Consumer Attitudes Towards Privacy and Security in Home Assistants

Nathaniel Fruchter
CSAIL, MIT
Cambridge, MA 02139
fruchter@mit.edu

Ilaria Liccardi
CSAIL, MIT
Cambridge, MA 02139
ilaria@csail.mit.edu

CHI'18 Extended Abstracts, April 21–26, 2018, Montreal, QC, Canada
ACM 978-1-4503-5621-3/18/04.
https://doi.org/10.1145/3170427.3188448

Abstract

Home assistants such as Amazon's Echo and Google's Home have become a common household item. In this paper we investigate if and what consumers have reported online (in the form of reviews) related to privacy and security after purchasing or using these devices.

We use natural language processing to first identify privacy and security related reviews, and then to investigate the topics consumers discuss within the reviews. We were interested in understanding consumers' major concerns.

Issues and/or concerns related to security and privacy have have been reported within reviews; however, these topics only account for 2% of the total reviews given for these devices. Three major concerns were highlighted in our findings: data collection and scope, "creepy" device behavior, and violations of personal privacy thresholds.

Author Keywords

home assistants; internet of things; privacy; security; consumer attitudes

ACM Classification Keywords

K.4.1 [Computers and Society]: Privacy

Device	Retailer	Reviews	% P&S
A-Echo	Amazon	3392	2.4
A-Echo	Best Buy	1174	1.5
A-Echo	Target	32	0
A-Dot	Amazon	65503	2.6
A-Dot	Best Buy	18437	1.8
A-Dot	Target	140	2.9
A-Show	Amazon	4202	7.4
A-Show	Best Buy	708	5.9
A-Show	Target	18	11
G-Home	Best Buy	8430	2.9
G-Home	Target	230	4.8
G-Home	Walmart	651	3.8
G-mini	Best Buy	6460	1.5
G-mini	Target	124	1.6
G-mini	Walmart	35	2.9
–	–	–	–
A-Echo	Overall	4598	2.1
A-Dot	Overall	84080	2.4
A-Show	Overall	4928	7.2
G-Home	Overall	9311	3.1
G-mini	Overall	6619	1.6

Table 1: Review text corpus by device, retailer, and percentage with Privacy & Security (P&S) discussion.

Introduction

The popularity of voice-activated personal assistants is growing to the extent that these devices are often consumers' first point of entry into the home Internet of Things (IoT) ecosystem. The adoption of such devices can only increase given the value that they create for users and developers [5, 10].

Home assistant devices often have a number of sensors or other features which can capture a great deal of data that was once thought to be private [6]; even the most intimate conversations may be recorded and perhaps sent to the cloud for storage and analysis. In addition to audio activity, smart home devices and their cloud-based services can capture other activities mediated through these systems, e.g. online purchasing behavior or news and entertainment listening preferences [11]. This data capture occurs passively (as a background process), often without individuals' knowledge or awareness due to the "always on" listening component that detects the device's activation phrase/keyword (*"OK Google", "Alexa"*).

However, the fact devices are constantly listening for these keywords does not imply that they are constantly transmitting "private" conversations captured around the device. They will typically transmit only sentences which are spoken after the keyword is detected, unless the keyword is spoken again. Once captured, however, this data could be used for other purposes beyond the reasonable expectation of consumers. It may even be sought by law enforcement agencies for investigatory purposes [9].

In this work we seek to determine if consumers write about privacy or security issues when they review, investigate, or discuss home assistants on popular e-commerce websites (Best Buy, Amazon, Walmart and Target). We are interested in understanding and identifying consumers' opinions, possible concerns, and general views about these devices. We are interested in understanding: **(1)** Do people report/discuss privacy and/or security topics when reviewing a product? **(2)** When privacy and/or security topics are present, what are people's major concerns?

There is clear and growing concern within the technical and policy communities regarding the privacy and security implications of IoT devices [7]. Past research has shown consumers to be unaware, uninterested, or uninformed when it comes to privacy and security issues [1, 12], whether due to misconceptions [15] or lack of expertise [13, 12]. While privacy and security have often been reported as mattering to consumers, they are not considered or are often forgotten when deciding which items to buy. In fact, consumers have been shown to choose devices and features which infringe upon their privacy and security rights in order to gain functionality [8].

Study Design and Methodology

In order to investigate the presence and the substance of P&S discourse, we analyzed consumer review data related to the five most popular home assistant devices on the market as of Fall 2017 [3].

