market. We see how the number of skills has rocked in recent years, with the Amazon Alexa skill ecosystem growing from just 135 skills in early 2016 to about 125k skills in early 2021. Along with the growth in skills, there is increasing concern over the risks that third-party skills pose to users' privacy. In this paper, we perform a systematic and longitudinal measurement study of the Alexa marketplace. We shed light on how this ecosystem evolves using data collected across three years between 2019 and 2021. We demystify developers' data disclosure practices and present an overview of the third-party ecosystem. We see how the research community continuously contribute to the market's sanitation, but the Amazon vetting process still requires significant improvement. We perform a responsible disclosure process reporting 675 skills with privacy issues to both Amazon and all affected developers, out of which 246 skills suffer from important issues (i.e., broken traceability). We see that 107 out of the 246 (43.5%) skills continue to display broken traceability almost one year after being reported. As a result, the overall state of affairs has improved in the ecosystem over the years. Yet, newly submitted skills and unresolved known issues pose an endemic risk.

CCS CONCEPTS

• **Human-centered computing** → **Sound-based input / output**; *Natural language interfaces*; • **Security and privacy**;

KEYWORDS

security and privacy, voice assistants, alexa skills, smart speakers

ACM Reference Format:
Jide Edu, Xavier Ferrer-Aran, Jose Such, and Guillermo Suarez-Tangil. 2022. Measuring Alexa Skill Privacy Practices across Three Years. In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3485447.3512289

WWW '22, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9096-5/22/04…$15.00
https://doi.org/10.1145/3485447.3512289

1 INTRODUCTION

Smart Personal Voice Assistants (SPA) have become very popular systems mostly due to their interactive technology. This allows users to easily interface with networked appliances as well as to consume all kinds of online services using natural language [7, 29, 31, 36]. SPA incorporate voice-driven applications generally developed by third parties, referred to as *skills* in Amazon Alexa and *actions* in Google Assistant. Like in mobile apps, skills play an essential role, extending SPA capabilities by offering a wide range of services. The number of skills has multiplied in recent years. For instance, the Amazon Alexa skill ecosystem has grown from just 135 skills in early 2016 [27] to over 100k skills by late 2020 [12]. This rapid surge in numbers can be attributed to the continuous proliferation of SPA worldwide: about 4.2 billion SPAs are being used around the world [37]. This is projected to double in the next years.

Despite SPA popularity, there is increasing concern over what risks third-party skills may pose to users [1, 2, 23, 24, 28]. Skills widen the attack surface of SPA [13], as malicious actors may develop potentially harmful software that could affect the security and privacy of the users. Recent studies looked at various issues in third-party skills, including publishing potentially harmful skills [20, 33, 38], performing unjustified data collection [12], covertly eavesdropping conversations [16], and performing squatting attacks (purposely setting a skill name that sounds as another skill but is spelled differently to hijack its invocation) [21].

Although recent studies delved into various attack vectors in third-party skills, it is unclear to what extent current attacks permeate through the markets. Lessons learned from other platforms like smartphones [8, 15] indicate that SPA operators will struggle to keep the pace in the fight against misbehaving skills. This prompt us with the following open question: how effective are SPA market operators in helping protect users? One key feature that needs to be considered when answering this question is the strong dependency SPA hold with the cloud. *Skills are hosted on remote Web services controlled by the skills' developer.* This makes it easy for developers to modify the skill's functionality after its publication.

To usher how SPA markets are protected in a drifting landscape, it is imperative to study and evaluate the effectiveness of existing measures against malicious threat actors over time. This paper is the *first* to measure the changes in Alexa skill developer privacy practices over time, our measurement ranging from 2019 to 2021. We focus on the following research sub-questions: i) *Has the overall state of affairs regarding data practices in the third-party skill ecosystem improved over time?* (§3), ii) *Is the collection of personal information explained better nowadays?* (§4), iii) *What influence*

changes over time and has there been an improvement in the review and certification process? (§5), iv) *Are skills effectively bypassing the permissions system?* (§6). To answer these questions, we design a methodology to perform a data practice measurement (§2), which offers an independent assessment of the skill marketplace.

2 OUR MEASUREMENT METHODOLOGY

We build a Web scrapper to collect data from the Amazon Alexa marketplace at different points in time. With the data collected, we first *characterize* the market, then analyze skills statically (namely, *traceability analysis*) and dynamically (*interrogation analysis*) while performing a *differential analysis* to highlight changes over time.

Characterization. Amazon operates separate online marketplaces that cater to a variety of segments. The United States (US), United Kingdom (UK), India (IN), Australia (AU), Canada (CA), Germany (DE), Japan (JP), Italy (IT), Spain (ES), France (FR), and Mexico (MX) are among the 11 online markets with third-party skills.

Web Scrapper. We built a Web scrapper with a framework that recursively crawls all markets with third-party skills and extracts metadata from the skills. The Web scrapper visits the different skill categories while building a collection of links corresponding to each skill. It then iterates through the collected skill links to visit the skill's website and extract the skill attributes. However, we see a lack of coverage when deploying traditional crawling methodologies. Additionally, due to a vested interest in protecting its data, Amazon discourages scraping [3] and currently implements different anti-scraping techniques, making it challenging to scrape Alexa skill markets. These include the use of captchas, email verification, and blacklisting of IPs.[1] Besides, some of the Alexa markets have varying page structures. Amazon instead encourages using their APIs such as the Product Advertising API and Marketplace Web Service Products API to query the marketplace. However, Amazon is selective regarding the information that can be accessed from these APIs. To overcome some of the anti-scraping measures: i) We limit the rate at which we generate our requests; ii) We mimic human behavior; iii) We make our requests through a pool of IP addresses and proxies; iv) Lastly, we design our scraper to handle, and react to, exceptions such as "ElementNotVisibleException" which occurs when the scrapper tries to find an element not visible within the skill page, or "NoSuchElementException" when elements unexpectedly become not available. All in all, we collected three snapshots of all market segments — one in May 2019, one in July 2020, and the last one in April 2021, respectively.

Search-based Crawling. Amazon organizes their skills into categories and subcategories, with each having a limit of 400 pages. There are 23 categories and 66 subcategories in total. However, none of the markets has the entire skill index listed. We get around this restriction by conducting tailored searches within each marketplace's subcategories. Note that a single skill can fall into several categories and be hosted in multiple marketplaces. We identify unique skills through the skill identifier embedded into the URL. Our latest scrape shows a total of 124,026 skills and 50,526 developers in 2021. In 2020 we see 111,796 skills and 46,804 developers, and in 2019 we see 84,856 and 31,238, respectively. This already shows a sharp increase in skills and developers over the years. For a full breakdown of this per marketplace and per category, see Appendix B and Appendix C, respectively.

Feature Extraction. Unlike other platforms like Android, Amazon Alexa runs the skills in the cloud, and the code of the skills is not publicly available. We use the following attributes (which we scrape from the skill's website) to characterize every skill: invocation utterances, permissions, the category, developer's information, privacy policy, terms of use, skill's name, skill description, cost, rating information, and reviews.

Traceability Analysis. We look at privacy policies to understand how developers disclose and justify the data permissions they request. For this, we leverage English-speaking privacy policy statements annotated to train a Machine Learning (ML) model that automatically identifies the traceability between the data operations performed by the skill and the data actions defined in their privacy policies as in [12]. We focus only on the 5 English-speaking markets: US, UK, IN, AU and CA, representing >80% of skills. Skills are not executed at this stage, as data operations in Alexa are protected by the Amazon API that requires explicit permission from users before being invoked. The permissions are listed in the market and consent is given at installation time. The data actions defined in the privacy policies are extracted using Natural Language Processing (NLP). Our model yield an average F1-score of 96.5% as in [12].

Interrogation Analysis. We dynamically interact with the skills by systematically engaging into a synthetic conversation following the method in [16]. Our tool comprises a range of components design to meaningfully interact with a skill (including utterance extraction, question understanding, answer generation and behavior exploration) as described in detail in Appendix E. Our tool has 81% coverage, similar to the coverage reported in [16].

