Smart Home Privacy Policies Demystified: A Study of Availability, Content, and Coverage

Sunil Manandhar and Kaushal Kafle, *William & Mary;* Benjamin Andow, *Google LLC;* Kapil Singh, *IBM T.J. Watson Research Center;* Adwait Nadkarni, *William & Mary*

https://www.usenix.org/conference/usenixsecurity22/presentation/manandhar

This paper is included in the Proceedings of the 31st USENIX Security Symposium.

August 10–12, 2022 • Boston, MA, USA

978-1-939133-31-1

Open access to the Proceedings of the 31st USENIX Security Symposium is sponsored by USENIX.

Smart Home Privacy Policies Demystified: A Study of Availability, Content, and Coverage

Sunil Manandhar[§], Kaushal Kafle[§], Benjamin Andow [† *], Kapil Singh[‡], Adwait Nadkarni[§]
[§] *William & Mary; {sunil, kkafle, nadkarni}@cs.wm.edu*
[†] *Google LLC; andow@google.com*
[‡] *IBM T.J. Watson Research Center; kapil@us.ibm.com*

Abstract

Smart home devices transmit highly sensitive usage information to servers owned by vendors or third-parties as part of their core functionality. Hence, it is necessary to provide users with the context in which their device data is collected and shared, to enable them to weigh the benefits of deploying smart home technology against the resulting loss of privacy. As privacy policies are generally expected to precisely convey this information, we perform a systematic and data-driven analysis of the current state of smart home privacy policies, with a particular focus on three key questions: (1) how hard privacy policies are for consumers to obtain, (2) how existing policies describe the collection and sharing of device data, and (3) how accurate these descriptions are when compared to information derived from alternate sources. Our analysis of 596 smart home vendors, affecting 2,442 smart home devices yields 17 findings that impact millions of users, demonstrate gaps in existing smart home privacy policies, as well as challenges and opportunities for automated analysis.

1 Introduction

Privacy concerns are an important hinderance in the adoption of smart home devices [57, 79, 88]. These concerns are well-founded, because smart home devices frequently transmit operational data to remote servers to enable functions ranging from basic status updates to trigger-action automation, which may contain *highly private evidence of the user's activities in their personal environment*. For example, when a door lock notifies the user of its status change by the way of sharing updates with its remote server, the lock's state may also be used by the vendor to track the user's schedule. As device data may be used to profile the user, it is fair to expect that users deserve to be informed about how their information may be used, *through a legally binding document*, which will build user confidence and increase smart home adoption.

A privacy policy performs this function, as it is the primary, legally binding, medium for conveying the data collection and sharing practices of an organization to users. In fact, prior work in Web and mobile privacy has attempted to understand the privacy posture of organizations by analyzing their privacy policies [7, 8, 81, 97], often revealing significant contradictions in the text, and inconsistencies with actual practices. With the goal of understanding the baseline privacy guarantees promised to smart home users, we initially pursued a similar route by analyzing the privacy policies of popular smart home vendors using state-of-the-art tools [7, 8]. However, our initial investigation encountered several obstacles due to the intricate (and *under-studied*) aspects of the smart home domain, which make *automated analysis infeasible at present*.

First, we observed that smart home privacy policies, i.e., vendor-provided policies that apply to all the smart home devices/services they produce (and device data they manage), are not as easily *available* for analysis as compared to mobile app or website privacy policies. That is, while it is relatively easy to find mobile app privacy policies on app markets that mandate them (e.g., Google Play [40]), we had to scour several sources (e.g., vendor websites, app markets, mobile apps) for policies that apply to smart home devices, even for popular vendors (e.g., Honeywell). Second, we realized that even if we could obtain the policies, existing tools would be unable to capture the contextual privacy implications of device data in them, without sufficient insight into their *content* and *organization*-related intricacies. Particularly, the various types of smart home device data and the contexts in which they are described are relatively unknown compared to typical private data that existing tools (e.g., PolicyLint [7]) are trained to analyze, as we elaborate in Section 7. Finally, we observed the need to assess policy *coverage* in terms of describing the collection/sharing practices for all the device data that the vendor collects. Without this knowledge, it would be difficult to reason about the results of an automated policy analysis. To summarize, our initial investigation revealed that it may not be trivial to directly analyze smart home privacy policies with existing tools, as the smart home is a separate application domain with its own intricacies, motivating this data-driven evaluation to uncover insights that would enable automated analysis tools for this domain in the near future.

[*] This work was completed when the author was at IBM Research.

Contributions: We describe the first large-scale evaluation of *smart home (device) privacy policies*, i.e., vendor-provided privacy policies that apply to smart home devices/services, performed with the goal of (1) understanding their understudied characteristics, and (2) extracting critical insights for guiding the development of practical analyses, targeted regulations, and end-user tools, for the holistic improvement in the privacy of the smart home ecosystem. In doing so, we develop an empirical foundation for privacy disclosure analysis in the smart home, through the following contributions:

- **Large-scale Study of Home Privacy Policies:** We present the first large-scale study of smart home privacy policies aimed at demystifying the unique characteristics of the domain along three areas: (1) *availability*, (2) *content*, and (3) *coverage*. We study the privacy policies of *vendors* integrated with 7 most popular smart home platforms, i.e., 596 vendors representing 2,442 smart home devices.
- **Systematic Study Methodology and Datasets:** We develop a semi-automated study methodology that thoroughly investigates the state of smart home device privacy policies, and yields insights that lay the groundwork for developing automated methods. We constructed several datasets that will be useful for developing automated tools that can reason about device data, including a labeled set of 284 device privacy policies and a precise vendor-device_type map describing the devices sold by the 596 vendors. The datasets are available in our github repository [87].
- **17 Novel and Impactful Findings:** Our findings demonstrate severe gaps in the current state of smart home privacy policies. Particularly, device privacy policies are extremely hard to obtain ($\mathcal{F}_1$–$\mathcal{F}_5$); e.g., we were able to obtain policies that apply to smart home devices for only 48.99% of the studied vendors ($\mathcal{F}_3$), with 10.57% not providing privacy policies at all ($\mathcal{F}_1$). Further, policies do not precisely describe device data ($\mathcal{F}_6$, $\mathcal{F}_{10}$, $\mathcal{F}_{11}$), and if they do, the descriptions are often inconsistent with actual state of data potentially collected by vendors ($\mathcal{F}_{13}$–$\mathcal{F}_{16}$). Our findings impact vendors whose products are used by millions as indicated by our impact metrics. We have disclosed our findings to all affected vendors (see Appendix 8).
- **Describing the need for (and path to) contextualization:** This study demonstrates a clear need for new policy analysis frameworks that are *focused on the characteristics of the smart home*, just as new frameworks were warranted for mobile apps. We show that existing NLP-based tools may incorrectly reason about a majority of the policies in our dataset ($\mathcal{F}_8$), and experimentally demonstrate the need to consider the smart home as a separate problem domain, and the tangible benefits from *contextualizing existing tools* ($\mathcal{F}_{17}$). Finally, we describe how our methodology, the labeled dataset, and the insights from this study, serve as a starting point for developing contextualized automated policy analyses for smart homes.

2 Motivation

Smart home *device data* consists of events observed or facilitated from internet-connected motion sensors, door locks, and cameras that form direct evidence of user activity; e.g., the door lock unlocks when the user comes home, and lights turn OFF when it's the user's bedtime. Such data can be used to infer incredibly private user behavior and profile users without consent, leading to real-world consequences. For example, an insurer could use device data from a water flow sensor to infer a maintenance issue (e.g., a small leak), and deny any future claims associated with water-damage. This example is not hypothetical; in November 2020, Yonomi and LexisNexis Risk Solutions announced a partnership wherein they would build IoT solutions to enable insurance companies to "discover devices in the home and share data from those devices with the insurer", which the insurer would use to affect claims [44]. This means that device data will provide entities such as insurers with a persistent window into the user's home, behavior, and lifestyle habits, i.e., as aptly stated by LexisNexis's director of IoT: *"We're moving from having a snapshot in time to an ongoing assessment of what's happening in a home"* [44].

Considering the risk of intrusive behavior profiling, it is critical to ensure that smart home device vendors disclose the privacy implications of using their devices to consumers through accurate privacy policies, in order to obtain *informed consent*. With the goal of understanding the data-use practices associated with smart home device data, we first set out to *automatically* analyze smart home privacy policies by using prior NLP-based frameworks for analyzing mobile/Web privacy policies (i.e., PolicyLint [7]). However, we encountered three challenges (described as follows) in terms of the relatively under-studied *availability*, *content*, and *coverage* of smart home policies, which not only make automated analysis infeasible, but also hard for consumers to obtain or reason about policies, motivating our empirical study.

1. Erratic provisioning of privacy policies: We observed that automatically acquiring smart home privacy policies at scale is non-trivial, due to the disparity in how the policies are distributed to consumers, if at all. Smart home devices can be purchased independently of any virtual marketplaces (e.g., the vendor's website, online stores such as Amazon). Hence, there is no central repository of privacy policies that apply to smart home devices, unlike domains such as mobile apps analyzed by prior work where app markets mandate policy links [39]. Thus, it is infeasible to develop an automated mechanism for obtaining such policies at scale, without first studying *how* the policies that apply to *devices* are provisioned to users, which motivates our first research question:

> **RQ**$_1$: *How difficult is it for consumers to obtain privacy policies that apply to their smart home devices?*

2. Unexplored device data: While prior work on mobile privacy policy analysis generally focuses on PII and other

well-studied private data types (e.g., SSN, credit card number), the range of privacy-sensitive smart home device data (e.g., motion detector/door lock status), and how it is described in policies, are relatively unexplored. For instance, we observed that certain policies describe the collection of device data for a broad category called "device data", while others are more precise and describe specific device data such as "camera stream" or "motion". Thus, we would need to know *how and in what contexts smart home data is described in these policies* to develop automated tools for analyzing them, which motivates our second research question:

RQ$_2$: *How precisely is the collection and sharing of device data described in smart home product privacy policies?*

3. The unknown coverage of smart home device policies: While it may be possible to avoid privacy risks in mobile apps by preventing them from collecting/transmitting data (e.g., IMEI, location), transmission of smart home device data is unavoidable, as it is necessary for enabling the inherently connected functionality (e.g., third-party integration, remote-control). Hence, it is important for privacy policies to disclose the collection and sharing practices *for each device and device data object* that the vendor sells, to provide the consumer with a complete perspective of what is at stake. Analyzing coverage may require out-of-band ground-truth (e.g., vendor-specific lists of devices), which may be non-trivial to automatically acquire, motivating. our third and final research question:

RQ$_3$: *How comprehensive are smart home product privacy policies in describing the collection/sharing of device-data?*

3 Study Overview

We address **RQ$_1$**–**RQ$_3$** using a semi-automated, data-driven methodology that enables us to develop a grounded understanding of the current state of smart home policies, as illustrated in Figure 1. We now provide a brief overview of this methodology, followed by a summary of the metrics we consider to approximate the impact of our findings.

1. Policy Availability Analysis (RQ$_1$, Section 4): As shown in Figure 1, we begin by identifying a representative set of smart home *vendors* whose policies we seek to study, by scraping the *integrations* advertised on the websites/marketplaces hosted by 7 popular smart home platforms. We subsequently perform an exhaustive search for device/product privacy policies for each vendor, which spans several resources, such as vendor websites, search engines, app stores, and mobile apps.

2. Policy Content Analysis (RQ$_2$, Section 5): We begin by defining a set of content labels based on existing regulatory requirements (e.g., GDPR, CCPA) as well as intuition gained from an exploratory pass over a subset of the dataset. We then label the device privacy policies using open coding, and identify several key semantic traits, as well as gaps in how vendors define device data usage.

3. Policy Coverage Analysis (RQ$_3$, Section 6): Given a privacy policy that precisely describes device data, we develop a methodology that estimates policy *coverage* in terms of describing data usage for *all types of devices sold by that vendor*. To enable such an analysis, we define a methodology to build a *vendor-device_type* map that describes the device-types (e.g., camera, switch) sold by each vendor whose policies we seek to evaluate, and the minimal data attributes that each device_type exhibits. We use the vendor-device_type map to discover anomalies in coverage, which further motivate the need for using such out-of-band context in policy analysis.