Review Data

We obtained reviews from large U.S. online shopping websites which sold at least one of the devices (see Table 1) and verified purchasers[1]. For each device, we collected all review text. The metadata collected included the review ID, review type (verified purchaser), review date, and review score (star rating). A web scraping system based on Python's *requests* client and *BeautifulSoup* HTML parser

[1] We did not include Google's first-party store as the Google Store does not solicit consumer reviews for its devices; it only advertises reviews aggregated from other sites, such as the ones in our sample.

privacy	security
individual	fear
leak	protection
breach	violence
permission	physical
loss	threat
storage	terrorism
data	cyber
surveillance	hack
spy	government
monitor	police
violation	crime
violate	abuse
legal	ethic
law	freedom
secret	insurance
confidential	harm
private	damage
nsa	vulnerability
fbi	unauthorized
creepy	snowden
third party	cybersecurity
track	firewall
privacy policy	virus
license agreement	malware
terms of service	spyware
	antivirus

Table 2: List of 53 Privacy & Security (P&S) keywords used.

was used to locate each product's review set. This set of reviews was then filtered down to those tagged as *verified* so as to analyze reviews based on real purchases.

Data Processing and Analysis
We used a combination of natural language processing techniques and human-based methods to identify a subset of our review corpus that reported privacy and security issues.

Automated Language Analysis
We created a dictionary with 53 keywords to be used to identify P&S reviews. These keywords, found in Table 2, are a broad set of privacy and security related terms. Several iterations were used to create the appropriate set of keywords. The authors discussed and identified common terms often associated with privacy or security concerns, opinions and attitudes. A set of keywords generated from prominent privacy and security related events and media coverage were also included in the dictionary.

This keyword dictionary was designed to be over-broad. Given that each review thread that contained a keyword would be read by at least two researchers, the false positives that were erroneously tagged could later be discarded. We were more interested in identifying *possible* privacy and security-related discussions rather than discarding relevant ones in the tagging process. False negatives would impact the validity of our research.

The text of each captured item was processed through stemming and lemmatization functions to ensure that all derivative forms of the keywords were considered during the tagging process [4]. We then used part-of-speech tagging functions to identify the subset of reviews which had at least one privacy or security keyword in their text.

We also augmented the results of our keyword analysis using a topic modeling approach [2]. We created an LDA[2] model trained on our stemmed review dataset using the *gensim* and *nltk* Python packages. This model enabled us to check the accuracy of the keyword analysis and gain a sense of the predominant topics of discussion and sentiment in reviews for all devices.

Results

We gathered 109,536 reviews from four major online retailers: Target, Walmart, Amazon, and Best Buy. Our corpus includes reviews that were written between November 6, 2014 and January 1, 2018. Table 1 shows a breakdown of all reviews collected for each device from each retailer.

Presence of privacy and security related reviews
We searched for 53 keywords (Table 2 within the 109,536 captured reviews and found an upper-bound subset of 2,237 (2.04%) *P&S reviews* which mentioned at least one keyword. This proportion does not significantly vary by device (p>0.1) except for the Amazon Echo Show, which has roughly double the number of P&S Reviews (6.1% of its reviews; p<0.01) compared to its competitors (this disparity is discussed further in the next section).

Aside from the Echo Show, our topic model confirms this observation, finding a P&S-related topic vector in 5,908 (5.39%) of reviews. This is also an upper bound; given the sparsity of P&S topics in the corpus, the model was not able to differentiate between physical and digital security issues.

While we found that privacy and security related issues are present within our corpus, our results suggest these top-

[2]Latent Dirichlet Allocation (LDA) is a statistical technique that allows sets of observations–such as words in a document–to be explained as a weighted mixture of individual topics.

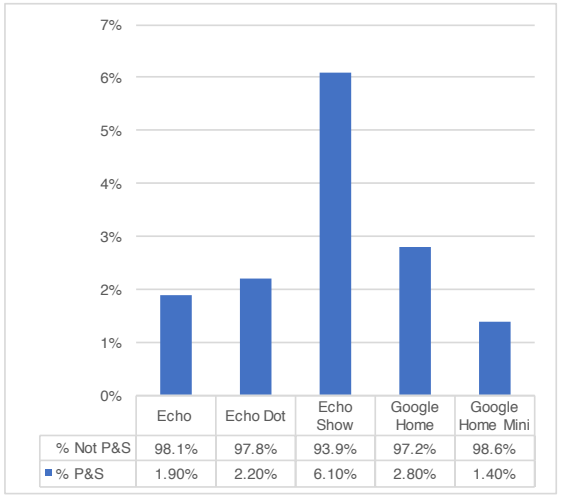

Figure 1: Distribution of privacy and security reviews compared to other reviews reported on each of the devices.

ics related to home assistants are rarely voiced, or *openly* reported by consumers in their online reviews[3]. Figure 1 shows the distribution of privacy and security topics as well as the remainder of unrelated reviews for each device.