Differential Analysis. We finally study how a skill changes by computing a differential of the representation of the skill at two points in time. Let the state of a skill be $S(f, t, d)$, where f is any of the features obtained during the *feature extraction* process (typically, the permissions although our methodology supports a wide range of features), t is the result of the traceability analysis (typically, complete, partial or broken), and d is the result of the interrogation analysis (typically, data collection practices through conversations). We define the differential of two states as $\mathcal{D} = S_{t_1}(f, t, d) - S_{t_2}(f, t, d)$, where t_1 and t_2 are two points in time and $\mathcal{D}$ represents the Levenshtein distance between the set given as inputs. For instance, a skill i that requests a new permission p in 2021 (over 2020) and its traceability changes from complete to broken results in the following: $S^i_{2021} - S^i_{2020} = [\texttt{insert}(p), \texttt{substitute}(complete, broken)]$

3 PERMISSIONS IN THE SKILL ECOSYSTEM

We next present a characterization of Alexa skills through the lens of our dataset. In particular, we focus in this section on permissions as a reliable proxy to understand data collection practices [26, 32].

Distribution of permissions by skills. Table 1 shows that more than 97% of skills have not been requesting permissions over the years. These skills are very simple, not needing user data (e.g, "Good Morning!" skill). However, the majority of the skills that request permissions appear listed in an English-speaking marketplace. Notably, the skills that declare permissions display an increasing trend over time. In particular, we see 0.41% of skills requesting more than

[1] Recent judgment in US shows that such scraping from public services is legal [17, 30].

Table 1: Number of permissions request over time.

No	2021		2020		2019	
	Skills	%	Skills	%	Skills	%
0	120,848	97.44%	109,120	97.61%	83,427	98.32%
1	2172	1.75%	1882	1.68%	1082	1.28%
2	625	0.50%	511	0.46%	241	0.28%
3	239	0.19%	188	0.17%	83	0.10%
4	80	0.06%	57	0.05%	14	0.02%
>=4	62	0.05%	38	0.03%	9	0.01%
Total	124,026	100.00%	111,796	100.00%	84,856	100.00%

Table 2: Distribution of permissions per category.

	Music	Games	Lifestyle	Education	Health	Productivity	Business	Shopping	Food	Travel	Social	Weather	News	Home	Utilities	Sports	car	Local	Kids
2021	459	319	293	247	223	220	210	175	163	153	152	98	93	92	67	60	37	36	32
2020	376	215	264	204	171 ▼	187	162	157	161	165	128	86	67	92	57	56	31	34	14 ▼
2019	122	157	119	117	231	73	105	52	74	80	58	48	23	42	32	16	10	15	24

one permission in 2019, rising to 0.71% in 2020 and to 0.82% in 2021. We see similar trends as the number of permissions increases, e.g.: there are 53 (62 − 9) skills more that are asking for >= 4 permissions in 2021 when compared to 2019. This increase is on average for all marketplaces, and we note that it is imbalanced. For instance, in 2021, the number of skills asking for more than four permissions in the IN marketplace increases by 70%, while the number of skills asking for three permissions increases by 133%.

Distribution of permissions per category. Table 2 shows the distribution of permissions by category over the years for the most relevant categories. In general, we see an increment in the number of skills requesting more permissions over time. For example, in 2021, we see a 43.2% rise in the number of skills in the *Lifestyle* category, asking for at least three permissions compared with 2020. This is an over 146% increase w.r.t. the number of skills in this category in 2019. Conversely, two categories—*Health* and *Kids*—have had a drop from one year to another (denoted as ▼ in Table 2). The number in the *Kids* category reduces by 41% in 2020 from the number we see in 2019. This indicates that there are certain categories that are under more persistent scrutiny.

To better understand the relationship between skills in a category and the number of permissions requested, we selected and further analyzed the top categories with many skills asking for more than two permissions. Our finding shows that out of the 21 skills requesting for more than three permissions under the *Education and Reference* category in 2021, 15 (71.4%) skills are developed by *VoiceXP* all asking for four permissions—*Mobile Number, Email Address, Full Name* and *Device Address*. Furthermore, 9 (60%) of these skills have no reviews or ratings. Likewise, in 2020, *VoiceXP* also has 12 (75%) of 15 skills with more than 3 permissions in the *Education and Reference* category. Similarly, in the *Music & Audio* category, 50% of the 20 skills requesting more than two permissions are also published by a single developer—*Alpha Voice*. These skills request *Device Address, Lists Read Access*, and *Lists Write Access* with 40% having a single review or rating.

Distribution of permissions by type. Table 3 shows how the different permissions are distributed across the years. The most requested permissions are *Device Address*, *Email Address* and *Device Country & Postal Code* generally used to offer services based on the user's location. For example, the *Device Address* is asked for by 772 skills (570 developers) in 2021, 753 skills (567 developers) in 2020, and 519 skills (381 developers) in 2019. In contrast, *Amazon Pay* is the least asked permission which is requested by 111 skills published by 75 developers in 2021 and 83 skills by 61 developers in 2020. Skills requesting for *Location Service* increase from 2% in 2019 to 4% in 2020 while those asking for *Device Address* and *List Access* reduces by about 9%, respectively. Overall, we see more skills asking for *Location Service*, *Email Address*, *Name*, *Reminders* and *Mobile Number* across the years. On the contrary, fewer skills are now requesting for *List Access*, *Device Address*, and *Device Postal Code*. This could potentially be due to developers being increasingly more concrete on the type of personal information they collect.

Table 3: Distribution of permissions by type.

Permission	2021			2020			2019		
	D	N	%	D	N	%	D	N	%
Device Address	570	772	16%	567	753	19%	381	519	28%
Email Address	445	761	16%	345	544	14%	137	160	9%
Device Country	394	707	15%	381	644	17%	305	378	20%
Name	287	524	11%	223	400	10%	79	118	6%
Reminders**	282	482	10%	205	263	7%	64	82	4%
Alexa Notifications**	275	555	12%	249	461	12%	117	165	9%
List Access	183	415	9%	183	417	11%	155	347	19%
Location Services	177	203	4%	140	156	4%	35	37	2%
Mobile Number	152	231	5%	112	162	4%	35	37	2%
Amazon Pay	75	111	2%	61	83	2%	31	31	2%
Timers**	15	15	0.3%	8	8	0.2%			
Skill Personalization	5	6	0.1%						
Total	2860	4782	100%	2474	3891	100%	1339	1874	100%
Unique	1887	3178	66%	1714	2676	69%	1022	1429	76%

** Amazon does not expects developers to disclose their collection in the privacy policy, **D** = Number of developers, **N** = Number of skills.

Note that in Table 3, *Name* refers to the aggregate of the *First Name* and the *Full Name* permissions and *List Access* is the aggregate of *List Read Access* and *List Write Access* permissions. Also, *Alexa Notifications* permission is now deprecated.

4 TRACEABILITY

The type of traceability is identified by comparing the permissions requested by the skill through the Amazon Alexa API with the data practices covered in the privacy policy. Traceability is evaluated as broken, partial or complete, as in other related works [12].

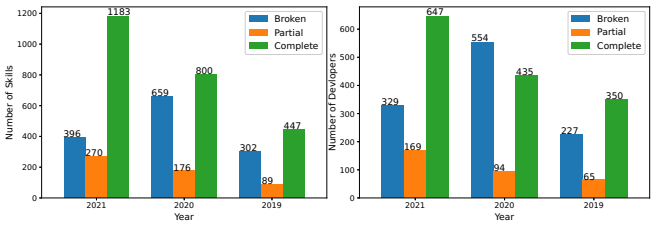
Complete: A skill offers complete traceability if it provides adequate information in its privacy policy document about its data practices, i.e., the data action defined in the privacy policy document can be completely mapped to the access data permissions.

Partial: A skill offers partial traceability if not all its data permissions are covered in its privacy document. Likewise, when data practices in a privacy document are not well mapped with the skill's data permission, the skill is evaluated to have partial traceability.

Broken: A skill has broken traceability if it has no data implication in its privacy policy document.

4.1 Traceability per Skills and Developers

Figure 1b shows that developers' data disclosure practices were poor in 2020 compared to 2019. About 35% of developers have skills with broken traceability compared to 51% in 2020. Instead, traceability improved considerably in 2021 compared to the previous years



(a) Skills from 2019-2021 **(b) Developers from 2019-2021**

Figure 1: Traceability results for English-speaking markets.