Impact Metrics: We construct 4 metrics to reason about the *impact* of our findings in terms of the popularity of the affected smart home vendors. As there is no public data on the market share of smart home vendors, we assess vendor-popularity using statistics from two representative sources: (1) *Amazon*, as it is a popular source of smart home devices, and (2) *Google Play*, as it often hosts *companion apps* for devices.

Metric 1: Amazon Ratings – This metric represents the most popular device sold by the vendor on Amazon, in terms of the count of ratings, which enables us to estimate the impact of insufficient disclosure even if the vendor only sold one (i.e., their most popular) device. We obtain this metric by searching for the vendor's name on Amazon, and selecting the product with the highest ratings count from the first page returned.

Metric 2: "Best Seller" badge – Amazon assigns a *Best Seller* badge to products that outperform others in the same category. We mark a vendor as a *Best Seller* if at least one of its devices is found to have the badge.

Metric 3: "Amazon's Choice" badge – Amazon recommends certain products to consumers, labeling them as *Amazon's Choice*. Consumers are likely to be steered towards them.We mark a vendor as *Amazon's Choice* if at least one of its devices is found to possess the badge.

Metric 4: Google Play install count – This metric represents the total install-count of the vendor's companion app on Google Play. If the vendor is represented by more than one app, we select the most relevant app with the highest download count, to get an estimate of the most users affected.

We include multiple metrics to provide insights based on various factors that contribute to vendor popularity. These factors include price, quality, sales performance, availability, and usage of vendor products. For instance, Amazon states that "Amazon's Choice highlights highly rated, well-priced products available to ship immediately" [5]. Similarly, Amazon Best Seller badge is provided based on calculation of sales [4], whereas, Amazon Ratings takes into account reviews and verified purchase status [6]. Thus, our impact metrics aim to approximate the popularity of vendors by considering several factors that signify popularity among consumers, as such metrics have been observed to lead to improved sales conversion rates [3]. However, these metrics are not free from the risk of vendor manipulation, as we discuss in Section 10, and hence, their inclusion only indicates importance, and not a quantifi-

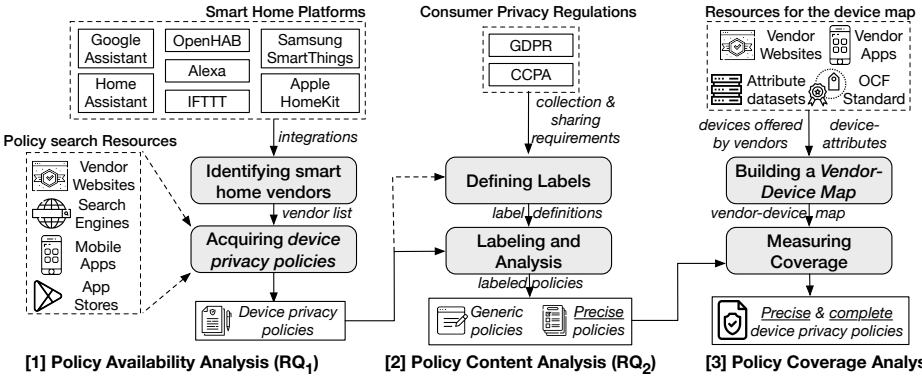


Figure 1: An overview of our systematic methodology for studying the current state of smart home privacy policies ($\mathbf{RQ}_1$–$\mathbf{RQ}_3$).

cation of impact. Finally, note that we used impact metrics to *assess* popularity *after* the conclusion of the study, i.e., the metrics have no impact on the vendors selected for the study.

While our metrics are collected automatically, one of the authors validated each result manually to ensure correctness. Finally, while not all vendors are represented on Amazon or Google Play, these two sources are complementary, and cumulatively ensure coverage of 519/596 (i.e., 87.08%) vendors.

4 Policy Availability Analysis ($\mathbf{RQ}_1$)

This section empirically characterizes the availability of privacy policies from a large set of smart home vendors. The significance of our analysis is two-fold. First, privacy policies that are not easily accessible may not be compliant with regulatory requirements, e.g., CCPA [56] mandates specific ways in which privacy policies must be provisioned. Second, our analysis demonstrates the complex approach needed for locating and acquiring smart home device privacy policies, laying the basis for their automatic acquisition.

4.1 Methodology

We develop a data-driven methodology to drive our semi-automated acquisition of policies, complemented by automatic tools to extract relevant information. Our analysis is organized along two key tasks. First, as privacy policies are generally specific to companies, and not individual devices, we identify a representative set of smart home vendors. Second, we exhaustively locate privacy policies for the said vendors and identify *smart home device privacy policies*, i.e., which apply to smart home devices/products.

Identifying Smart Home Vendors: To acquire a list of relevant vendors we rely on the following intuition: Since the demand for automation and inter-operability is one of the primary drivers of the smart home market, the cumulative list of devices integrated into popular platforms would represent the devices users are most likely to use. Therefore, we systematically compile a list of representative smart home vendors from the integration-lists published by 7 popular automation frameworks, namely Alexa [2], Google Assistant [41], IFTTT [50], SmartThings [85], Apple HomeKit [46], OpenHAB [67], and HomeAssistant [45]. To resolve integrations to brands/vendors, we first normalize the lists (i.e., remove irrelevant terms such as "smart" and "outlet"), and condense the remaining names into individual brands (e.g., Wemo Light Switch and Wemo Coffeemaker to "Wemo"). Then, to identify the primary Web domain for each vendor, we automatically search for the brands on Google, scrape the top ten results, and resolve URLs to vendors by matching the domain name with the vendor name. Finally, we manually confirm the vendor-domain match, and filter out non-smart-home vendors using the website content. We use this methodology to maximize inclusion of vendors that (a) offer a smart home device, and (2) have a website (i.e., are actively advertising products). This process may exclude certain vendors that do not have a web presence but still sell devices (e.g., through brick and mortar stores). This exemption aligns with this study, given its focus on analyzing vendors that provide a mechanism to inform consumers *prior* to purchasing devices, which can only be accomplished through some form of online presence (e.g., Web or mobile app).

Acquiring Device Privacy Policies: Our goal is to obtain policies that apply to *smart home devices* and the data they transmit, i.e., unlike policies that cover other artifacts, such as websites or mobile apps. Hence, we define an *exhaustive* methodology for locating such device privacy policies by searching 4 vendor-resources, in the order that a user *planning to buy* a smart home device is likely to follow, i.e., ordered by the ease of obtaining the policy: (1) the vendor website, (2) a web search, (3) the mobile app store (i.e., Google Play), and (4) the vendor's mobile app.

Our approach for locating privacy policies varies by the resource. For vendor websites, we search for "Privacy Policy" or "Legal" links on the main page, whereas Google search involves searching for a combination of terms, such as the vendor name and web domain, "privacy policy" and "product policy", and filtering search results aligned to the vendor. For mobile app stores, we search with the vendor name and for all apps that are found, we subsequently look for any links to policy documents. If no policy is found and the app has a short registration process, we manually register to see if the privacy

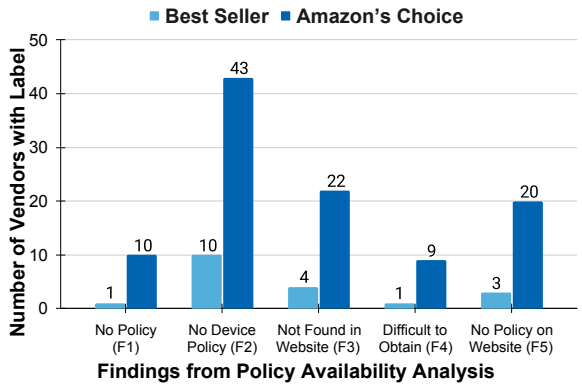


Figure 2: *Amazon's Choice* and *Best Seller* metrics ($\mathcal{F}_1$–$\mathcal{F}_5$).

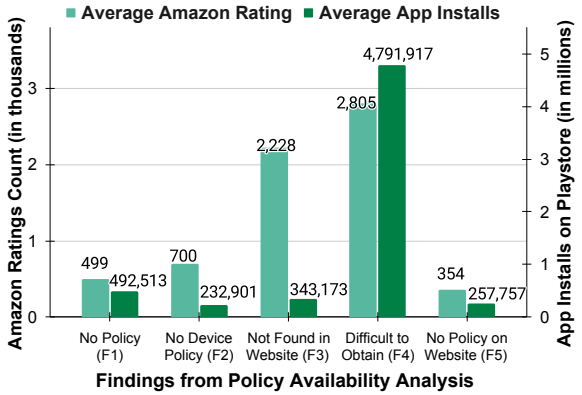


Figure 3: *Amazon Ratings* and *App Install* metrics ($\mathcal{F}_1$–$\mathcal{F}_5$).

policy dialog box is shown after registration. For every policy we collect, we perform the following test to check whether the policy applies to *devices*: We extract the *policy preamble*, and the section describing *data collection*, and check for any reference to "Products" or "devices", either explicit (i.e., as specific terms) or implicit (i.e., by specifying that it "covers everything" for the vendor). We end the search once a device privacy policy is found, or all resources have been examined.

4.2 Policy Availability Findings ($\mathcal{F}_1$–$\mathcal{F}_5$)

Using the methodology in Section 4.1, we obtained 3,678 unique integrations from the 7 popular automation platforms, after removing duplicates across platforms (e.g., Wemo Switch in both SmartThings and IFTTT). These integrations represent unique products/services provided by each vendor across all platforms; however, a vendor may have multiple products (and hence integrations representing them). Thus, we resolved websites for these integrations, and also condensed them into unique brands (e.g., condensing "Wemo Switch" and "Wemo Coffeemaker" to "Wemo") as described in Section 4.1, as our analysis explores the privacy posture of vendors/brands, leading to 1,365 *vendor websites*.

To ensure that we analyze vendors relevant to the smart home, we further manually analyzed the content of each of these 1365 vendor/brand websites to filter out non-smart home brands that fall into one of two categories: (1) vendors that

Table 1: Sources of the 292 device privacy policies.

Source	Number of device policies
Vendor websites	188 (64.38%)
Google Search	41 (14.04%)
Google Play Links	21 (7.19%)
Mobile Apps	42 (14.38%)
Total	292 (i.e., 100%)

do not sell smart home products, but nevertheless integrate their products with smart home platforms, such as Facebook and Evernote [29], and (2) vendors whose websites simply advertise or discuss smart home products belonging to other vendors, or are under construction and/or devoid of any information, such as Mattel [59] (which advertises other vendors' products) and Gidbo [38] (which has no information on its website) respectively. As these categories of vendors do not sell smart home products or provide relevant information that would facilitate privacy analysis, we chose to exempt them from our analysis, resulting in the final dataset of 596 confirmed smart home vendors, which together represent 2442 unique smart home devices (as per vendor websites). As described previously, our approach for deriving this list of 596 vendors is motivated by user demand for automation and interoperability, and is biased towards the products that users are most likely to use as they are integrated into popular platforms.

We located the privacy policies of these 596 vendors using machines in the US. While we did not perform locale-specific analysis, we observed that vendors make the same policies available across locations, with separate sections for geography-specific regulations (e.g., for Europe, CA) (See $\mathcal{F}_8$). We spent anywhere between 2-15 minutes per vendor, with the longest searches generally resulting in failure to find a device privacy policy. We were able to locate device privacy policies for 292/596 vendors. Our analysis of policy availability led to five findings, with considerable impact, illustrated in Figures 2 and 3 as per our impact metrics (Section 3).

Finding 1: No Policy – 10.57%, **i.e.,** 63/596 **of smart home vendors do not provide privacy policies, i.e., not even for their websites** ($\mathcal{F}_1$) – These 63 vendors sell smart home devices belonging to 27 unique types that include privacy-sensitive devices such as security cameras and baby monitors. Our impact metrics demonstrate that consumers are extremely likely to purchase devices from these vendors, with 10/63 labeled as *Amazon's Choice*, and one labeled as *Best Seller* (Figure 2). Moreover, the companion apps from these vendors have over 492*k Google Play Installs* and 499 Amazon ratings on average (Figure 3), which indicates that the lack of privacy policies may have affected a substantial user population. Due to the absence of privacy policies, these vendors potentially violate CCPA (Section 999.305 [20]) if certain additional criteria are satisfied (e.g., annual gross revenue [56]).