Privacy and security concerns

We wanted to determine if there exist any trends or patterns in the limited set of reviews which do raise privacy or security concerns. Privacy is the predominant topic among this subset, occurring in just under 10% of P&S reviews. Specifically, reviews focus on the issues of **data collection** and individual **privacy thresholds** (most often couched in terms of "**creepiness**" by reviewers).

[3]Future work would include an analysis and/or empirical study to determine if consumers consider P&S issues but do not report them, if they are unaware of them, or simply just push them aside.

Concern: Data collection

"Amazon has not been forthcoming with what data this picks up and forwards to the cloud, especially when Alexa is not actively being used." [R269-A]

Reviewers reported concerns about the amount, scope, and type of data collected by their assistants. Reviews with data-related topics or keywords (n=214, 9.6%) tended to express uncertainty and frustration about data collection practices (*"I feel a bit insecure about it using The Cloud [...] it will get hacked one of these days"; "When you sync Alexa to your phone it extracts all of your personal contact information... customer service rep admitted that she could see all my contacts"* [R1213-T, R1647-A])[4].

A small number of reviews (P&S reviews which discussed *data* or *data collection*; n=31, 1.4%) worried about the lack of clarity about the scope of data collection. ("*If Amazon cares to provide any legal statement or policy that it does not record [...] then I would feel better [...] it creates a lot of user doubt and mistrust*"; "*Amazon should indicate exactly what Alexa "hears" and saves to Amazon servers*" [R1183-A, R933-A]). These frustrations often lead users, such as the previous reviewer, to abandon or return the assistant (*"Now I don't know whether to return it or not because now my personal data is on this device"*).

Concern: "Creepy" behavior

"Amazon just wants to use it to make more money - I wanted a smart alarm clock but purchased a SPY." [R850-A]

Many reviewers expressed concern about how their home assistants changed or violated their personal conceptions of privacy in the home environment; 134 (5.9%) mentioned the concept of "creepiness" in some form. A common worry

[4]Reviews are cited by noting the document number and are suffixed with a store code (A: Amazon, BB: Best Buy, T: Target, W: Walmart).

is that conversations will be overheard (*"an always listening device will prove to not a good thing. Do you ever have private conversations?*" or will be transmitted to outside parties (*"amazing spy for NSA"* [R1324-A, R950-BB]).

This often manifests itself in discussions of the home assistant as a "spy" (mentioned in n = 55, 2.4%; *"compact design but I do not like the fact that it might be spying on me"; "but other than big brother spying on me... its a great product"* [R2188-BB, R1627-BB]) or as a tool for intrusive marketing (*"I get random advertisements on my computer purely based off of conversations I have had on my home"; "still listening and likely being assimilated by Amazon to target future purchases"* [R1627-A, R619-A]).

Concern: Privacy thresholds

Other reviewers more explicitly discuss events which pass their personal privacy thresholds and this was especially common in reviews with more than one privacy-related keyword (n=894, 39.9%). These anecdotes deal with personal experiences with the device (or with worries spurred on by outside news reports in a few select instances). For example, one reviewer stopped using their device after hearing a recording of their son ("it was just the sound of my son talking... I unplugged it the next day and put it back" [R1101-A]). Another cited worries about a pending court case: *"law enforcement is trying to get Amazon to surrender data... if that happens it just opens the door to more and more privacy invasion*" [R343-A].

However, a substantial number of reviews indicate that a loss of privacy is an adequate tradeoff given their device's functionality. Among the 1584 (70.8%) P&S reviews an explicitly positive topic model, 99 (4.4%) still discuss creepiness and privacy thresholds. Some are on the fence, recognizing concerns but still endorsing the product: *"creepy privacy concerns about the big G listening all the time that I don't dismiss, but I'm still in for using one"*; *"If you accept the creepy always listening part, then we can recommend the permanent Google presence in your house"* [R2181-BB, R2221-W]. Finally, some even make the tradeoff more explicit (*"it works great but so creepy"; "creepy but amazingly priceless"; "Let em spy. I like the convenience more than my privacy"*[R645-A, R2159-BB, R1927-BB]).