(see Section 5 to understand the factors impacting these changes, including the responsible disclosure of 675 skills we did in the second half of 2020). Naturally, as developers become more (or less) aware of the importance of disclosing their data collection practices, the traceability of the skills they develop (see Figure 1a) also change in a similar fashion. However, when we put things in perspective, we see that the number of developers with sound data practices (complete vs broken & partial) disclosure has only risen from 55% in 2019 to 57% in 2021.

4.2 Traceability per Category

To understand how traceability changed across types of skills, we look at the market category. Specifically, we compute the traceability by category in the five English-speaking marketplaces. Next, we evaluate the different categories based on the number of concerns (broken and partial) normalized by the number of well-defined policies (complete). As shown in Table 4, the *Kids* category is ranked first in category with issues in 2021 and the *News* category in 2020. They have the highest ratio of skills with inadequate privacy disclosure to those that are well defined. The *Music & Audio* category has the largest number of complete traceability skills across the years, which is also a sizable proportion of skills within the category.

Table 4 also shows that traceability improves in category such as *Business & Finance*, *Movies & TV*, and *Music & Audio*. For instance, the *Business & Finance* category is currently ranked 18th out of the 21 categories. This is an improvement from the previous rank of 12th in 2020 and 8th in 2019. We now see bad privacy practices in 51 skills compared to 152 skills in the same category with complete traceability. Also, the *Movies & TV* category ranked 10th in 2019 and 5th in 2020, now ranked 19th. We similarly observe categories where traceability has gone worse. An example is the *Utilities* category currently ranked 6th in 2021 from 9th in 2020, and 11th in 2019. Our findings here confirm the hypothesis drawn in Section 3 that certain categories are under heavier scrutiny, but it also shows that the effectiveness of having more complete traceability and less broken (or partial) in a category changes from one year the another.

4.3 Traceability by Permission Type

We also look at the traceability of skills per permission requested. Table 5 shows the distribution of traceability across the different types of permission for the skills that request permissions and warrant a privacy policy in the English-speaking marketplace. The permissions are first grouped into broken, partial, complete, with respect to the policies of the skills where these permissions are requested. In 2021, a total number of 2,852 permission are requested

Table 4: Traceability by category (markets in English).

	2021				2020				2019			
Category	R	B	P	C	R	B	P	C	R	B	P	C
Kids	**1**	2	4		19	2		4	6	10	2	9
Novelty & Humor	2	17	5	8	2	24		8	3	11	3	8
Weather	3	36	17	37	4	45	5	22	4	15	7	15
Food & Drink	4	53	25	61	3	71	20	40	13	19	5	27
News	5	12	34	36	**1**	27	24	11	7	10	2	10
Utilities	6	23	8	30	9	26	4	18	11	10	4	13
Games	7	66	**56**	151	8	121	**42**	101	5	66	**36**	75
Smart Home	8	24	8	41	7	44	2	27	20	16		27
Local	9	5	4	11	14	7	4	9	**1**	7	2	3
Connected Car	10	9	2	14	18	7	2	11	2	1	1	
Social	11	17	5	30	16	22	3	25	17	9		12
Travel & Transp.	12	36	15	73	6	77	8	45	14	15	4	22
Health & Fitness	13	32	25	87	10	73	13	55	12	**131**	34	**170**
Shopping	14	13	43	91	15	20	**42**	59	9	12	3	13
Productivity	15	58	17	139	11	94	12	81	19	16		26
Lifestyle	16	**68**	35	198	17	122	12	147	16	36	11	65
Education & Ref.	17	61	21	201	13	**125**	7	113	15	42	9	67
Business & Finance	18	28	23	152	12	75	7	64	8	37	7	40
Movies & TV	19	3		8	5	8	5	5	10	2		1
Music & Audio	20	45	19	**367**	20	79	12	**292**	18	47	4	78
Sports	21	4		38	21	7		38	21			7
Total		612	366	1,773		1,076	224	1,175		512	134	688
Unique		396	270	1,183		659	176	800		302	89	447

B = Broken, **P** = Partial, **C** = Complete, **R** (Rank) ~ (B+P)/(C+1)

Table 5: Distribution of traceability across different permissions in the 5 English-speaking marketplaces across 3 years.

Permission	2021				2020				2019			
	R	B	P	C	R	B	P	C	R	B	P	C
Device Address	586	**162**	88	336	**559**	**255**	58	246	**374**	**141**	**40**	**193**
Device Country	**598**	87	48	**463**	528	148	25	**355**	273	87	20	166
Email Address	558	82	119	357	385	123	83	179	86	30	7	49
List Access	296	105	20	171	281	140	10	131	227	96	6	125
Name	419	64	**134**	221	322	93	**101**	128	87	25	30	32
Mobile Number	182	27	32	123	127	42	21	64	27	6	6	15
Location Services	137	41	45	51	94	46	21	27	15	7	4	4
Amazon Pay	76	7	16	53	51	9	11	31	13	1	3	9
Total	2,852	575	502	1,775	2,347	856	330	1,161	1,102	393	116	593
Unique	1,849	396	270	1,183	1,635	659	176	800	838	302	89	447

R = Requested, **B** = Broken, **P** = Partial, **C** = Complete.

(622 by skills with broken traceability, 485 by skills with partial traceability, and 1,509 by skills that exhibit complete traceability). We see that *Amazon Pay* is the least asked permission which is requested by 76 skills in 2021 and 51 skills in 2020, and also tends to be requested more by skills that have complete traceability. In contrast, *Location Services* permission requested by 137 skills in 2021, 94 in 2020, and 15 skills in 2019 is found more in skills that exhibit broken traceability. This means that the type of permission matters when it comes to the justification of the collection practices and the desired data flow patterns. This could be effectively leveraged to implement a better triage mechanism during a vetting process. We discuss the implications of over-privileged skills in Section 7.

4.4 Profiling Developers

Table 6 shows the number of developers per type of traceability considering the 5 English marketplaces across the years.

Complete: In 2021, there are 638 (56%) developers with *all* their skills showing complete traceability. This implies that all their skills have statements in their privacy policies clearly stating and justifying their request's permissions. This is higher than the 423 (40%) developers we see in 2020 and the 347 (54%) in 2019.

Broken: There are 540 developers in 2020 with *all* their skills broken. This accounts for about 51% of the developers. Their skills do

Table 6: Developers' disclosure practices.

Year	D	B	P	C	B+P	B+C	P+C	P+B+C
2019	638	223	64	347	1	3		
2020	1068	**540**	90	423	**3**	**11**		**1**
2021	1133	323	**161**	**638**	2	3	**5**	**1**
Total	2839	1086	315	1408	6	17	5	2
Unique	1349	666	182	740	4	13	5	1

D = Developer, **B** = Broken, **P** = Partial, **C** = Complete

not generally offer an adequate explanation when we analyze the skills, their privacy statements, and their reviews. The number is much lower in 2021 as we find only 323 (29%) developers with all their skills exhibiting broken traceability.

Partial: We see 161 developers with *all* their skills with partial traceability in 2021. This accounts for 14% of the developers. They appear to have a lax attitude when writing privacy policies and informing users of how the personal data requested is used. We see the highest number of skills with partial traceability in 2021 compared to the 8% and 10% in 2020 and 2019, respectively.

Mixed: We see a handful of developers with a mix of broken (B), partial (P), and complete (C) (see B+P, etc. in Table 6). There is an interesting case of a developer *Blutag Inc.* in the P+B+C case. It has 74 skills, 1 broken (dead link), 35 partial, and 38 complete.

While we see an increasing trend towards having more complete traceability over time, we still see more broken skills in 2021 than in 2019. Also, partial traceability seems to be the issue.

5 FACTORS IMPACTING TRACEABILITY

Next, we explore several hypotheses on what could have influenced the changes we see over the years. In particular, we analyze: i) the impact of new skills on the ecosystem. ii) how existing skills' traceability has changed over time, iii) the impact of change in skills' permissions in the ecosystem, iv) the effect of the responsible disclosure we did to Amazon and third-party developers.