Finding 2: No Device Policy – 43.52% **do not have policies that apply to smart home products** ($\mathcal{F}_2$) – We found

that out of the 517 vendors that have at least one privacy policy in English (i.e., excluding 63 vendors with no policy ($\mathcal{F}_1$) and 16 that provided non-english policies), 225 (or 43.52%) do not provide privacy policies that discuss their smart home products or devices, but only discuss website, mobile app, or account-related data. For example, none of the 3 privacy policies for *FirstAlert* [31] discuss the collection or sharing of device data. *FirstAlert* is a leading brand of smart safety devices, including their OneLink smart smoke and carbon monoxide alarm *contains a microphone* (hence collects audio) and integrates with both Apple HomeKit and Amazon Alexa. Similarly, we also found that *Panasonic* [69], which sells several smart home devices, including cameras, baby monitors, and smart air conditioners, also released 3 different privacy policies, none of which apply to devices. Several of these 225 vendors are high-impact, i.e., 43/225 are labeled as *Amazon's Choice* and 10/225 labeled as *Best Seller* (Figure 2), and together have 232,901 *Google Play Installs* and 700 *Amazon Ratings* on average (Figure 3). This finding indicates that *device privacy policies are not widely available*, preventing both their access by consumers, as well as analysis by researchers. As CCPA (in Section 999.308 [21]) and GDPR (in Article 13, 14 [34,35] mandate the disclosure of *all* categories of data collected to the consumers, these vendors are potentially in violation.

Finding 3: Not Found on Website – Only 64.38% of vendors that released a device privacy policy made it available from their website ($\mathcal{F}_3$) – In the absence of a centralized source of device privacy policies, users and researchers would be likely to look for them on the one common interface that most vendors provide: their website. However, as shown in Table 1, only 188/292 (64.38%) device privacy policies were obtained from vendor websites, whereas the remaining 35.62% (104/292) were found from other sources. While this shows that the majority of the vendors are making their device privacy policy more accessible, we find that users may have to jump through multiple hops to simply obtain device privacy policy. For instance, we encountered 41 instances, where we did not obtain the privacy policy in the website but found it after a Google Search. There could be several reasons for this, e.g., the privacy policy could be hidden in different subpage or shown only on the registration page. We also noticed 4 instances, where policies were obtained from a separate (i.e., parent company) domain (e.g., Allegion [4] hosts for Schlage [5]). Several of these 104 vendors are high-impact, as shown in Figures 2 and 3. This finding indicates that the current distribution of smart home privacy policies is severely fragmented and as a result, an automated approach that only obtains policies from websites would fail to obtain 35.62% of the policies.

Finding 4: Difficult to Obtain – Device privacy policies can be *difficult* to obtain ($\mathcal{F}_4$) – In 42/292 (14.38%) cases, we needed to *execute the vendor's companion app* to view

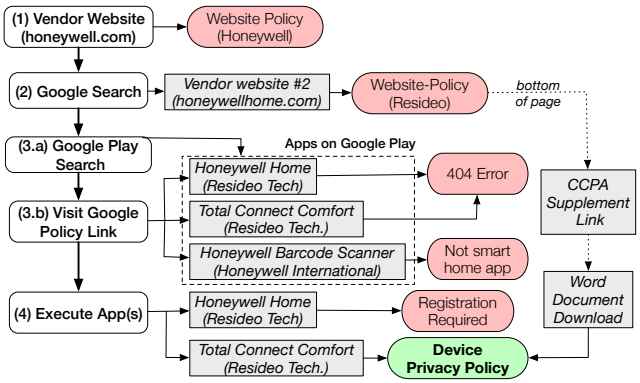


Figure 4: Obtaining Honeywell's device privacy policy.

the device privacy policy. These vendors are high impact, with 4,791,917 *Google Play Installs* and 2,805 *Amazon Ratings* on average (Figure 3), and with 9/42 labeled *Amazon's Choice* and 1 *Best Seller* (Figure 2). In many cases, it took significant labor to locate the device privacy policy due to several confounding factors, as exemplified by the convoluted path taken to obtain Honeywell's policy, as shown in Figure 4, and described below:

Honeywell's website and Google search results led us to two different *website* privacy policies that did not cover devices. Further, searching on Google Play led us to three top apps, of which the barcode scanner was eliminated as irrelevant. Interestingly, the other two apps are published by "Resideo Technologies", which is a third-party vendor that has *licensed* the *Honeywell* brand name, which may not be apparent to users. We were met with `HTTP 404` errors on attempting to access the privacy policy links accompanying the Google Play listings of both apps [48] [47].

Upon executing the apps, as our last resort, we discovered that the *Honeywell Home* app required registration to display the privacy policy. The fact that we were required to disclose PII (e.g., email, name for registration) before finding the device privacy policy is concerning from a privacy standpoint and against the general spirit of privacy policies. We finally obtained the *device privacy policy* for Honeywell's smart home devices upon executing the *Total Connect Comfort* app. When we revisited Resideo's website privacy policy [72], at the very bottom of the page, we found a link to the CCPA "Supplementary Privacy Statement". Clicking on this link *downloaded a Microsoft Word document* that discussed sensor data.

We learn two lessons from this experience: (1) As Figure 4 illustrates, the distribution of device privacy policies is convoluted, involving complicating factors such as unknown vendor relationships, broken links, account registration requirements, and unwarranted document downloads (i.e., users may not expect or want to download a Word doc, simply to view a policy), and, (2) The convoluted distribution, diversity of artifacts involved, and the complex combination of analyses required may pose challenges for automated privacy policy extraction. Finally, in contrast to this example, we found that several popular vendors do make their device privacy policies

easily available on their websites (e.g., Scout Alarm [80]).

Finding 5: No Policy on Website – 26.84% of the vendors do not make their *website privacy policies* easily available ($\mathcal{F}_5$) – We observed that 26.84% (160/596) of vendors did not post privacy policy links on their homepage, marked with "Privacy" or other similar phrases, which may violate certain regulations (e.g., CCPA [21]). As shown in Figures 2 and 3, several of these vendors have high impact.

5 Policy Content Analysis (RQ_2)

The goal of our *content* analysis is to understand both (1) the *semantics* of the content, i.e., *what* information is being disclosed and (2) the *structural composition*, i.e., *how* it is being presented in the policy. In this section, we identify properties of smart home (device) privacy policies that are vital for understanding these aspects, codify them into labels, and use the labels to annotate the content of the policies.

5.1 Methodology

We perform a systematic analysis of the 284/292 device privacy policies identified in Section 4, excluding 8 policies out of 42 cases where we executed the mobile apps to manually obtain privacy policies (as discussed in $\mathcal{F}_4$), as the apps prevented us from retrieving the policy text (e.g., by copying or taking screenshots), and because the policies were unavailable in the static content/apk resources. We first identify properties vital for understanding the semantics and structural composition of the policies, and codify them as labels, and then, label the 284 smart home device privacy policies, and analyze the results to identify semantic and syntactic intricacies.

Label Definition: We define a set of labels with the goal of sufficiently identifying domain-specific privacy contexts in smart home device privacy policies. For this, we initially explored the use of labels created for other domains (e.g., mobile apps [7, 98]), however, they did not disclose physical privacy context specific to smart homes (e.g., indoor privacy is perceived differently to outdoors). Privacy regulations also do not yield labels with sufficient granularity to accommodate our goal to study physical aspects. Thus, we founded the labels based on our knowledge of regulations (e.g., CCPA, GDPR) and prior work on privacy policies, while complementing them with domain-specific enhancements necessary to disclose gaps in device privacy policies, leaning towards precision in the spirit of good disclosure.

To elaborate, we define two unique granularities of labels: (1) *document labels* that apply as a property of the entire policy; and (2) *content labels* that apply to specific text fragments. To instantiate these labels, we began by analyzing the CCPA [56] and GDPR [70] regulatory documents to identify any requirements that may impact privacy policies. For example, both CCPA (in Section 999.308 [21]) and GDPR (in Article 13, 14 [34, 35]) mention that businesses should disclose a list of categories of personal information collected and shared, categories of third-parties to whom the data is shared, and the purpose of collection/sharing to the consumers. Based on this requirement, we define a content label for annotating text that refers to data collection practices (i.e., *collection*, *sharing*, and *collect_purpose/share_purpose*). Moreover, this requirement also inspired document_labels that denote the specificity of the device data types mentioned in the data collection practices (*collection_granularity* and *share_source_granularity*), which speak to the precision at which the privacy policy discusses the categories of personal information. For example, a policy that collects "usage data from sensors" will have *collection_granularity* set to *broad* while another that collects "audio data" will have *collection_granularity* set to *attribute*, to allow characterization of vendors that sell devices with multiple sensors, which may pose separate privacy costs/risks. Similarly, we include the document label *share_source_granularity* that describes the precision at which shared data is discussed and the label *share_destination_granularity* that captures with whom the personal data is shared. Lastly, we include document labels (such as contains_children_privacy, contains_data_retention, and contains_storage/transfer) motivated by their separate discussions in the regulations. For example, Recital 38 [36] along with various other articles in GDPR discusses special protection for children, whereas Section 1798.120(c) [56] in CCPA discusses disclosure practices regarding minors.

As CCPA and GDPR documents generally outline only *what* information needs to be disclosed, but not *how* it must be disclosed, we expand our analysis via an initial exploratory pass over a subset of policies (consisting of 18 policies). For example, we identified the two primary document formats (a) *monolithic*, which present all information in a single free-flowing body of text, and (b) *sectioned*, which break apart components of the policy content into disparate sections that cover different topics (e.g., sharing, collection). Hence, we define a label to capture this property (i.e., *contains_section*). Similarly, we observed that some policies display their collection/sharing practices in a table rather than plaintext (hence the label *contains_table*).

Our approach results in 24 labels that are motivated by our goal of evaluating disclosure practices in the specific context of the smart home, in the spirit of good disclosure (see Table 3 in Appendix B for the full list). These labels enable a contextualized analysis of privacy policies in the smart home domain, hence motivating future improvement and standardization of privacy disclosures for the smart home.

Labeling Privacy Policies: For this task, we take a multi-pass approach to improve the ease of the annotation task while reducing risks of errors. First, we begin by segmenting the policies (when possible) into up to 11 sections (i.e. preamble, collection, sharing, purpose/use, data retention, storage/transfer, children, contact, cookies, and extra). The goal behind this step is to ease the process behind our deeper annotation session and reduce sources of imprecision. Second, we anno-

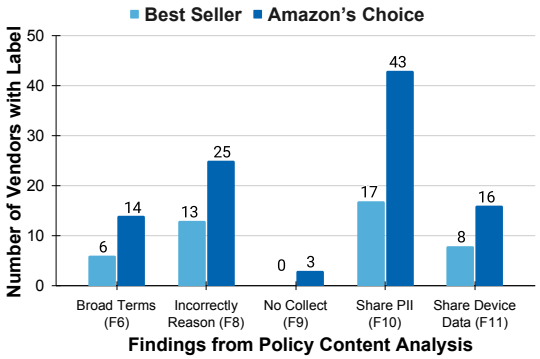


Figure 5: *Amazon's Choice* and *Best Seller* metrics for Content Analysis findings.

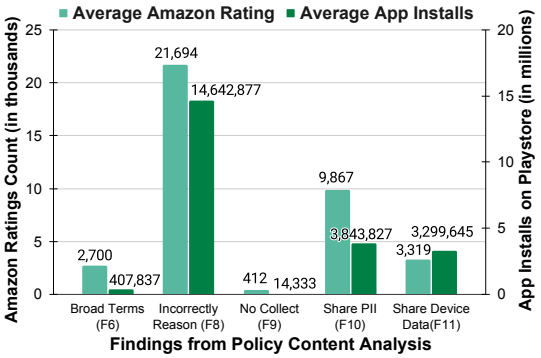


Figure 6: *Amazon Ratings* and *App Install*s for Content findings.

tate each document with document label annotations, such as whether it contains a table, summary, or sections. Third, we perform a deep and systematic analysis of the content by reading through the privacy policies using the section/segment guides. To ensure a complete analysis, we read through the entire policy, including seemingly irrelevant sections.