Notably, we believe that increased discussion of privacy thresholds accounts for the disproportionate review count for the Echo Show. The Show is the only device in our sample to contain a video camera and screen. Manual review of the review sample indicates that over 100 (2.05%) reviews discuss the device's camera, video calling, and "drop in" conferencing features (*"[Drop In is] horrible from a privacy perspective"; "media folks have stated that the "Drop-in" feature of the Echos is "creepy" and can see their point [but] it is a helpful tool for the elderly"* [R1418-A; R1329-A]).

Discussion and Future Work

In this study, we sought to understand how consumers communicate P&S concerns with the popular home assistant class of IoT devices. By leveraging a corpus of consumer product reviews from four major online retailers, we were able to provide one of the first pictures of how consumers discuss P&S concerns of home assistant devices.

We can conclude that, for the most part, consumers who review home assistants tend to not discuss privacy or security concerns. Where consumers do discuss concerns, they do so in regards to the type and amount of data collected by the assistants. They also detail conceptions and potential violations of their privacy thresholds, often drawing a line between "creepy" and "non-creepy" assistant behavior.

Privacy thresholds have previously been discussed in other contexts where data is collected. Shklovski et al. analyze

privacy perceptions in the mobile app space and note users' perceived creepiness, yet simultaneous continued use of, mobile applications which conduct continual data collection [14]. However, this paradigm remains little-explored in the larger IoT realm. Deeper knowledge of users' attitudes can help designers and engineers improve these new forms of IoT-centric interaction. It can also help designers and policymakers with the challenge of obtaining meaningful consent for data usage and collection as the number of devices with non-traditional (or nonexistent) user interfaces increases in the home [11]. A more in-depth behavioral or *in-situ* study of user perceptions coupled with an analysis of a broader swath of the home IoT market could serve as some of the first steps towards this goal.

Acknowledgements

The authors would like to thank Brandon Karpf for his contribution to previous versions of this work. Nathaniel Fruchter and Ilaria Liccardi were supported by the William and Flora Hewlett Foundation grant 024127-00004.

REFERENCES

1. Alessandro Acquisti and Jens Grossklags. 2005. Privacy and rationality in individual decision making. *IEEE Security & Privacy* 2 (2005), 24–30.
2. Charu C Aggarwal and ChengXiang Zhai. 2012. *Mining text data*. Springer Science & Business Media.
3. Strategy Analytics. 2017. Smart Speakers: Sales Head towards 24 Million in 2017 Despite Confusing Array of Choice. (October 2017). `http://sa-link.cc/1lq`
4. Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.
5. Louis Columbus. 2016. Roundup Of Internet Of Things Forecasts And Market Estimates. (2016).
6. Federal Trade Commission and others. 2015. Internet of things: Privacy & security in a connected world. (2015).
7. Brian Krebs. 2016. KrebsOnSecurity Hit With Record DDoS. (2016).
8. Pedro G. Leon, Ashwini Rao, Florian Schaub, Abigail Marsh, and Lorrie F. Cranor. 2015. Why people are (Un) willing to share information with online advertisers. *Technical Report CMU-ISR-15-106, Carnegie Mellon University* (2015).
9. Christopher Mele. 2016. Bid for Access to Amazon Echo Audio in Murder Case Raises Privacy Concerns. *The New York Times* (Dec 2016).
10. Florian Michahelles and Stephan Karpischek. 2010. What can the Internet of Things do for the citizen (CIoT)? *IEEE Pervasive Computing* 10 (2010), 102–104.
11. Scott R. Peppet. 2015. Regulating the Internet of Things : First Steps. *Texas Law Review* 93, 85 (2015).
12. Lee Rainie and Maeve Duggan. 2016. *Americans' opinions on privacy and information sharing*. Technical Report. Pew Research Center.
13. Joel R Reidenberg. 2005. Disagreeable Privacy Policies: Mismatches Between Meaning and User's Understanding. *Erasmus* November (2005).
14. Irina Shklovski, Scott D. Mainwaring, Halla Hrund Skúladóttir, and Höskuldur Borgthorsson. 2014. Leakiness and Creepiness in App Space : Perceptions of Privacy and Mobile App Use. *Proc. CHI'14* (2014).
15. Rick Wash. 2010. Folk Models of Home Computer Security. In *Proc. SOUPS'10*. ACM, Article 11, 11:1–11:16 pages.