5.1 Effect of New Skills on Traceability

We investigate the effect of new skills on traceability. As shown in Figure 2 there are 996 new skills added between 2019 and 2020 that ask for permissions that warrant privacy policies. Similarly, there are 399 new skills added between 2020 and 2021 that ask for permissions that Amazon expects developers to disclose their collection in the privacy policy. Interestingly, this data shows that more skills with complete traceability have been added over the years than skills that exhibit broken or partial traceability. In particular, 518 (52%) skills in 2020 and 256 (64%) skills in 2021 are new skills added with complete traceability.

However, the number of skills with issues is also on the rise. In particular, 478 (48%) skills with privacy issues were added between 2019 and 2020, and 143 (36%) of these skills were added between 2020 and 2021. One good example is the "air monitor" skill by *AirMonitor* added in 2021. This skill collects *Device Address and Location Services*. However, the skill exhibit broken traceability as the privacy policy links direct users to a dead page. Although the overall state of affairs is improving, many newly submitted skills still have privacy issues. We therefore, posit that the vetting process could still be improved. Also, the research community (with studies like ours, as we show in Section 5.4) has made a commendable effort to contribute to the market's sanitation.

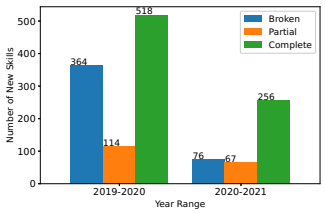


Figure 2: Traceability of newly added Skills.

Table 7: Change in traceability across time.

	2020										
2019	Traceability	B		P		C		PR		SR		Total
		N	%	N	%	N	%	N	%	N	%	
	B	199	65.9%	2	0.7%	6	2.0%	5	1.7%	90	29.8%	302
	P	16	18.0%	58	65.2%	1	1.1%	1	1.1%	13	14.6%	89
	C	80	17.9%	2	0.4%	275	61.5%	3	0.7%	87	19.5%	447
	Total	295	35.2%	62	7.4%	282	33.7%	9	1.1%	190	22.7%	838
	2021											
2020	Traceability	B		P		C		PR		SR		Total
		N	%	N	%	N	%	N	%	N	%	
	B	315	47.8%	38	5.8%	162	24.6%	57	8.6%	87	13.2%	659
	P		0.0%	163	92.6%	2	1.1%		0.0%	11	6.3%	176
	C	5	0.6%	2	0.3%	763	95.4%	3	0.4%	27	3.4%	800
	Total	320	19.6%	203	12.4%	927	56.7%	60	3.7%	125	7.6%	1635

B = Broken, **P** = Partial, **C** = Complete, **SR** = Skills Removed, **PR** = Permission Removed

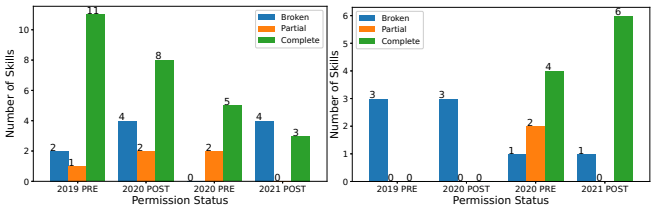
5.2 Traceability across Existing Skills

We investigate how the traceability of existing skills has changed over time. This could allow us to measure the effect of Amazon continuous vetting techniques. Table 7 shows how the traceability has changed over time. We could see that out of the 302 broken skills in 2019, 199 (65.9%) were still broken in 2020, 90 (29.8%) skills were removed, 2 (0.7%) have partial traceability, and 6 (2%) were complete in 2020. However, 80 skills that exhibit complete traceability in 2019 were broken in 2020. On further analysis, we found that this change in traceability is due to a lack of access to the skills' privacy documents. The policy links are either dead or take users to a dead page. A possible explanation for this might be that developers no longer maintain these skills. This could also be one of the reasons why skills remain broken over the years.

Equally, out of 659 broken skills in 2020, 162 (24.6%) were complete in 2021, 57 (8.6%) had their permission removed, 87 (13.2%) complete removed, and 315 (47.8%) were still broken. We see 5 (0.6%) skills that were previously complete in 2020 becoming broken in 2021 also due to dead link. An example is the "Kids Booklet" by *WebRecycles Inc* that collects *Device Country and Postal Code*. The traceability changed from complete in 2020 to broken in 2021. Even so, the traceability of skills with bad privacy practices in 2020 improved considerably in 2021. Only 320 skills were still broken from the same set of skills we see in 2020 compared to 659. The result shows that both Amazon and developers have worked to improve the traceability of skills in this ecosystem.

5.3 Effect of Change in Permission(s)

Does traceability change because permissions change? To answer this question, we first study changes in the use of permissions. In Figure 3a, we see an increase in the number of permissions per skill and how it negatively impacts their traceability. The "PRE" and



(a) Increase in permission (b) Decrease in permission

Figure 3: Traceability of skills before (Pre-) and after (Post-) increase and decrease in number of permissions.

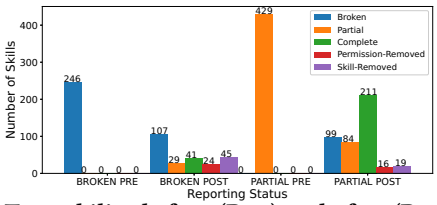


Figure 4: Traceability before (Pre-) and after (Post-) reporting the skills with issues

"POST" suffix in Figure 3a indicates pre-increase and post-increase, respectively. Between 2019 and 2020, 14 skills asked for additional permissions. In 2019, 11 (79%) skills had complete traceability, while 3 (21%) have inadequate privacy disclosure practices, including partial traceability. However, the number of skills with privacy issues increases by 100% to 6 in 2020 after the skills requested more permissions. A similar trend can be seen between 2020 and 2021, where the number of skills with insufficient privacy disclosure increases by 100%. For example, "Salah Time" skill by *Arshad* collects *Device Country and Postcode* in 2020 and exhibits complete traceability. It then collects *Device Address, Location Services, Reminders* in 2021 and exhibits partial traceability. The traceability difference from complete to partial is due to the use of the same privacy policy, even when different data is collected.

We next look at the opposite angle and study changes in traceability as the number of permissions decrease. Figure 3b shows the traceability of skills before (PRE) and after (POST) they reduce the number of permissions they requested. Note that we exclude those skills that have their permission wholly removed to avoid biasing the result. As we can see, there is no change in the number of skills with privacy issues between 2019 and 2020, even after the number of permissions requested reduces. However, we can see an improvement in traceability when the number of permissions requested by skills reduces between 2020 and 2021. We see a 50% increase in the number of skills with complete traceability from 4 to 6 skills. Nevertheless, these results need to be interpreted with caution because of the small skills involved. However, low number of skills does not mean a low interactions (or "installations"). Among the skills is the popular "Uber" skill by *Uber.com* with hundreds of reviews and possibly hundreds of interactions.

5.4 Effect of Responsible Disclosure

To enhance the security of the skill ecosystem, we safely report our findings. We perform a responsible disclosure process, starting from mid-August 2020, as follows. First, we notify all skill developers who are not engaging in good data practices whenever we have their contact details. Second, we also report our findings to Amazon and have confirmed that the skill store team has taken action. Thus, we measure the effect of the responsible disclosure.

From the data in Figure 4, we can see that out of 246 skills with broken traceability reported (*BROKEN PRE*), 111 (45.12%) no longer pose a threat to users at the time of writing: 45 (18.29%) of these skills have been removed and are no longer available on Alexa, 24 (9.76%) have their permission(s) removed, and 41 (16.67%) of them now have complete traceability. Overall, 356 (52.74%) out of 675 reported skills no longer threaten the users. This result corroborates our earlier findings in Section 5.1 that while traceability has improved, there are still skills with privacy issues across markets. Likewise, it shows how Amazon could benefit from enabling more actionable research mechanisms to study privacy issues.

6 BEYOND PERMISSIONS

Amazon enforces access to personal data through permission via their APIs, as explained before. However, prior work [16] shows that skills could bypass this system and request personal data directly from the user via conversations: they found 100 skills across the US market in 2020 asking for personal data via conversation using an interactive system called SkillExplorer. To understand how conversational skills may have changed, we interact with those available in the US market in 2021 and compare our findings with the results obtained in [16] in 2020. We implement SkillExplorer as in [16] to automatically interact with skills, because its original implementation is not publicly available, nor was shared upon request. Refer to Appendix E for details about the implementation and the evaluation of its equivalent accuracy.