Two authors independently labeled the 284 privacy policies. We obtained an average Cohen's Kappa score of 0.95, denoting high inter-coder agreement over all labels. The Kappa score for each label are reported separately in Table 3 in Appendix B to highlight the relative disagreements for each label for increased reproducibility [90]. The two coders resolved disagreements through discussion.

5.2 Content Analysis Findings ($\mathcal{F}_6$–$\mathcal{F}_{12}$)

Based on our analysis of the structure, we found that 276/284 (i.e., 97.18%) privacy policies are split into different sections describing collection, purpose of collection, and sharing practices. 279/284 (i.e., 98.23%) contained mixed content, discussing website or mobile apps along with smart home products (e.g., Arlo [9]). Further, 275/284 (i.e., 96.83%) provide a preamble that briefly explains what the policy covers. We also found that 20/284 (i.e., 7.0%) include a table describing the privacy practices for different categories of data.

Our content analysis led to 7 findings, with considerable impact (illustrated in Figures 5 and 6). Note that two findings ($\mathcal{F}_7$ and $\mathcal{F}_{12}$) are *positive*, i.e., where the vendors followed best practices, and hence are excluded from the impact analysis.

Finding 6: Broad Terms – 26.05% **of the policies describe collection using *broad* terms rather than discussing specific device types or device data** ($\mathcal{F}_6$) – We found that 91/284 (32.04%) of vendors use broad terminology (e.g., "service usage information," or "sensor information") to refer to smart home device data when describing their collection practices. As we discuss later ($\mathcal{F}_{16}$), further analysis revealed that 17 were actually described precisely given the fact that the associated vendor only sold a single device, reducing the number to 74/284 (26.05%) generic policies. Considering that smart home vendors advertise 3 different device types on average as found later in Section 6.2, this broad terminology may not sufficiently describe the privacy practices of a vendor in the context of the individual devices they sell. For instance, Electrolux [27] sells smart home appliances, such as refrigerators and laundry machines, and their privacy policy explicitly mentions that the policy applies to the use of their "Smart Appliance." However, Electrolux's policy contains the following coarse statement that covers all devices but does not provide any useful information: *"...Electrolux will collect information about your use of the Smart Appliance and the App."*. Such vague descriptions of device data collection may not be in compliance of CCPA [21], given that CCPA requires the disclosure of all applicable categories of data (CCPA Section 1798.110 [18]), and sensor data may fall under several different categories currently not represented in the aforementioned privacy policies, such as biometric information and audio/visual information (CCPA Section 1798.140, subdivision (o) [19]). Further, as regulations evolve to recognize precise types of smart home data, such vague policies will fall further out of compliance. In contrast, several vendors make device-specific data and its collection explicit in their policies; e.g., Wyze [91] and NetAtmo [63] both discuss individual devices (e.g., indoor cameras, doorbells, and security alarms).

Finding 7: Device Data – 70.42% **of device privacy policies specify collection at the granularity of device data** ($\mathcal{F}_7$) – We found that 183/284 (64.43%) policies mention some device data when describing collection, which increased to 70.42% (200/284) after including the 17 that were precise given that the vendor only sold one device (see $\mathcal{F}_6$ and $\mathcal{F}_{16}$). For example, Wyze [92] thoroughly describes the data collected by its sensors as "device event data" (e.g., when it is turned ON/OFF) as well as "additional data" (e.g., heart rate by the weight scales), as shown in Figure 10 in Appendix A.

Finding 8: Incorrectly Reason – **Existing state-of-the-art privacy policy analysis tools may incorrectly reason about** (196/284) 69.01% **of smart home device privacy policies due to structural and semantic challenges** ($\mathcal{F}_8$) – 58/284 policies present content in one or more ways that existing analysis tools [7, 8, 43, 81, 99] cannot reason about due to their lack of consideration of the context of surrounding statements or their reliance on unstructured natural lan-

guage text. Particularly, 25/58 have locale-specific sections such for California residents and EU Citizens (e.g., Fitbit), 25/58 provide an external link to regulation-specific notices (e.g., Samsung SmartThings), and 20/58 display their collection/sharing practices using tables. Further, we found that tools like PolicyLint [7] cannot reason about *device data attributes* when policies describe data at attribute-granularity. For example, PolicyLint only identifies two policy statements that used device attributes (i.e., containing "device event information" and "video"), which is a negligible subset of the total number of such statements in our labeled dataset, as it contains over 183 policies that describe attributes. We study PolicyLint's effectiveness at recognizing smart home data types without insight into our smart home dataset through a case study of Named Entity Recognition (NER) in Section 7.

Finding 9: No Collect – 8 **vendors explicitly state that they *do not collect any information* in their privacy policy** ($\mathcal{F}_9$) – We found that 8 vendors that sell various devices or provide smart home services explicitly state that they do not collect data from users within their privacy policy, e.g., Nuheat [64], which sells thermostat and floor heating systems. However, NuHeat has a mobile application on Google Play and also advertises integration with Alexa, IFTTT, and Google Assistant, which may indicates data transmission to support remote access or third-party integration via REST API calls. It is certainly also possible that the vendor does not store any information, but simply forwards the request to integrated platforms (e.g., IFTTT), and hence may not consider this practice as data "collection". However, this understanding may not be consistent with the broad formal definition of "collection" adopted by regulations such as the CCPA (subdivision (e), Section 1798.140 [19]) that may consider such forwarding as within scope. Finally, this finding motivates the need to precisely understand what constitutes data collection, and for researchers to investigate the privacy practices of vendors that deny collection but enable network-based services.

Finding 10: Share PII – 186/284 **or** 65.49% **of device privacy policies only discuss sharing practices broadly for "PII" or "personal data", without explicitly including or excluding device data** ($\mathcal{F}_{10}$) – We found that vendors do not precisely discuss what they share, and often only disclose that they share 'PIIs' without explicitly mentioning whether that includes any device data or other precise categories of personal data (e.g., geolocation). Several highly popular vendors display such characteristics, as seen in Figures 5 and 6. If these vendors share device data even for a "business purpose" (subdivision (a), Section 1798.115 of CCPA [17]) as opposed to selling, but simply do not describe it categorically, then they may be in violation of the CCPA requirement that stipulates that vendors should include sharing information *for all categories* (e.g., biometric, geolocation, audio/visual) of personal information disclosed to the third-parties in the past 12 months [21], which implicitly includes data from devices (e.g., cameras, doorbells), given CCPA's broad definition of what constitutes as personal information (subdivision (o), Section 1798.140 [19]). In contrast, several popular vendors precisely describe the specific device data that they share, such as Ecobee [26], which describes how it shares data regarding electricity use to enable its partners and utility vendors to make intelligent decisions regarding electricity production and conservation. Finally, we note that certain vendors may provide sharing information regarding device data in a disparate set of non legally binding resources that may not be immediately evident or accessible, as we describe using Ring's example in Appendix D.

Finding 11: Share Device Data – 24.64% **of device privacy policies discuss sharing device data with varying degrees of precision, but generally do not specify with whom the data is shared** ($\mathcal{F}_{11}$) – We found that 70/284 (24.64%) vendor policies explicitly provide information pertaining to sharing of device data, which indicates that smart home device data is indeed being shared by a significant minority of vendors. Of these, however, only 35/70 (50%) discuss device data at the precise attribute level, while the remaining discuss all device data together (i.e., under the term "usage"). Moreover, only 24/70, i.e., 34.28%, explicitly discuss the *destination* of data, i.e., name at least one third-party partner they share data with. Most of these 70 vendors are highly popular, as shown in Figures 5 and Figure 6. In contrast, emerging vendors such as Foobot [32] provide lists of vendors and platforms (e.g., Alexa, IFTTT, Google Home) that they share device data with.

Finding 12: No Share – Only 2.1% **of vendors do not discuss sharing data and only** 3.87% **state that they do not share data** ($\mathcal{F}_{12}$) – The 6/284 (2.11%) policies that do not discuss sharing at all may either be due to the vendors actually not sharing any data or a poorly written policy. These vendors sold a variety of privacy-sensitive smart home devices, such as security cameras, video doorbells, and door locks. Of the 11/284 (3.87%) vendors that explicitly state that they do not share, 9 stated they do not share with third-parties, whereas the remaining 2 stated that they did not share data for marketing and promotional purposes, which is a positive deviation from the general trend observed in our analysis.

6 Policy Coverage Analysis ($\mathbf{RQ}_3$)

To comprehensively address privacy concerns, policies must not only discuss the collection of device data at attribute granularity, but also explicitly describe collection practices for *all* devices sold by the vendor. We propose a methodology to empirically evaluate such *coverage* ($\mathbf{RQ}_3$) in the 284 device privacy policies obtained in Section 5.

6.1 Methodology

The ideal approach to determine coverage would be to derive the *ground truth* about data transmission from the network traces for each device sold by a vendor, and contrasting this

ground truth with the claims in the policy. However, deriving such ground truth experimentally would be prohibitive in terms of device costs, and the manual effort of interacting with 1000s of smart home devices represented by our 284 vendor policies. Indeed, prior work has demonstrated that while one can scale up the network analysis of IoT devices to obtain coarse information (e.g., the destination servers) [49, 71], uncovering the extent of actual data transmission is hard to scale, even for a handful of devices [62].

Therefore, for a scalable analysis, we propose a data-driven approach that *approximates coverage* by using out-of-band information about the devices sold by vendors. That is, we define a coarse but effective test: A vendor's policy is *complete* if it describes the data related to *all the devices that they sell*. This test allows us to identify gaps in the policy, in terms of both (1) missing devices, and (2) missing device attributes. To enable this test, we first construct a *vendor-device_type map*.

Constructing the Vendor-Device_Type Map: We construct the *vendor-device_type map* in two steps: First, for each of the 284 vendors with device privacy policies, we manually obtain a list of *devices-types* (e.g., camera, doorbell) that they sell/advertise, by manually analyzing their websites (e.g., from sections labeled "Products" or "Devices"). Second, we map each device-type to a specific set of *minimal device attributes* associated with it (e.g., "camera recording (audio/video)" for the security camera), using device-attribute mappings provided under the Open Connectivity Foundation (OCF) standard [66] (used by Iotivity [51]) and by prior work [58].

Evaluating Privacy Policy Coverage: For each policy, we obtain every device-type sold by that vendor from the *vendor-device_type map*, and for each device, check for the presence of the minimal attributes in the privacy policy (e.g., lock status for the door lock). If the policy includes all the minimal device attributes necessary (i.e., *even if imprecisely or indirectly*), we consider it as complete. For example, we marked as complete several policies that only specify the collection of "audio/video" data in general, without specifying the device associated with it (e.g., indoor camera, doorbell, baby monitor). This permissive check allows us to obtain a *conservative estimate* of policies that have incomplete coverage.

6.2 Coverage Analysis Findings ($\mathcal{F}_{13}$–$\mathcal{F}_{16}$)

We identified a total of 130 device types when constructing the vendor-device map for our list of 284 vendors. Each vendor was associated with 3 device types on average with 110 vendors selling only one type. We performed an in-depth coverage analysis of 200/284 policies, i.e., barring 8 policies that explicitly stated that they do not collect any data ($\mathcal{F}_9$), 2 that did not discuss data collection, and 74 that were classified as using *broad* terms to describe data collection ($\mathcal{F}_6$). Our analysis of policy coverage led to 4 findings with considerable impact (Figures 7 and 8). Note that $\mathcal{F}_{14}$ is exempted from the impact analysis as it is not statistical.

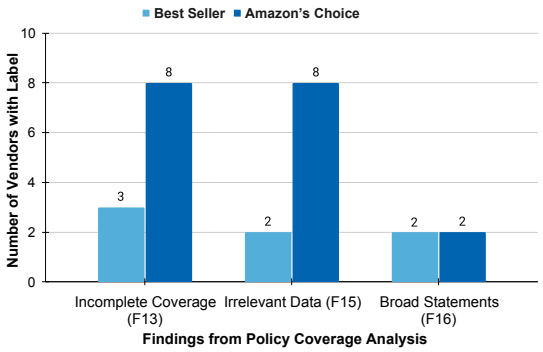


Figure 7: *Amazon's Choice* and *Best Seller* labels for Coverage Analysis findings.

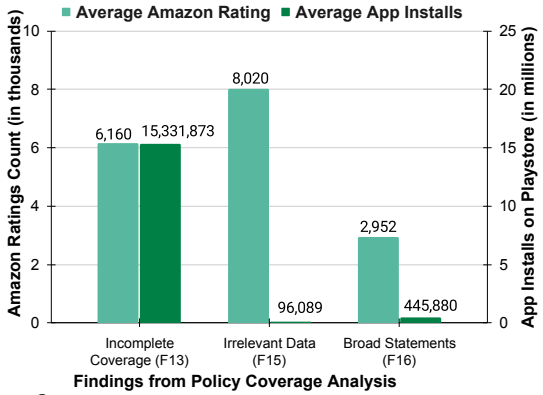


Figure 8: *Amazon Ratings* and *App Install* counts for Coverage Analysis findings.

Finding 13: Incomplete Coverage – 50/200 (25%) **of the privacy policies that precisely discuss device data are incomplete, i.e., only discuss a subset of their available devices** ($\mathcal{F}_{13}$) – We found 50 instances of incomplete policies when we semi-automatically compared the vendor-device_type map against the devices advertised by the vendors. Figure 9 shows the top 10 (out of 46) device data types that were missing from these vendor policies, of which several produce privacy-sensitive device data (e.g., cameras, motion sensors). For example, Owlet's [68] privacy policy was classified as discussing data collection at attribute granularity in Section 5, as they state that they collect *heart rate* information through their smart sock. However, Owlet also sells baby monitors (as indicated in our *vendor-device_type map*), but their policy does not provide information regarding the collection of attributes associated with this type (e.g., audio/video). Owlet's policy is representative of most policies with incomplete coverage, where vendors completely miss out on discussion of certain devices they sell. In contrast, we found instances of policies that exhaustively discussed each products they sell (e.g., Netatmo [63] describes all of its device types including camera).

Finding 14: Privacy Implication – Vendors do not differentiate their privacy disclosures for devices that produce similar data but have vastly different privacy implications ($\mathcal{F}_{14}$) – We analyzed 26 vendors that sold devices

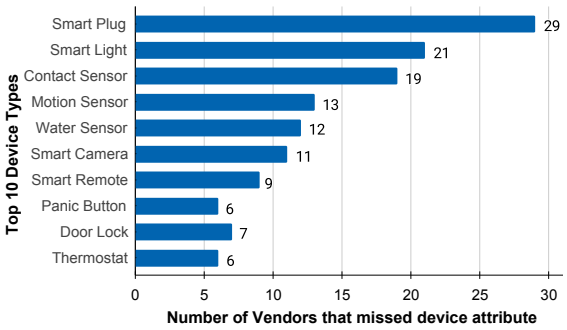


Figure 9: Top 10 device types omitted in privacy policies.

with different privacy implication (e.g., baby monitor and video doorbell) and found that 23 policies provided a generic description for the collection of *audio/video* data from devices. For example, Arlo [9] sells various devices that collect video data (e.g., video doorbells and baby monitors), but provides a common description, i.e., merely states that Arlo collects and *stores* video data. However, video from *baby monitors* may be far more privacy-sensitive than video from the *video doorbell* from the user's perspective, given that the former is recorded within the private confines of the home. Hence, users may not want Arlo to store video from a baby monitor even if they agree with long-term cloud storage for the video doorbell. Treating all video data similarly may make the privacy sensitivity of certain devices ambiguous to users.

Finding 15: Irrelevant Data – 29 **vendors describe the collection of device data precisely, but discuss irrelevant data due to remnants from *templates*** ($\mathcal{F}_{15}$) – Several vendors provide data collection statements that contain *unrelated* data attributes. For example, GeoSmartPro [37] only sells smart fans, yet it discusses the collection of height, weight, body fat mass index, BMI from fitness trackers and images and videos from smart cameras. In our content analysis in Section 5, we originally marked this example as evidence of collection at attribute level. However, when cross-verifying the statement with our *vendor-device_type* map, we found that $8/29$ cases contained statements describing device data that could not be associated with the vendor and contained structurally similar components and similar text, and hence, were potentially remnants from unmodified templates. The use of blanket statements from templates has also been observed previously in mobile app policies [7]. However, remnants in the policies can also simply mean that vendors were in the process of updating their policies or products. $\mathcal{F}_{15}$ exemplifies the ambiguity with which data collection is described in smart home privacy policies, which would not only make it difficult for users to understand policies, but would also complicate automated analysis (e.g., leading to a false positive in terms of data collection precision).

Finding 16: Broad Statements – Broad statements used to describe device data collection may not always denote insufficient precision or incompleteness ($\mathcal{F}_{16}$) – We discovered 17 cases where the privacy policies were initially classified as "broad" due to how they described data collection (in $\mathcal{F}_6$), but were precise given the actual devices they represent. For example, Milight [61], which only states that "online status" is collected, was classified as broad. However, our vendor-device_type map revealed that Milight (and 16 others) represented only one device-type, and were hence precise with respect to it; e.g., Milight only sells light bulbs, for which the minimal attribute is indeed "online status". Discussing this minimal attribute makes Milight's policy precise, as well as complete in terms of coverage, since that is the only attribute that needs to be discussed. This finding highlights the need to integrate out-of-band contextual information (e.g., vendor-device mapping) to disambiguate statements and enable an accurate analysis.

7 The Need for (and Path To) Contextualizing Policy Analyses for the Smart Home

This study uncovers several challenges and characteristics of smart home privacy policies that are sufficiently different from mobile or Web policies, whether in terms of their acquisition, content, or coverage, which may make direct application of existing tools difficult. For instance, as discussed in $\mathcal{F}_8$, we confirm that PolicyLint [7] cannot reason about smart home device data. Similarly, we ran Polisis [43] on our dataset of 284 device privacy policies, and as expected, Polisis could not reason about device data, and device data attributes such as "motion" were classified in the "Others" category (see Figure 11 in Appendix A for Polisis's results on the SmartThings privacy policy). Thus, we argue that smart home device privacy policies are a separate *application domain* from an analysis perspective, and one would need to *contextualize* existing techniques to effectively analyze them.

In this section, we first test the validity of this argument with a case study, i.e., by evaluating the effects of smart home-specific contextualization on the task that directly impacts an analysis tool's ability to detect relevant data types: *Named Entity Recognition (NER)*. We then discuss how the data, insights, and observations generated through this study may be leveraged for enabling such contextualized policy analysis frameworks for the smart home in the future.

7.1 Case Study: PolicyLint

We select PolicyLint [7], a state-of-the-art open source privacy policy analyzer (also used in PoliCheck [8]), and test the performance of its NER model on smart home device privacy policies, under two contrasting conditions: (a) using PolicyLint's NER model as is, and (b) by *augmenting the NER model with the smart home context*, i.e., training it with annotated sentences from smart home device policies.

Methodology: We created a smart home-specific NER dataset by annotating 600 policy statements from our dataset of 284 smart home privacy policies, using the annotation methodology as laid out by PolicyLint for creating training and test sets (see Appendix C). We shuffled this annotated dataset

Table 2: NER Performance of the *spaCy*, *PolicyLint*, *PolicyLintHome* models on the SmartHomeTest dataset

Metrics	spaCy (baseline)	PolicyLint	PolicyLintHome
	Overall NER Performance		
Precision	25.25	57.49	76.60
Recall	5.70	69.17	76.25
F1-Score	9.31	62.79	76.43
	Recognition of Data Objects		
Precision	-	65.68	75.29
Recall	-	71.77	76.20
F1-Score	-	68.59	75.75
	Recognition of Entities		
Precision	38.33	62.77	79.39
Recall	14.46	71.06	82.38
F1-Score	21.00	66.66	80.86

and used 500 statements for training (the *SmartHomeTrain* dataset) and 100 for testing (the *SmartHomeTest* dataset).

PolicyLint extends the spaCy NER model [86] by training it on an annotated mobile privacy policy sentences. We obtained this extended model from the authors, which we simply term as *PolicyLint*. We also obtained the stock spaCy model (i.e., en_core_web_lg [86]) to use as the baseline. Finally, we created a new model contextualized to the smart home, *PolicyLintHome*, by training it on the original PolicyLint annotations as well as the *SmartHomeTrain* dataset. All 3 models were tested on the *SmartHomeTest* dataset.

Results: Table 2 shows the performance of the 3 NER models on the *SmartHomeTest* dataset, including the overall performance and that over data objects and entities separately. We observe that in all cases, the PolicyLint model fares better than the baseline, but not nearly as well as the contextualized PolicyLintHome model, leading to our next finding:

Finding 17: Contextualization – The smart home signifies another application domain for policy analysis, and contextualized models fare much better than general-purpose models, or models trained on other domains. ($\mathcal{F}_{17}$) – The additional smart home context allows the PolicyLintHome model to obtain 19.11% higher precision (i.e., 76.6 vs 57.49), 7.08% higher recall (76.25 vs 69.17), and 13.64 higher F-1 score (76.43 vs 62.79) in comparison to the PolicyLint model, which clearly demonstrates the benefit of contextualization, at least for the task of NER, which has direct bearing on whether the model recognizes device data. Our adapted model recognizes several entities (e.g., "Fitness Tracker" and "Smart Camera"), and data objects (e.g., "heating system status", "connectivity status") that were not detected by the PolicyLint model, or classified as "Other" by Polisis. This impact is purely because of the additional smart home context provided by the integration of the SmartHomeTraining dataset, and not any changes to the model itself. The effects of adaptation to the smart home domain are evident here, just as the effects of adaptation to the mobile domain were evident in PolicyLint [7] as it demonstrated similar performance (i.e., $75-80\%$ F-1 score) when tested with mobile app privacy policies for which it is adapted.

7.2 Enabling Policy Analysis Frameworks for the Smart Home

While augmenting PolicyLint's NER model allowed us to improve its performance for smart home policies, our analysis reveals key challenges for enabling an end-to-end automated analysis of smart home privacy policies. This section discusses three such challenges, and how the artifacts, data, and lessons from this study may be leveraged by future research.

1. *Acquiring* and *Identifying* Smart Home Device Policies: Our study demonstrates that device privacy policies are not easily available for analysis due to a fragmented delivery system ($\mathcal{F}_2$, $\mathcal{F}_3$). To this end, our methodology in Section 4.1 lays the groundwork required for developing new integrated techniques to locate device privacy policies. Moreover, analysis frameworks also need to be able to *identify* the relevant device privacy policies, from a heterogeneous collection consisting of Web and mobile policies. Our dataset of smart home device policies enables future classifiers that effectively differentiates between device and other privacy policies.

2. Accurately Analyzing Device Data: Future research may build/adapt frameworks that consider *contexts* in which device data may be used, even beyond our experimentation with PolicyLint, by using our labeled dataset of 284 smart home device privacy policies with $7,494$ labels (e.g., by analyzing the device data usage context in collection/sharing statements). Further, our labeled dataset may also be used to test/evaluate the performance of models built to analyze smart home privacy policies (e.g., as in our analysis in Section 7).

3. Incorporating out-of-band context: We found that content analysis of device privacy policies is not effective in a vacuum, i.e., without using additional *out-of-band information* such as vendor-device relationships collected from external sources (Section 6). This insight motivates the need to gracefully incorporate such information into automated analysis for enhancing accuracy (e.g., detecting missing ($\mathcal{F}_{13}$) or irrelevant ($\mathcal{F}_{15}$) devices) and precision (e.g., avoiding false positives resulting from single-device policies in $\mathcal{F}_{16}$).

8 Vendor Disclosure

We manually informed our findings to 506 vendors, out of which we contacted 471 vendors via email between November 4 - 11, 2021, and 26 through contact forms on December 29, 2021. Note that we informed 9 vendors explicitly mentioned in this paper on June, 2020. We excluded 57 cases, where we could not contact the vendors because of reasons such as site or contact address being inaccessible, or when submission of contact form required product info (e.g., Zooz [84]). Listing 1 in Appendix shows the template of an email.