We interact with 35k skills in the US market, excluding skills without unique invocation names, as SkillExplorer can not handle them [16]. We find 65 skills requesting personal information via conversation. This is 35% less than the ones found in 2020 [16]. In particular, 58 (85%) skills collect users' name, 4 skills collect zip code, 3 (5%) request for user's birthday, 2 (3%) collects user's phone number, and a skill collects user's location.[2] We then examine the traceability exhibited by these 65 skills found in 2021. The results show that 37 (57%) of these skills have broken traceability between the data collected via conversation together with the Amazon Alexa API, and the data practices mentioned in their privacy policies, if any. Besides, 3 (5%) exhibit partial traceability and only 25 (38%) have complete traceability. Out of the 37 skills with broken traceability, we see 29 (78%) skills completely lacking a privacy policy document, five skills with a policy link that redirects us to a dead page, and three skills without any mention about data practices in their privacy policy. Interestingly, most of the conversational skills requesting personal information via conversations do not ask for permission via Alexa API.

One example is "F1 forecast" by *Jordan Perkins* in the US market, which informs its users of the latest news and updates in the F1 world. When the skill is invoked with the utterance "Alexa, ask f one forecast where Charles Leclerc qualified at the last race", it requests to know the user's address, which seems irrelevant. The skill also lacks a privacy policy to state the purpose behind this. We also find skills like "Name Expansion" by *Jackson Jacob* asking

[2] 3 of the skills requests for more than one personal data.

for the name of the user. Since the skill function expands the user's name, we deem this relevant as the skill needs it to offer its services. However, while the skill has a privacy policy link, accessing it takes us to a dead page and thus also exhibits broken traceability.

What has changed over time? The above measurement shows a unique view of the underlying issues behind conversational skills requesting for personal information. However, to understand changes over time, we further explore the 100 skills (developed by 89 distinct developers) collecting personal data via conversation provided to us by the authors of SkillExplorer [16]. This dataset was collected from the US marketplace in early 2020. Out of the 100 skills having privacy issues in 2020, we only find 25 conversational skills available in 2021 (which are also included in the 65 skills our tool finds as stated above). Interestingly, from the 100 reported skills in 2020, only 3 skills have been taken down and 72 are still listed in the Amazon market but unavailable. Amazon does not allow users to interact with those skills for several reasons. First, Alexa suggests a different skill (albeit with a similar name) to the one being invoked in 36 of the 72 skills. Second, Alexa replies that it does not 'understand what you want' for 34 of the 72 skills. Note that we invoke skills through the Amazon Alexa simulator that supports text interaction using a simple command: "Alexa open *[Skill Name]*". For another skill, Alexa replies that it is 'having trouble accessing' it. Finally, the remaining skill is listed as 'not currently available'.

Out of the 25 skills that are available in both 2020 and 2021, 16 skills (64%) still request users to provide personal information via conversation. Furthermore, only 5 (31%) skills exhibit complete traceability with the personal data collected via conversation. One has partial traceability, and the rest 9 (60%) have broken traceability. In particular, 7 have no privacy policies, one has a dead link and one has a privacy link that points to a porn site. This is the case of the "Praise Me" skill published by *Jackson Jacob*, which is available across all five English-speaking countries.

With most of the available skills still having broken traceability and no privacy policies, our takeaway is that, even if the number of skills collecting personal information via conversation is not that high, especially when compared with skills collecting information via permissions, more efforts are needed to sanitize the marketplace.

7 DISCUSSION

Increase in the number of skills and developers. There is a sharp increase in the number of third-party developers contributing to the skill ecosystem. As reported in Section 2, the number of developers rose by 62% from 31K in 2019 to 51K in 2021. This number is a sizable set and indicates that the ecosystem is ramping up. Unfortunately, while the growth offers users more functionality, it could potentially usher in a new level of threats and actors that could attack it. After all, developers have different motivations for publishing skills [12], and being able to identify malicious developers is critical to secure users' privacy. Amazon should implement mechanisms to validate skill developer's identity for easy attribution, which, at least from the marketplace, is currently not possible.

Improved skill review and certification. Amazon's privacy requirements for skill developers mandate that a skill must come with an adequate privacy policy if it collects personal information [4]. But, unfortunately, we see skills having a privacy policy only to fulfill Amazon requirements and not to create awareness and control of data practices, which are essential to help users protect their privacy. It is apparent that some developers are approaching privacy policy requirements as a tick box exercise disregarding users' privacy. Ensuring that privacy policies are relevant, accessible, and understandable will go a long way in providing transparency about skill data practices and help users set the available privacy controls.

Overall, as already discussed, there seems to be an improvement in the review conducted as part of the skill certification process judging by the improved traceability of new skills added recently (cf. Figure 2). Thus, it would be interesting to repeat studies like [11] that had shown the review and certification process not working a couple of years ago by attempting to ethically inject skills in the marketplace. Also, we have seen Amazon taking action by removing several skills with privacy issues that were reported to them. Although many skills are requesting more permissions in 2021 (c.f Table 2), there is an improvement in skills traceability over the years. Notwithstanding, many newly added skills still exhibit broken traceability, which suggests there is still room to improve the review process. The pipeline of tools and the insights used in this paper can help in that endeavor.

Better but still not good enough. In 2021, we see bad privacy practices in about 666 skills (36% of those that request permissions in the English-speaking marketplaces). This is an improvement from the 835 skills (51%) we observe in 2020 and also, in proportion, from the 391 (47%) we see in 2019. We also see how the research community has supported the sanitation of the market. All this seems to suggest an improving trend in terms of traceability, despite the high number of skills still exhibiting bad privacy practices in 2021. Notably, we see that 107 out of 246 skills (43.5%) continue to display broken traceability almost one year after being reported to both Amazon and the respective developers as part of our work (c.f. Section 5.4). We see that including or removing new permissions has a clear impact on traceability, with skills increasing the number of permissions across the years negatively impacting their traceability and skills decreasing the number of permissions impacting their traceability positively. Furthermore, looking at the privacy issues based on skill categories, we see a large number of skills in the *Lifestyle* and *Games* category exhibiting broken traceability and partial traceability, respectively. These two categories comprise skills that offer services related to the user's behavioral pattern, daily interaction, consumption, work, activity and other interests that could potentially describe them. In contrast, the *Music & Audio* subcategory has the most significant number of complete traceability skills, which is also a sizable proportion of skills within the subcategory. Note that this category is related to industries with a larger tradition of offering services on the Web, where privacy has been under scrutiny for longer.

Permissions vs Conversation. At the moment, data collection via conversation does not offer the same level of transparency compared to data collected via the Alexa API. This is mostly because data collected through the Alexa API is enforced by permissions and this way users can easily withdraw their consent. In fact, some

skills direct users to visit Alexa companion apps to grant it access to the personal data they need. A good example is the "Barkibu" by *Barkibu*, which says "In order to send you an email report at the end of the consultation process, Barkibu will need access to your email address [...] Visit home screen in your Alexa app and grant me permission." We note that the vast majority of skills collecting data via conversation lack a privacy policy. Instead, those collecting via Alexa API permissions have a much higher proportion of complete traceability. This may suggest that there is stricter scrutiny when developers use the API to collect data. However, the API currently supports a limited number of permissions, and developers may require other information like *age, gender,* etc. Also, there may be questions about the user experience when forced to use other modalities than voice, and/or the usability of such controls.

Unconvincing justifications and control over flows. We identify over-privileged skills in 7% of 100 randomly selected skills we manually interacted with from the set of 1,183 skills with complete traceability in 2021. While these skills state the data they collect and justify their use, this justification may not be compelling. One good example is the "Artificial Intelligence (AI) Facts" skill developed by *by rbashish* in the UK market. According to the description page, the skill tells users facts and figures about AI. The skill requests access to the user address with the pretext to offer a better service. While the skill exhibits complete traceability (it acknowledges collecting the data and its purpose), this collection seems not needed: the skill only answers trivial contextless questions, giving the same answers regardless of the location given. Note that the skill is a single interaction skill that terminates after performing one task.

Our findings show that even when a skill adequately discloses their data practices, there are over-privileged skills. This means that using state-of-the-art tools like SkillVet [12] and SkillExplorer [16] may not be enough, and there is a need for more sophisticated mechanisms to detect this threat model. Hence, the research community, especially Amazon, should look beyond traceability and consider data relevance in the skill review process. Future research should also look into how to implement a framework that let users state their desired data flow patterns. Similar frameworks have been effectively deployed in smartphones for IoT apps [14, 18], and recent work has studied users' desired data flow in SPA [2].

Limitations. Although the study has successfully highlighted how the Alexa skill ecosystem has evolved over three years, it has certain limitations. One important limitation is that we only conduct the traceability analysis for English-speaking marketplaces, as the traceability model is trained with English-speaking privacy policy statements. However, the English market represents over 80% of the skills, so our results cover a very significant part of the ecosystem. In addition, the automated analysis tools we use rely on NLP and ML and thus, inherit their limitations. Nevertheless, we believe high accuracy (93%) achieved by the traceability model and comprehensive coverage level (81%) of SkillExplorer [16] is a good starting point for a meaningful analysis. Unfortunately, SkillExplorer only works well with skills that have unique invocation utterances. It is challenging to explicitly specify which skill to invoke when many skills use the same invocation utterance. While we ensure that the skill of interest is enabled before invoking it, this workaround does not always work as Alexa will only invoke one of the skills based on its predefined algorithm. This implies that even for the best effort, some skills will still not be activated.

Finally, we focus on the Alexa skill ecosystem as it has, by far, the highest number of skills when compared with other SPA like Google assistant, Siri and Cortana [19], but follow-up work on those SPA is a much needed avenue for future research.

8 RELATED WORK

Skill Measurement, Privacy and Security. In this paper, we conduct the first longitudinal study of skills across time. Previous measurements such as [12, 16, 21, 22] just provided a snapshot in time. In [21], the authors crawled skills across seven markets, studied the feasibility of conducting squatting attacks, and provided an initial look at privacy policy effectiveness. In [12], we perform a of developers' practices, including how they collect and justify the need for sensitive information by designing a methodology to identify over-privileged skills with broken privacy policies. Researchers in [22] also used skills description as a baseline to detect inconsistent privacy policies, but not the actual data permissions collected by the skills through Alexa API. Likewise, the authors in [16] conducted a measurement study to understand skills' behaviour by building an interactive system, which identifies skills that request personal information through conversations bypassing developer specifications. Another stream of work like [34] assesses the attack surface of SPA by looking into the sensitivity of the voice commands the SPA skills accept. We do not just look at how sensitive a voice command is; we also look into whether such a command is intended to collect personal information from the user. In addition, we check how well the skill discloses its data collection practices.
Privacy traceability analysis. The traceability analysis is related to our previous work in [12] that focuses on identifying poor data disclosure practices by third party skill developers in their privacy policies. It presents an automated tool called *SkillVet* that leverages machine learning and natural language processing techniques to identify skill traceability. Our previous work also motivates the problem behind collecting permissions using bad privacy practices. This study extends our previous works by shedding light into how traceability has changed over time and what brings about these changes. Other studies have also looked at privacy traceability in areas such as Online Social Networks [9, 10], Social Media Aggregators [25], and Smartphone Applications [39].

9 CONCLUSION

We measured the Amazon Alexa data practices across three years, highlighting how the ecosystem has evolved. We examined the developers' data disclosure practices and presented the landscape in the third-party ecosystem. While the overall ecosystem has improved, newly submitted skills still pose an important risk to users' privacy. The vetting in Amazon marketplace appears to suffer from important flaws, although the research community has made a commendable effort to improve the market's sanitation. Amazon would benefit from adopting actionable mechanisms for researchers to study and analyze privacy and security issues in Alexa. As future work, we would like to expand our understanding of the privacy practices in this domain and devise usable ways of improving them.

Acknowledgments

This research was funded by EPSRC under grant EP/T026723/1 and the "Ramon y Cajal" Fellowship RYC-2020-029401-I. Edu was supported by the PTDF for his PhD.

REFERENCES

[1] Noura Abdi, Kopo Ramokapane, and Jose Such. 2019. More than Smart Speakers: Security and Privacy Perceptions of Smart Home Personal Assistants. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*. USENIX Association, Santa Clara, CA, 451–466.

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

[3] Amazon. 2019. Conditions of Use & Sale. https://www.amazon.co.uk/gp/help/customer/display.html?nodeId=GLSBYFE9MGKKQXXM. [Online; last accessed June-2021].

[4] Amazon. 2019. Security Testing for an Alexa Skill. https://developer.amazon.com/docs/custom-skills/security-testing-for-an-alexa-skill.html. [Online; last accessed 03-July-2019].

[5] Amazon. 2020. Build Skills with the Alexa Skills Kit. https://developer.amazon.com/en-US/docs/alexa/ask-overviews/build-skills-with-the-alexa-skills-kit.html. [Online; last accessed 15-October-2020].

[6] Amazon. 2020. Host a Custom Skill as a Web Service. https://developer.amazon.com/en-US/docs/alexa/custom-skills/host-a-custom-skill-as-a-web-service.html. [Online; last accessed 29-May-2021].

[7] Tawfiq Ammari, Jofish Kaye, Janice Y Tsai, and Frank Bentley. 2019. Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Trans. Comput. Hum. Interact.* 26, 3 (2019), 17–1.

[8] Benjamin Andow, Adwait Nadkarni, Blake Bassett, William Enck, and Tao Xie. 2016. A study of grayware on google play. In *2016 IEEE Security and Privacy Workshops (SPW)*. IEEE, 224–233.

[9] Pauline Anthonysamy, Matthew Edwards, Chris Weichel, and Awais Rashid. 2016. Inferring Semantic Mapping Between Policies and Code: The Clue is in the Language. In *Proceedings of the 8th International Symposium on Engineering Secure Software and Systems - Volume 9639* (London, UK) *(ESSoS 2016)*. Springer-Verlag, Berlin, Heidelberg, 233–250. https://doi.org/10.1007/978-3-319-30806-7_15

[10] Pauline Anthonysamy, Phil Greenwood, and Awais Rashid. 2013. Social networking privacy: Understanding the disconnect from policy to controls. *Computer* 46, 6 (2013), 60–67.

[11] Long Cheng, Christin Wilson, Song Liao, Jeffrey Young, Daniel Dong, and Hongxin Hu. 2020. Dangerous Skills Got Certified: Measuring the Trustworthiness of Skill Certification in Voice Personal Assistant Platforms. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*. 1699–1716.

[12] Jide Edu, Xavi Ferrer Aran, Jose Such, and Guillermo Suarez-Tangil. 2021. SkillVet: Automated Traceability Analysis of Amazon Alexa Skills. *IEEE Transactions on Dependable and Secure Computing* (2021), 14. https://doi.org/10.1109/TDSC.2021.3129116

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

[14] Earlence Fernandes, Justin Paupore, Amir Rahmati, Daniel Simionato, Mauro Conti, and Atul Prakash. 2016. FlowFence: Practical Data Protection for Emerging IoT Application Frameworks. In *25th USENIX Security Symposium (USENIX Security 16)*. USENIX Association, Austin, TX, 531–548. https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/fernandes

[15] Julien Gamba, Mohammed Rashed, Abbas Razaghpanah, Juan Tapiador, and Narseo Vallina-Rodriguez. 2020. An analysis of pre-installed android software. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1039–1055.

[16] Zhixiu Guo, Zijin Lin, Pan Li, and Kai Chen. 2020. SkillExplorer: Understanding the Behavior of Skills in Large Scale. In *29th USENIX Security Symposium (USENIX Security 20)*. USENIX Association, 2649–2666. https://www.usenix.org/conference/usenixsecurity20/presentation/guo

[17] D. Neuburger Jeffrey. 2021. Supreme Court Vacates LinkedIn-HiQ Scraping Decision, Remands to Ninth Circuit for Another Look. https://www.natlawreview.com/article/supreme-court-vacates-linkedin-hiq-scraping-decision-remands-to-ninth-circuit

[18] Yunhan Jia, Qi Alfred Chen, Shiqi Wang, Amir Rahmati, Earlence Fernandes, Zhuoqing Mao, and Atul Prakash. 2017. ContexIoT: Towards Providing Contextual Integrity to Appified IoT Platforms. NDSS. https://www.ndss-symposium.org/wp-content/uploads/2017/09/ndss2017-08-2-jia_slides.pdf

[19] Bret Kinsella. 2019. Google Assistant Actions Total 4,253 in January 2019. https://voicebot.ai/2019/02/15/google-assistant-actions-total-4253-in-january-2019-up-2-5x-in-past-year-but-7-5-the-total-number-alexa-skills-in-u-s/. [Online; last accessed June-2021].