In total, we received 113 responses as of Jan 2022, after excluding 25 emails that were not delivered. The majority of responses i.e., $71/113$ were automated replies that acknowl-

edged the receipt of the message, without any further follow-up messages. In 6 responses, vendors informed us that they were in the process of addressing the issues identified, such as creating or revising their privacy policies. Additionally, 9 vendors directed us to their updated privacy policies. For example, when we sought clarification related to sharing practices as discussed in $\mathcal{F}_{10}$ and $\mathcal{F}_{11}$, EufyLife [82] provided instructions to obtain updated privacy policy from the app, which now discusses categories of information shared in the CCPA section. Moreover, 15 vendors clarified their data handling practices via email or provided additional information elaborating on their privacy practices, such as privacy FAQs, and in some cases, expressed the intention to include the said information in their privacy policies. For example, when we sought clarification regarding their data sharing practices and disclosure ($\mathcal{F}_{10}$), Starlinghome [83] explained that none of their devices collect usage data from smart home devices, but they are open to including additional details in their privacy policies to clarify as such. On the contrary, 11 vendors simply pointed us to their privacy policies without any changes, which we confirmed did not address any of the issues discovered in our study. For instance, 5 vendors who were informed about the unavailability of privacy policies applying to their smart home products (i.e., the *device policy* described in our study) ($\mathcal{F}_2$) responded with either the website or mobile app privacy policies, which did not contain any information on their treatment of device data. Finally, of the 43 vendors that had no privacy policy ($\mathcal{F}_1$), none responded to our disclosure, indicating the need for meaningful enforcement.

9 Discussion

Smart home technology is in its its incipient stages, and hence, there is opportunity for improving it from the standpoint of enabling privacy guarantees for user data. Our study highlights significant inconsistencies in smart home privacy policies, further motivating key improvements in how we standardize and regulate privacy disclosure, as we discuss in this section.

1. Bolstering informed consent through standardized distribution of policies: Our accessibility analysis highlights the user burden in effectively evaluating the privacy repercussions of IoT devices ($\mathcal{F}_1$–$\mathcal{F}_4$), to the extent of having to download and execute multiple companion apps to simply obtain device privacy policies. Our findings demonstrate the need for explicit distribution guidelines for smart home vendors such that consumers can make an informed choice based on the privacy practices of smart home devices prior to device purchase.

2. Strengthening transparency by improving precision and completeness: Our findings motivate the need to improve precision and completeness at which device data is discussed in smart home privacy policies, encouraging better disclosure practices for this domain. We find that vendors may not describe device data at all ($\mathcal{F}_2$), or may provide an imprecise ($\mathcal{F}_6$, $\mathcal{F}_{10}$), or incomplete ($\mathcal{F}_{13}$) description. Considering the privacy-sensitive nature of smart home domain, this further prompts vendors to improve transparency regarding their data handling practices. Furthermore, applying a carpet policy for a data type (e.g., video feeds) may be insufficient, no matter how precise the policy is, given that a data type may apply to devices with disparate privacy implications ($\mathcal{F}_{14}$). Considering the privacy sensitive nature of the smart home domain, consumers might also benefit if regulations are supplemented to require description of physical/digital contexts within which a data type is expected to be used.

3. Facilitating automated policy analysis via tool-enabled standards and practices: Our study motivates the need for the effective standardization of privacy policies, so that automated tools can be developed to reduce cost and effort for both vendors and consumers. We recognize that numerous external factors may lead to imprecise privacy policies despite vendors' intention to be transparent about their privacy practices. For example, small businesses may have little incentive or human resources to continually update their privacy policies as per the regulatory requirements and changing company processes. Our findings describe a spectrum of vendors ranging from those that do not provide any policies ($\mathcal{F}_1$) to those that discuss privacy practices for each of their devices (e.g., Wyze [92]). While stringent regulation is helpful, effective governance is challenging given the fragmented and heterogeneous nature of IoT ecosystem. Hence, clear standardization of privacy policies (e.g., a machine-readable standard) that is amenable to automated tools may be an effective approach for enforcing best-practices, and enabling consumers as well as researchers to automatically discover gaps in privacy policies.

4. Privacy Policy Enforcement: This work analyzes smart home privacy policies, i.e., "what vendors claim", which is complementary (but orthogonal) to an analysis of the privacy practices of vendors, i.e., of "what vendors actually do". The former requires an analysis of disclosure practices, as this paper does, while the latter requires a thorough analysis of several avenues for exfiltration, including smart home apps, the network, and cloud presence. In the latter area, prior work has developed systems that detect private data leaks by analyzing IoT apps (e.g., IoTWatch [10], SAINT [22]), or network traffic (e.g., Ren et al. [71], as we later discuss in Section 11. Our work complements such analysis of vendor app/network behavior by providing additional context to it in the form of disclosure practices, i.e., prior work leverage the data and findings of this study to validate *legally binding* collection/sharing claims in privacy disclosures. This is a natural future direction for privacy analysis in the smart home, as has been done in other domains (e.g., PoliCheck [8] builds upon PolicyLint [7] to validate app behavior in conjunction with privacy policy analysis). Finally, behavioral analysis has its limitations, and may not be able to detect when determined vendors purposefully violate their claims made in privacy policies by exfiltrating content after it reaches the cloud [33].

10 Threats to Validity

The methodology and findings of this study must be examined while considering the following threats to validity:

1. Manual effort, human error: Our study is semi-automated and consists of manual components that are subject to human error. We have taken several steps to mitigate the threats arising from this aspect, particularly in the interest of scientific rigor and reproducibility, e.g., using two coders for labeling, confirming *every finding* manually, and describing the methodology in precise detail for reproducibility.

2. Privacy disclosures *after* purchase: Vendors may provide privacy policies *inside the box*, i.e., after the purchase. However, we believe that retroactive disclosure contrasts with the spirit of informed consent (since the consumer is already invested into the device), and hence, do not consider this aspect.

3. Google Play: For our availability analysis, we only explore Android apps as Google Play has the largest market share, and it is extremely unlikely for a vendor to have a device privacy policy provided in their iOS app but not the corresponding Android app. However, we note that some vendors without device privacy policies may fall within this category.

4. Impact metrics: We choose Amazon badges (Best Seller, Amazon's Choice) and ratings for estimating impact as they have been known to have had a tangible impact on user purchases, e.g., the Amazon's choice badges increased the sales conversion rate by 25% and Best Seller Badge boosted page views by 45% [3]. However, such metrics may be susceptible to seller manipulation and may not fully represent actual purchase characteristics. Therefore, while we use multiple impact metrics (i.e., Amazon badges, ratings and Google Play install counts) to approximate the popularity of a vendor, they should be considered along with the risk of vendor manipulation.

11 Related Work

This work lays the empirical foundation to bridge the domains of *privacy policies* and *smart home device privacy analysis*, and is particularly related to prior work in the two areas.

Privacy Policy Analysis: Prior work has studied privacy policies in terms of their readability and comprehension [15, 52, 60]. Other work has addressed the availability of privacy policies for mobile apps [14, 25, 42, 99], where policies are readily available in the users' app usage workflow. In contrast, we study the availability of smart home device privacy policies that do not have a clear route of distribution. Prior research has also analyzed the content of privacy policies to identify vagueness [8, 13], opt-out choices [65, 78], contradictions [7, 24, 94], purpose-centric statements [93], and compliance [14, 15]. Similarly, research has also analyzed the collection and sharing practices in privacy policies for mobile applications [7, 8, 89, 94, 99], and websites [11, 12, 16, 23, 43, 77, 95, 96]. In contrast, our work focuses on the disclosure of such practices within *device privacy policies*, exploring the intricacies introduced by the smart home domain.

Smart Home Privacy Analysis: This is the first, empirical, large-scale analysis of smart home *device* privacy policies, especially in terms of analyzing availability and coverage, and attempts to understand the state of *privacy disclosure* in the smart home. In doing so, we complement prior work that analyzes the *behavior* of smart home products. Particularly, prior work has analyzed network traffic to understand the exfiltration of sensitive data via devices [1, 49, 55, 71, 71]. For instance, Ren et. al. [71] study the behavior of 81 devices installed at US and UK based labs to understand destination traffic, device interactions, and sensitive data exposures, while Moghaddam et. al. [62] perform a similar in-lab analysis of TV streaming platforms. In contrast, Huang et. al. [49] study network traffic from end-user homes to understand device behavior and tracking risks associated with smart home devices. Similarly, Kumar et. al. [55] performed a large scale empirical analysis of network scans to study the current state of security of IoT devices across different geographical locations. In a similar vein, recent work [10, 22, 30] has analyzed the privacy exposure resulting from smart home apps, particularly IoT apps provided on platforms such as SmartThings. For instance, IoTWatch [10] collects privacy preferences during app installation and informs users about sensitive data leaks when the preferences do not match the app behavior. As discussed previously in Section 9, our work complements such behavioral analysis from the network and apps by providing additional context to it in the form of disclosure practices, allowing prior work to use the data and findings of this study to validate *legally binding* collection/sharing claims made by vendors, which is a promising future direction.

Finally, given the known limitations in the efficacy of privacy disclosures, prior work has proposed *privacy labels* [28, 53, 54] to make security and privacy information more consumable at the first point of contact (i.e., prior to purchase), with privacy policies available for additional details. Our work complements this approach by motivating holistic improvements in the distribution, content, and governance of IoT privacy policies, which will provide a legally-binding supplement to privacy labels for interested consumers.

12 Conclusion

Smart home privacy policies inform consumers how vendors use their smart home data. We described an empirical study of 596 device privacy policies affecting 2,442 devices, using a data-driven methodology that studies availability, content, and coverage of devices in policies. Our findings demonstrate how users' access to precise and clear privacy policies is hampered by the lack of clear standards of policy delivery, specification, and contextualization to the smart home. The labeled data and insights produced in this work lay the groundwork for future automated analysis of device privacy policies.

13 Acknowledgements

We would like to thank our shepherd, Selcuk Uluagac, and the anonymous reviewers, for their constructive feedback. The authors have been supported in part by the NSF-2132281 grant, the COVA CCI Research grant, and a COVA CCI Dissertation Fellowship. Any opinions, findings, and conclusions expressed herein are the authors' and do not reflect those of the sponsors.

References

[1] Abbas Acar, Hossein Fereidooni, Tigist Abera, Amit Kumar Sikder, Markus Miettinen, Hidayet Aksu, Mauro Conti, Ahmad-Reza Sadeghi, and Selcuk Uluagac. Peek-a-boo: I see your smart home activities, even encrypted! In *Proceedings of the 13th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, WiSec '20, page 207–218, New York, NY, USA, 2020. Association for Computing Machinery.

[2] Amazon. Amazon Alexa. https://developer.amazon.com/alexa, Accessed December 2018.

[3] Amazon Badges. Amazon Badges Impact. https://www.modernretail.co/platforms/they-dont-have-as-much-value-how-amazons-choice-lost-some-of-its-luster/, 2021. Accessed: Dec 30, 2021.

[4] Amazon Best Seller Badge FAQ. Dataset. https://www.amazon.com/gp/help/customer/display.html?nodeId=GGGMZK378RQPATDJ.

[5] Amazon Choice Badge FAQ. Dataset. https://www.amazon.com/b?ie=UTF8&node=21449952011.

[6] Amazon Ratings FAQ. Dataset. https://www.amazon.com/gp/help/customer/display.html?nodeId=GQUXAMY73JFRVJHE.

[7] Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. PolicyLint: Investigating Internal Privacy Policy Contradictions on Google Play. In *Proceedings of the USENIX Security Symposium*, 2019.

[8] Benjamin Andow, Samin Yaseer Mahmud, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Serge Egelman. Actions Speak Louder than Words: Entity-Sensitive Privacy Policy and Data Flow Analysis with PoliCheck. In *Proceedings of the USENIX Security Symposium*, 2020.

[9] Arlo. Arlo Camera. https://www.arlo.com/en-us/about/privacy-policy/, Accessed June 2020.

[10] Leonardo Babun, Z. Berkay Celik, Patrick Mcdaniel, and Arif Selcuk Uluagac. Real-time analysis of privacy-(un)aware iot applications. *Proceedings on Privacy Enhancing Technologies*, 2021:145 – 166, 2021.

[11] Jaspreet Bhatia and Travis D. Breaux. A Data Purpose Case Study of Privacy Policies. In *Proceedings of the IEEE International Requirements Engineering Conference (RE)*, 2017.