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

[21] Christopher Lentzsch, Sheel Jayesh Shah, Benjamin Andow, Martin Degeling, Anupam Das, and William Enck. 2021. Hey Alexa, is this Skill Safe?: Taking a Closer Look at the Alexa Skill Ecosystem. In *Network and Distributed Systems Security (NDSS) Symposium*. NDSS. https://www.ndss-symposium.org/wp-content/uploads/ndss2021_5A-1_23111_paper.pdf

[22] Song Liao, Christin Wilson, Long Cheng, Hongxin Hu, and Huixing Deng. 2020. Measuring the Effectiveness of Privacy Policies for Voice Assistant Applications. In *Annual Computer Security Applications Conference* (Austin, USA) *(ACSAC '20)*. Association for Computing Machinery, New York, NY, USA, 856–869. https://doi.org/10.1145/3427228.3427250

[23] David Major, Danny Yuxing Huang, Marshini Chetty, and Nick Feamster. 2021. Alexa, Who Am I Speaking To?: Understanding Users' Ability to Identify Third-Party Apps on Amazon Alexa. *ACM Transactions on Internet Technology (TOIT)* 22, 1 (2021), 1–22.

[24] Nicole Meng, Dilara Keküllüoğlu, and Kami Vaniea. 2021. Owning and Sharing: Privacy Perceptions of Smart Speaker Users. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–29.

[25] Gaurav Misra, Jose Such, and Lauren Gill. 2017. A Privacy Assessment of Social Media Aggregators. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017* (Sydney, Australia) *(ASONAM '17)*. Association for Computing Machinery, New York, NY, USA, 561–568. https://doi.org/10.1145/3110025.3110103

[26] Alexios Mylonas, Marianthi Theoharidou, and Dimitris Gritzalis. 2014. Assessing Privacy Risks in Android: A User-Centric Approach. In *Risk Assessment and Risk-Driven Testing*, Thomas Bauer, Jürgen Großmann, Fredrik Seehusen, Ketil Stølen, and Marc-Florian Wendland (Eds.). Cham, 21–37.

[27] Sarah Perez. 2016. Amazon Alexa now has over 1,000 Skills, up from 135 in January. https://techcrunch.com/2016/06/03/amazon-alexa-now-has-over-1000-skills-up-from-135. [Online; last accessed May-2020].

[28] Alexander Ponticello, Matthias Fassl, and Katharina Krombholz. 2021. Exploring Authentication for Security-Sensitive Tasks on Smart Home Voice Assistants. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*. USENIX Association, 475–492.

[29] Amanda Purington, Jessie G. Taft, Shruti Sannon, Natalya N. Bazarova, and Samuel Hardman Taylor. 2017. "Alexa is My New BFF": Social Roles, User Satisfaction, and Personification of the Amazon Echo. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI EA '17)*. ACM, New York, NY, USA, 2853–2859. https://doi.org/10.1145/3027063.3053246

[30] Cornell Law School. 2021. VAN BUREN v. UNITED STATES 940 F. 3d 1192. https://www.law.cornell.edu/supremecourt/text/19-783

[31] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I Hong. 2018. " Hey Alexa, What's Up?" A Mixed-Methods Studies of In-Home Conversational Agent Usage. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 857–868.

[32] Gian Luca Scoccia, Anthony Peruma, Virginia Pujols, Ivano Malavolta, and Daniel E. Krutz. 2019. Permission Issues in Open-Source Android Apps: An Exploratory Study. In *2019 19th International Working Conference on Source Code Analysis and Manipulation (SCAM)*. 238–249.

[33] Filipo Sharevski, Peter Jachim, Paige Treebridge, Audrey Li, Adam Babin, and Christopher Adadevoh. 2021. Meet Malexa, Alexa's malicious twin: Malware-induced misperception through intelligent voice assistants. *International Journal of Human-Computer Studies* 149 (2021), 102604.

[34] Faysal Hossain Shezan, Hang Hu, Jiamin Wang, Gang Wang, and Yuan Tian. 2020. Read Between the Lines: An Empirical Measurement of Sensitive Applications of Voice Personal Assistant Systems. In *Proceedings of The Web Conference 2020*. Association for Computing Machinery, New York, NY, USA, 1006–1017. https://doi.org/10.1145/3366423.3380179

[35] Stanfordnlp. 2021. https://stanfordnlp.github.io/CoreNLP/. Accessed May 2021].

[36] Jose Such. 2017. Privacy and autonomous systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*. 4761–4767.

[37] Lionel Sujay Vailshery. 2021. Number of digital voice assistants in use worldwide from 2019 to 2024 (in billions). https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/. [Online; last accessed May-2021].

[38] N. Zhang, X. Mi, X. Feng, X. Wang, Y. Tian, and F. Qian. 2019. Dangerous Skills: Understanding and Mitigating Security Risks of Voice-Controlled Third-Party Functions on Virtual Personal Assistant Systems. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1381–1396.

[39] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N. Cameron Russell, and Norman Sadeh. 2019. MAPS: Scaling Privacy Compliance Analysis to a Million Apps. *Proceedings on Privacy Enhancing Technologies* 2019, 3, 66–86. https://doi.org/10.2478/popets-2019-0037

A DATA STATEMENT

To support other researchers interested in repeating and reproducing our measurements, we make our dataset publicly available as well as our implementation of [16] at https://github.com/xfold/Are-We-There-Yet-Alexa-Market-Comparison

B SKILLS AND DEVELOPERS

Table 8 shows the breakdown of the total number of Alexa skills across Amazon marketplaces. From the table, it can be seen that there are 124,026 skills published in 2021. This is 10.94% higher than the 111,796 skills published in 2020 and 46% higher than the 84,856 skills published in 2019. In addition, more skills were published between 2019 and 2020 (26,940 skills) than between 2020 and 2021 (12,230). Across all years, English-speaking marketplaces have the largest skills, representing over 80% of the skills. The Spanish market has the highest increment in the number of skills changing by almost 300% from 1,286 in 2019 to 5,435 in 2021. Likewise, we see more skills in the IN marketplace in 2020 (31,246) compared with 28,672 in 2021. Overall, there is a high percentage increase in the number of skills added to the non-English speaking markets (71%) than the English speaking markets (25%) over the years.

Likewise, there are 50,526 developers in 2021. This is an 8% increase from the 46,804 recorded in 2020. Overall, there is a 62% rise in the number of developers we see from 2019 to 2021. As we see with the total number of skills during the years, the Spanish marketplace also has the highest increment in developers changing by a similar percentage of 300%. From the table, it can be seen that the highest number of developers is also located within the English-speaking marketplaces.

Table 8: Number of Skills and Developers from 2019 to 2021.

	2021		2020		2019	
Market	Skills	Devs	Skills	Devs	Skills	Devs
US	68,667	29,394	55,736	25,483	51,338	19,507
UK	37,056	15,998	34,618	15,066	29,094	12,078
IN	28,672	11,781	31,246	13,316	20,989	9,197
CA	27,093	11,662	26,027	11,509	24,700	10,773
AU	24,512	11,603	24,062	10,762	23,123	10,123
DE	10,631	4,018	10,287	3,713	8,928	3,165
ES	5,435	2,856	5,010	2,543	1,286	716
IT	4,649	2,331	4,203	2,049	2,210	1,095
JP	3,637	1,437	3,545	1,377	2,679	1,056
FR	2,863	1,407	2,288	1,194	1,341	641
MX	2,486	1,563	1,972	1,212	897	540
Total	215,701	94,050	198,994	88,224	166,585	68,891
Unique	124,026	50,526	111,796	46,804	84, 856	31,238

C SKILL CATEGORY

We look at the skill categories to understand how skills are grouped and study what changes have occurred over time, specifically by looking at the newly added and removed skills across the years. Table 9 shows by category the number of skills that have been added to the ecosystem over the years. Also, it shows the skills that have been removed from the ecosystem over the years. We see that the *Game & Trivia* is the category with the highest number of newly added skills between 2019 and 2020. It contains about 27% of the total added skills. Also, *Music & Audio* is the top category with the highest number of newly added skills between 2020 and 2021. It contains 6,796 (25.5%) new skills.