[12] Jaspreet Bhatia and Travis D. Breaux. Semantic Incompleteness in Privacy Policy Goals. In *Proceedings of the IEEE International Requirements Engineering Conference (RE)*, 2018.

[13] Jaspreet Bhatia, Travis D. Breaux, Joel R. Reidenberg, and Thomas B. Norton. A Theory of Vagueness and Privacy Risk Perception. In *Proceedings of the IEEE International Requirements Engineering Conference (RE)*, 2016.

[14] Jasmine Bowers, Bradley Reaves, Imani N. Sherman, Patrick Traynor, and Kevin Butler. Regulators, Mount Up! Analysis of Privacy Policies for Mobile Money Services. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, 2017.

[15] Jasmine Bowers, Imani N Sherman, Kevin Butler, and Patrick Traynor. Characterizing Security and Privacy Practices in Emerging Digital Credit Applications. In *Proceedings of the ACM Conference on Security and Privacy in Wireless and Mobile Networks (WiSec)*, 2019.

[16] Carolyn A. Brodie, Clare-Marie Karat, and John Karat. An Empirical Study of Natural Language Parsing of Privacy Policy Rules Using the SPARCLE Policy Workbench. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, 2006.

[17] CCPA. CCPA - Business Purpose. https://leginfo.legislature.ca.gov/faces/codes_displaySection.xhtml?lawCode=CIV&sectionNum=1798.115., 2021. Accessed: Dec 30, 2021.

[18] CCPA. CCPA - Collection. https://leginfo.legislature.ca.gov/faces/codes_displaySection.xhtml?lawCode=CIV&sectionNum=1798.110., 2021. Accessed: Dec 30, 2021.

[19] CCPA. CCPA - Definitions. https://leginfo.legislature.ca.gov/faces/codes_displaySection.xhtml?lawCode=CIV&sectionNum=1798.140., 2021. Accessed: Dec 30, 2021.

[20] CCPA - Notice at Collection. Notice at Collection of Personal Information. https://govt.westlaw.com/calregs/Document/IC21FA97F1DA54DE591D2EA851633AA64, 2021. Accessed: Dec 30, 2021.

[21] CCPA - Privacy Policy. Privacy Policy Notice for CCPA. https://govt.westlaw.com/calregs/Document/I450BEC60B23D475082E55E9EC0AE6885, 2021. Accessed: Dec 30, 2021.

[22] Z Berkay Celik, Leonardo Babun, Amit Kumar Sikder, Hidayet Aksu, Gang Tan, Patrick McDaniel, and A Selcuk Uluagac. Sensitive Information Tracking in Commodity IoT. In *Proceedings of the 27th USENIX Security Symposium (USENIX)*, August 2018.

[23] Elisa Costante, Jerry den Hartog, and Milan Petković. What Websites Know About You: Privacy Policy Analysis Using Information Extraction. In *Proceedings of the International Workshop on Data Privacy Management and Autonomous Spontaneous Security (DPM)*, 2013.

[24] Lorrie Faith Cranor, Pedro Giovanni Leon, and Blase Ur. A Large-Scale Evaluation of US Financial Institutions' Standardized Privacy Notices. *ACM Transactions on the Web*, 2016.

[25] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. We Value Your Privacy... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy. In *Proceedings of the ISOC Network and Distributed Systems Symposium (NDSS)*, 2018.

[26] Ecobee. Ecobee Privacy Policy. https://www.ecobee.com/en-us/privacy-policy/, 2021. Accessed: Jan 25, 2021.

[27] Electrolux. Electrolux Appliances. https://www.electroluxappliances.com/, Accessed June 2020.

[28] Pardis Emami-Naeini, Yuvraj Agarwal, Lorrie Faith Cranor, and Hibshi Hanan. Ask the Experts: What Should Be on an IoT Privacy and Security Label? In *Proceedings of the 41st IEEE Symposium on Security and Privacy*, 2020.

[29] Evernote Website. Evernote note taking app. https://evernote.com/, 2021. Accessed: Jan 25, 2021.

[30] Earlence Fernandes, Justin Paupore, Amir Rahmati, Daniel Simionato, Mauro Conti, and Atul Prakash. FlowFence: Practical Data Protection for Emerging IoT Application Frameworks. In *Proceedings of the 25th USENIX Security Symposium (USENIX)*, pages 531–548, August 2016.

[31] First Alert. Smart Home Smoke and Carbon Monoxide Alarms. https://www.firstalert.com/product-category/smart-home/, Accessed June 2020.

[32] Foobot. Foobot Privacy Policy. https://foobot.io/privacy/, 2021. Accessed: Jan 25, 2021.

[33] Forbes. Report Claims Ring Employees Had Unfettered Access To Security Camera Footage. https://www.forbes.com/sites/paullamkin/2019/01/11/report-claims-ring-employees-had-unfettered-access-to-security-camera-footage/?sh=74086d4a206e.

[34] GDPR - Article 13. Article 13 of GDPR. https://gdpr-info.eu/art-13-gdpr/, 2021. Accessed: Dec 30, 2021.

[35] GDPR - Article 14. Article 14 of GDPR. https://gdpr-info.eu/art-14-gdpr/, 2021. Accessed: Dec 30, 2021.

[36] GDPR - Recital 38. Recital 38 of GDPR. https://gdpr-info.eu/recitals/no-38/, 2021. Accessed: Dec 30, 2021.

[37] GeoSmartPro. GeoSmartPro Fans. https://www.geosmartpro.com/airgo, Accessed June 2020.

[38] Gidbo Website. Gidbo Website. https://gidbo.com, 2021. Accessed: Jan 25, 2021.

[39] Google. Google Play Developer Policy Center: Privacy, Security, and Deception. https://play.google.com/about/privacy-security-deception/.

[40] Google. Google Play. https://play.google.com/store, Accessed December 2020.

[41] Google. Local Home SDK. https://developers.google.com/actions/smarthome/local-home-sdk, Accessed June 2019.

[42] Catherine Han, Irwin Reyes, Álvaro Feal, Joel Reardon, Primal Wijesekera, Narseo Vallina-Rodriguez, Amit Elazari, Kenneth A. Bamberger, and Serge Egelman. The Price is (Not) Right: Comparing Privacy in Free and Paid Apps. In *Proceedings on Privacy Enhancing Technologies (PETS)*, 2020.

[43] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G. Shin, and Karl Aberer. Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. In *Proceedings of the USENIX Security Symposium*, 2018.

[44] Stacey Higginbotham. The smart home will change insurance, but it will take time. https://staceyoniot.com/the-smart-home-will-change-insurance-but-it-will-take-time/, 2020.

[45] Home Assistant. Home Assistant Website. https://www.home-assistant.io/, Accessed June 2020.

[46] HomeKit. HomeKit Website. https://www.apple.com/shop/accessories/all-accessories/homekit?page=1, Accessed June 2020.

[47] Honeywell. HoneyWell Home Privacy Policy . https://www.honeywellhome.com/eula, Accessed December 2020.

[48] Honeywell. HoneyWell Total Connect Privacy Policy. https://www.honeywellhome.com/terms-and-conditions, Accessed December 2020.

[49] Danny Yuxing Huang, Noah Apthorpe, Frank Li, Gunes Acar, and Nick Feamster. IoT Inspector: Crowdsourcing Labeled Network Traffic from Smart Home Devices at Scale. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)/Ubicomp*, 2020.

[50] IFTTT. IFTTT helps your apps and devices work together. https://ifttt.com/, Accessed June 2018.

[51] IoTivity. IoTivity Wiki: IoTivity Initialization and Setting. https://wiki.iotivity.org/initialize_setting, Accessed June 2019.

[52] Carlos Jensen and Colin Potts. Privacy Policies as Decision-Making Tools: An Evaluation of Online Privacy Notices. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI)*, 2004.

[53] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W Reeder. A "nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, pages 1–12, 2009.

[54] Patrick Gage Kelley, Lucian Cesca, Joanna Bresee, and Lorrie Faith Cranor. Standardizing privacy notices: an online study of the nutrition label approach. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*, pages 1573–1582, 2010.

[55] Deepak Kumar, Kelly Shen, Benton Case, Deepali Garg, Galina Alperovich, Dmitry Kuznetsov, Rajarshi Gupta, and Zakir Durumeric. All things considered: An analysis of IoT devices on home networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 1169–1185, Santa Clara, CA, August 2019. USENIX Association.

[56] California State Legislature. California Consumer Privacy Act of 2018 ("CCPA"), 2018.

[57] Nicole Lindsey. Consumers Still Concerned About IoT Security and Privacy Issues. https://www.cpomagazine.com/data-privacy/consumers-still-concerned-about-iot-security-and-privacy-issues/, Accessed June 2019.

[58] Sunil Manandhar, Kevin Moran, Kaushal Kafle, Ruhao Tang, Denys Poshyvanyk, and Adwait Nadkarni. Towards a Natural Perspective of Smart Homes for Practical Security and Safety Analyses. In *Proceedings of the 41st IEEE Symposium on Security and Privacy (Oakland)*, San Fransisco, CA, USA, May 2020.

[59] Mattel Website. Mattel Website. https://www.mattel.com/, 2021. Accessed: Jan 25, 2021.

[60] Aleecia M. McDonald and Lorrie Faith Cranor. The Cost of Reading Privacy Policies. *I/S Journal of Law and Policy for the Information Society (ISJLP)*, 4, 2008.

[61] Milight. Milight Lights. https://www.milight.com/, Accessed June 2020.

[62] Hooman Mohajeri Moghaddam, Gunes Acar, Ben Burgess, Arunesh Mathur, Danny Yuxing Huang, Nick Feamster, Edward W Felten, Prateek Mittal, and Arvind Narayanan. Watching You Watch: The Tracking Ecosystem of Over-the-Top TV Streaming Devices. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 131–147, 2019.

[63] Netatmo. Netatmo Privacy Policy. https://view.netatmo.com/us/legals/app?gsc=true&goto=privacy, 2021. Accessed: Jan 25, 2021.

[64] Nuheat. Nuheat Website. https://www.nuheat.com/, Accessed June 2020.

[65] Ehimare Okoyomon, Nikita Samarin, Primal Wijesekera, Amit Elazari Bar On, Narseo Vallina-Rodriguez, Irwin Reyes, Álvaro Feal, and Serge Egelman. On the Ridiculousness of Notice and Consent: Contradictions in App Privacy Policies. In *Workshop on Technology and Consumer Protection (ConPro)*, 2019.

[66] Open Connectivity Foundation. OFC Solving the IoT Standards Gap. https://openconnectivity.org/, Accessed June 2020.

[67] OpenHab. OpenHab Website. https://www.openhab.org/, Accessed June 2020.

[68] Owlet. Owlet Baby Monitor. https://owletcare.com/, Accessed June 2020.

[69] Panasonic. Panasonic Smart Air Conditioner. https://www.panasonic.com/my/consumer/home-appliances/air-conditioner-learn/features-explanation/control-your-ac-with-your-smartphone-smart-wifi-network-adapter.html, Accessed June 2020.

[70] European Parliament and Council of the European Union. General Data Protection Regulation (EU) 2016/679 ("GDPR"), 2016.

[71] Jingjing Ren, Daniel J Dubois, David Choffnes, Anna Maria Mandalari, Roman Kolcun, and Hamed Haddadi. Information exposure from consumer iot devices: A multidimensional, network-informed measurement approach. In *Proceedings of the Internet Measurement Conference*, pages 267–279, 2019.

[72] Resideo. PRIVACY STATEMENT AND COOKIES. https://www.resideo.com/us/en/corporate/legal/privacy/english/?_ga=2.123804809.22730246.1592422631-1649725473.1592422631, Accessed June 2020.

[73] Ring. Ring CCPA Notice. https://shop.ring.com/pages/ccpa-disclosures, Accessed June 2020.

[74] Ring. Ring Law Enforcement Guidelines. https://support.ring.com/hc/en-us/articles/360001318523-Law-Enforcement-Legal-Process-Guidelines, Accessed June 2020.

[75] Ring. Ring Privacy FAQ. https://shop.ring.com/pages/privacy, Accessed June 2020.

[76] Ring. Ring Video Camera Website. https://shop.ring.com, Accessed June 2020.

[77] Norman Sadeh, Alessandro Acquisti, Lorrie Faith Cranor Travis D. Breaux, Aleecia M. McDonald, Joel Reidenberg, N. Cameron Russell Noah A. Smith, Fei Liu, Shomir Wilson Florian Schaub, James T. Graves, Pedro Giovanni Leon, Rohan Ramanath, and Ashwini Rao. Towards Usable Privacy Policies: Semi-automatically Extracting Data Practices From Websites Privacy Policies. In *Symposium On Usable Privacy and Security*, 2014.

[78] Kanthashree Mysore Sathyendra, Shomir Wilson, Florian Schaub, Sebastian Zimmeck, and Norman Sadeh. Identifying the Provision of Choices in Privacy Policy Text. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

[79] Jason Sattler. Privacy Concerns Cooling IoT Adoption in the US and Europe. https://blog.f-secure.com/privacy-concerns-cooling-iot-adoption-us-europe/, Accessed June 2019.

[80] Scout Alarm. Scout Alarm Privacy Policy. https://www.scoutalarm.com/pages/legal/privacy-policy, 2021. Accessed: Jan 25, 2021.

[81] Rocky Slavin, Xiaoyin Wang, Mitra Bokaei Hosseini, James Hester, Ram Krishnan, Jaspreet Bhatia, Travis D. Breaux, and Jianwei Niu. Toward a Framework for Detecting Privacy Policy Violations in Android Application Code. In *Proceedings of the International Conference on Software Engineering (ICSE)*, 2016.

[82] Smart Home Vendor EufyLife. Eufylife Website. https://us.eufylife.com/, 2021. Accessed: Dec 31, 2021.

[83] Smart Home Vendor StarlingHome. Starling Home Website. https://www.starlinghome.io/, 2021. Accessed: Dec 31, 2021.

[84] Smart Home Vendor Zooz. Zooz Website. https://www.support.getzooz.com/new/, 2021. Accessed: Dec 31, 2021.

[85] Smartthings Developers. Documentation. http://developer.smartthings.com/, Accessed June 2018.

[86] Spacy Model. Spacy en_core_web_lg model. https://spacy.io/models/en#en_core_web_lg, Accessed June 2021.

[87] Study of Smart Home Privacy Policies - Dataset. Dataset. https://github.com/Secure-Platforms-Lab-W-M/smart-home-privacy-policies.

[88] H. Tankovska. Smart home adoption: impact of privacy implications in North America 2018. https://www.statista.com/statistics/1076973/north-america-smart-home-adoption-privacy-implications/, 2020.

[89] Xiaoyin Wang, Xue Qin, Mitra Bokaei Hosseini, Rocky Slavin, Travis D. Breaux, and Jianwei Niu. GUILeak: Tracing Privacy Policy Claims on User Input Data for Android Applications. In *Proceedings of the International Conference of Software Engineering (ICSE)*, 2018.

[90] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1340, Berlin, Germany, August 2016. Association for Computational Linguistics.

[91] Wyze. Wyze Privacy Policy. https://wyze.com/privacy-statement, Accessed June 2020.

[92] Wyze. Wyze Website. https://wyze.com/, Accessed June 2020.

[93] Lu Yang, Xingshu Chen, Yonggang Luo, Xiao Lan, and Li Chen. Purext: Automated extraction of the purpose-aware rule from the natural language privacy policy in iot. *Security and Communication Networks*, 2021.

[94] Le Yu, Xiapu Luo, Xule Liu, and Tao Zhang. Can We Trust the Privacy Policies of Android Apps? In *Proceedings of the IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, 2016.

[95] Razieh Nokhbeh Zaeem, Rachel L. German, and K. Suzanne Barber. PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining. *ACM Transactions on Internet Technology (TOIT)*, 2013.

[96] Sebastian Zimmeck and Steven M. Bellovin. Privee: An Architecture for Automatically Analyzing Web Privacy Policies. In *Proceedings of the USENIX Security Symposium*, 2014.

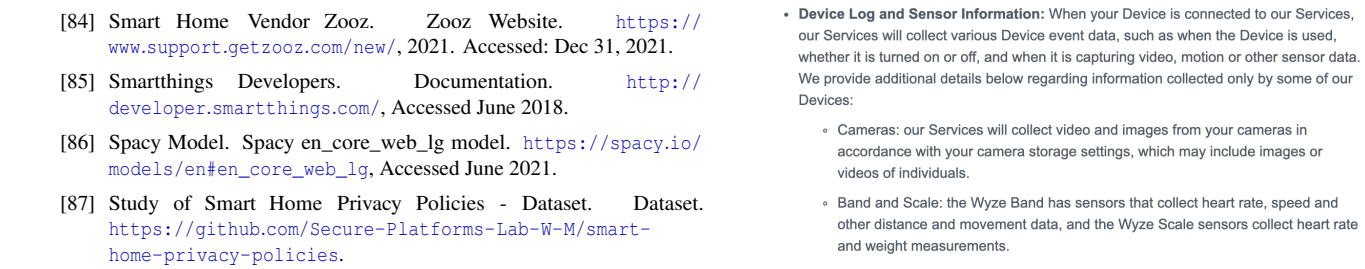
[97] Sebastian Zimmeck, Jie S. Li, Hyungtae Kim, Steven M. Bellovin, and Tony Jebara. A Privacy Analysis of Cross-device Tracking. In *Proceedings of the USENIX Security Symposium*, 2017.

[98] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel R Reidenberg, N Cameron Russell, and Norman Sadeh. Maps: Scaling privacy compliance analysis to a million apps. *Proc. Priv. Enhancing Tech.*, 2019:66, 2019.

[99] Sebastian Zimmeck, Ziqi Wang, Lieyong Zou, Roger Iyengar, Bin Liu, Florian Schaub, Shomir Wilson, Norman Sadeh, Steven M. Bellovin, and Joel Reidenberg. Automated Analysis of Privacy Requirements for Mobile Apps. In *Proceedings of the ISOC Network and Distributed System Security Symposium (NDSS)*, 2017.

A Privacy Policy Screenshots

Figure 10 shows the snippet Wyze's privacy policy that precisely discusses device data. Figure 11 is the snapshot of the response obtained from Polisis tool.



- **Device Log and Sensor Information:** When your Device is connected to our Services, our Services will collect various Device event data, such as when the Device is used, whether it is turned on or off, and when it is capturing video, motion or other sensor data. We provide additional details below regarding information collected only by some of our Devices:
  - Cameras: our Services will collect video and images from your cameras in accordance with your camera storage settings, which may include images or videos of individuals.
  - Band and Scale: the Wyze Band has sensors that collect heart rate, speed and other distance and movement data, and the Wyze Scale sensors collect heart rate and weight measurements.

Figure 10: Snippet from Wyze's privacy policy that describes how it collects data for its camera and weight scale devices

B Policy Labeling Appendix

Table 3 lists the labels defined and used for our content analysis. We note that unlike prior studies that study website or mobile app privacy policies, our study is aimed at holistic understanding of different aspects of smart home privacy policies. Thus, while our "Content Labels" are similar to the labels used by prior work [7, 8] (e.g., collection, sharing, purpose),

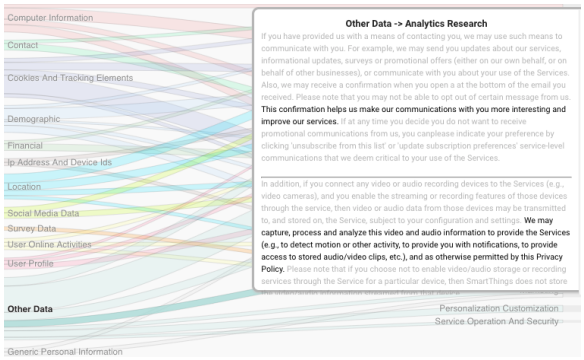


Figure 11: Result from Polisis tool for SmartThings privacy policy

the document labels (e.g., regulation_specific content, collection granularity) are unique to this work. For each label, the 'Label Origin' column describes how the labels were created, as described in Section 5 ("reg" refers to labels based on CCPA/GDPR requirements, "pol" refers to labels identified from our initial exploratory pass over a subset of policies, and "reg+pol" describes labels inspired from both).

C Methodology for Preparing NER Dataset

We created a smart home-specific NER dataset by annotating 600 policy statements from our dataset of 284 smart home privacy policies. To maintain fidelity to the original evaluation of PolicyLint, we used the same annotation methodology as laid out by PolicyLint for creating training and testing sets. That is, we extracted 400 statements using the 9 lexico-syntactic patterns indicated in the PolicyLint paper, and 200 random statements that discuss collection and sharing (using the labeled dataset resulting from Section 5). We annotated each statement with granular named-entity information, using the standard annotation methodology described in the PolicyLint paper. We shuffled this annotated dataset and used 500 statements for training (the *SmartHomeTrain* training dataset) and 100 for testing (the *SmartHomeTest* dataset).

D Sharing Statements in FAQs

Ring's [76] policy discusses sharing for "personal data" but is unclear about the sharing practices for some of the most privacy-sensitive device data collected (e.g., audio/video recordings). Ring's CCPA supplement [73] discusses audio/video data, but not in the context of the smart home. We found a FAQ [75] that discusses sharing device data (e.g., with law enforcement), which is further elaborated in a separate 'Ring Law Enforcement Guidelines" document [74].

E Email Template

We crafted our email based on the findings for each vendor. Listing 1 presents the generic outline of how we informed vendors about different findings reported in this paper.

Table 3: List of Content and Document labels and their corresponding Kappa scores calculated after labeling process

No.	Content Label	Paragraph Frequency	Kappa Score	Label Origin
1	collection	4323	0.82	reg+pol
2	sharing	1937	0.71	reg+pol
3	not_collection	133	0.74	reg+pol
4	not_sharing	118	0.66	reg
5	both (i.e., collection and sharing)	96	0.90	pol
6	not_both (i.e., does not collect or share)	9	0.94	pol
7	collect_purpose	2856	0.84	reg+pol
8	share_purpose	383	0.86	reg+pol
9	both_purpose	26	0.93	pol
No.	**Document Label**	**Document Frequency**	**Kappa Score**	**Label Origin**
10	(i) product policy_type	5	1	pol
	(i) mixed policy_type (site and product)	279	1	pol
11	effective_date	214	0.94	pol
12	(i) generic (not regulation specific)	223	0.99	pol
	(ii) CCPA regulation specific	46	0.97	pol
13	(iii) GDPR regulation specific	58	1	pol
14	contains_summary	14	1	pol
15	contains_sections	276	1	pol
16	contains_table	20	1	pol
17	contains_preamble	275	1	pol
18	contains_children_privacy	167	1	reg+pol
19	contains_data_retention	209	1	reg+pol
20	contains_storage/transfer	200	0.98	reg+pol
21	contains_contact	268	0.97	pol
22	(i) broad collection_granularity	91	0.97	reg+pol
	(ii) attribute collection_granularity	153	0.97	reg+pol
	(iii) not_collect collection_granularity	8	1	reg+pol
	(iv) undefined collection_granularity	2	1	reg+pol
23	(i) attribute share_source_granularity	35	1	reg+pol
	(ii) usage share_source_granularity	36	0.97	reg+pol
	(iii) not_share share_source_granularity	11	1	reg+pol
	(iv) undefined share_source_granularity	6	1	reg+pol
24	(i) specific share_destination_granularity	50	1	reg+pol
	(ii) only_purpose share_destination_granularity	191	1	reg+pol
	(iii) undefined share_destination_granularity	16	1	reg+pol

```
Subject: Issues Related to Smart Home Product Privacy Policy

To Whom It May Concern:

We are a team of security researchers from the <XYZ> in the Department of Computer Science at <XYZ>. We performed a systematic study to analyze privacy policies for different smart home vendors.

We found the following issues in the privacy policy <Privacy_Policy_Link>:
        1. <Finding Description>
        2. <Finding Description>

Any additional information that you may have that clarifies collection, purpose, and sharing practices of data originating from the device usage would be extremely helpful.

If you have recently updated your product privacy policy, kindly direct us towards the updated policy.

Thank you!
<EMAIL_SIGNATURE>
```

Listing 1: Email Template used to inform the vendors