Table 9: Number of skills per category added and removed across the years.

Category	2020-2021		2019-2020	
	New	Removed	New	Removed
Business & Finance	726	461	1971	297
Connected Car	74	47	72	4
Education & Reference	1940	1314	4555	1490
Food & Drink	369	390	847	146
Games & Trivia	4339	2499	10832	2337
Health & Fitness	543	380	1161	1168
Kids	578	293	1718	159
Lifestyle	3310	734	2979	2925
Local	87	36	309	46
Movies & TV	106	87	339	72
Music & Audio	6830	4367	5750	1755
News	3347	927	2840	2942
Novelty & Humor	758	555	1167	365
Productivity	621	302	1092	98
Shopping	169	110	255	38
Smart Home	1018	1093	2033	316
Social	571	188	1289	118
Sports	367	216	604	125
Travel & Transportation	270	268	703	125
Utilities	619	218	886	119
Weather	603	530	247	64

Looking at the removed skills over the years, we see that the *Music & Audio* category has the highest number of removed skills across the marketplaces. In 2021, this category contains more than 50% of the total skills removed in the US market with respect to 2020. We also examine the interplay between skills added into a category and the number of removed skills across time. As shown in Table 9, we see that *Smart Home*, and *Food & Drink* categories in 2021 have more skills removed than the number of skills added. Overall, fewer skills are removed between 2019 and 2020 than the number of publications between 2020 and 2021.

D TRACEABILITY BY NUMBER OF PERMISSIONS

To establish whether skills that request more permission are more traceable or not, we study the relationship between the number of permission requested by skills and their traceability. The data in Table 10 shows that there is a higher number of skills with complete traceability, asking for one or three permissions. Likewise, skills that request for 5 or more permissions tends to have a lower number of skills (8%) with broken traceability. It is apparent that over the years, there is no correlation between the number of permissions requested by skills and their type of traceability.

E SKILLEXPLORER TOOL IMPLEMENTATION AND EVALUATION

As in [16], our implementation follows a black-box approach to interact with skills, since the skills' code or executable is not available — recall that skills *run in the cloud instead of the users' device*, e.g., as an AWS Lambda function [5] or in a server controlled by the skill developer [6], and the only way to interact with them online

Table 10: shows the relationship between the number of permissions requests by skills and their traceability.

Perm Length	Traceability	2021		2020		2019	
		N	%	N	%	N	%
1	C	841	69%	590	51.7%	346	53.1%
	B	261	21%	474	41.5%	232	35.6%
	P	123	10%	77	6.7%	74	11.3%
	Total	1,225	100%	1,141	100.0%	652	100.0%
2	C	173	48%	103	35.6%	62	50.4%
	B	97	27%	129	44.6%	53	43.1%
	P	89	25%	57	19.7%	8	6.5%
	Total	359	100%	289	100.0%	123	100.0%
3	C	105	60%	68	48.2%	35	66.0%
	B	29	17%	40	28.4%	14	26.4%
	P	41	23%	33	23.4%	4	7.5%
	Total	175	100%	141	100.0%	53	100.0%
4	C	52	80%	35	64.8%	3	50.0%
	B	7	11%	14	25.9%	2	33.3%
	P	6	9%	5	9.3%	1	16.7%
	Total	65	100%	54	100.0%	6	100.0%
>=5	C	12	48%	4	40.0%	1	25.0%
	B	2	8%	2	20.0%	1	25.0%
	P	11	44%	4	40.0%	2	50.0%
	Total	25	100%	10	100.0%	4	100.0%

is via a conversation. Our tool comprises four key components. i) The utterance extraction where it extracts utterances from the skill page to initiate the conversation, ii) the question understanding section, to understanding the response from the SPA, iii) the answer generation unit that generates a suitable answer to the question extracted from the SPA response for further interaction, and lastly iv) the behavior exploration component that ensures that all routes of conversation are explored.

The utterance extraction: We extract the sample invocation utterances from the skill web page to activate the skill and initiate the interaction with Alexa. Developers are requested to provide sample utterances questions to help users understand how to use the skill. These can be located by looking at the "a2s-utterance-box-inner" tag in the source code of the skill web page on the Alexa store.

The question understanding unit: After the first extracted utterance is sent to a skill, Alexa responds with the feedback and output from the skill. The feedback could be an answer to a request or a request for further commands. Our tool is implemented in a way that it could adequately understand the type of feedback given by Alexa. To understand the response from Alexa, we use Standard CoreNLP parser [35] to process the response as it considers clause, phrase, and word level when generating the abstract syntax tree from a text. This allows detecting patterns within the text at a lower level which can help identify and categorize specific questions.

We consider five different types of questions: i) Wh-Questions – These types of questions are open questions that users answer based on their understanding. An example of this question is, "Tell me your firstname?"; ii) Yes/No questions – these are questions that expect "yes" or "no" answers. Examples include questions such as "Did you mean Lite Rock 105?", "Do you want to listen to another fact?" that could be answered by responding with either "yes" or "no"; iii) Instruction Questions – this type of question contains instructions on how to answer them. It commonly includes the word "say" or "ask". An example of this question type is "Please say repeat to hear the question again", where the user is instructed to say "repeat"; iv) Selection Questions – this type of question gives users options from where they can select. An example is "To get started; you can get a quote or listen to the daily briefing". Here, the user has two options to select from when generating their response; v) Mix questions – this type of question contains more than one type of the other question type. For example, the question "Please ask me for a cryptocurrency price by saying, what is the price of bitcoin? Or, tell me the price of Ethereum" comprises Wh-question, instruction question, and selection question.

The answer generation: After categorizing the questions, we generate a suitable answer for the question type. The answer to be developed need to keep the conversation going as much as possible. We can directly extract the answer from the questions themselves for the instruction, selection, and Yes/No questions. However, for the wh-question, we create a knowledge database to answer the question and explore the skill behavior. We likewise leverage kuki chat-bot (https://chat.kuki.ai/chat) due to its performance to answer other wh-question that are not cover in the knowledge database. Regarding the Mix question where multiple questions were detected, we prioritize answers as follows. If selection question and instruction exist simultaneously, we process both questions; if the Yes/No question exists, we answer with "yes" or "no".

The behavior exploration: for a specific Alexa response, there could be multiple answers. To ensure that all conversation flow routes are explored before moving onto the following utterance, we use a tree data structure to represent the exploration status and track which question has been visited. Each node of the tree is a single interaction that comprises an Alexa response and the generated answer. When the tool interacts with the skill, the tree is drawn simultaneously. Thus, the tool ensures that every execution path is explored and all nodes are visited.

We use the Alexa simulator in the developer console for the interaction. The simulator allows developers to test their skills as they can directly feed text input into a skill and observe its outputs. For a more detailed explanation of the interactive system, we refer the readers to [16]. An important observation is that while the interactive tool automatically enabled the skill on the Alexa store before invocation, Alexa still has issues understanding some of the skill invocation sample utterances. For instance, when we invoke the skill "Little Figure Skater Test" by *Modal Systems Ltd* with the sample utterance "Alexa, Start little figure skater test", Alexa responded with "Hmm, I don't know that one".

Evaluation: To check for the accuracy of our implementation, we conduct the same evaluation reported in [16]. In particular, we randomly selected 50 skills from different categories and manually interacted with them. The interaction generated 61 Mix questions, 14 Wh questions, 18 Yes/No questions, 11 Selection questions and 15 instruction questions and lasted for 5 hours. We then compared the output from the manual interaction with that generated by the interactive tool. The tool generates 97 outputs, which is 22 outputs less than the outputs from the manual interaction. The coverage implies that our tool has 81% coverage, similar to the coverage reported in [16]. Regarding the answer generation accuracy, all the Yes/No answers are correctly identified, and 9% of Mix questions were wrongly identified. On average, only 7% of answers are wrong.