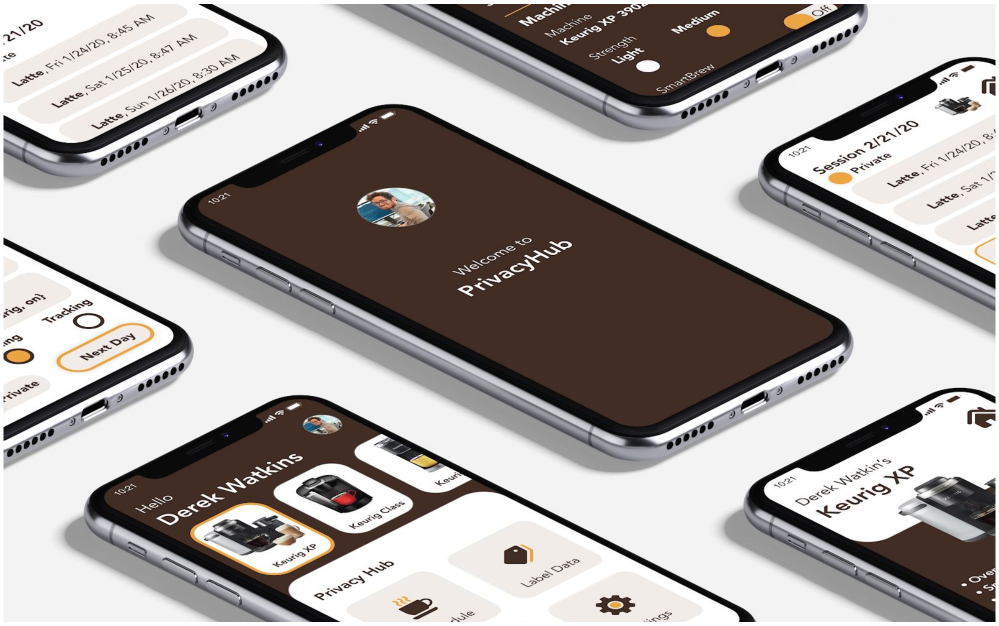
Privacy By Design:
Best Practices to Minimize Privacy Risks Posed by Smart Home Technologies

Sohini Kar
Ting Li
Maya Slavin

6.805 - Foundations of Internet Policy
December 9, 2020

Table of Contents



Executive Summary

Smart home technologies (SHTs) are a growing area of innovation in technology both domestically and abroad. SHTs include sensors, monitors, appliances, and devices in the home that are networked together to enable automation and remote control of everyday processes.[1] As users make the decision to place these gadgets in their homes, they can benefit from the convenience, functionality, and information that these devices provide. However, many users are unaware of the potential privacy harms posed by SHTs.

This paper is concerned with the ensuing harm from users' lack of control over what their SHTs record and share about them. Current corporate SHT data collection privacy policies are elusive; at best, they include inconvenient opt-out features and at worst, they completely disregard user control over personal information. The sheer quantity and invasive nature of the data collected by SHTs put users at risk of having the details of their personal lives used against them. This data can include sensitive information such as time-stamped logs of everyone who entered and left the house, records of movement between rooms, and audio recordings of everything that was said in a room, among countless other possibilities. When this information goes unregulated, private data is allowed to leave the home and may be used in unintended ways.

As personal information from SHTs leaves the home, consumers are subject to potential violations of privacy from two primary groups of actors: commercial producers and law enforcement. Commercial producers of SHTs can use the data collected to make inferences about the lifestyle, habits, and characteristics of their users. They can then make profits by selling the information to advertisers who will profile and target the user with specific ads. In the worst case scenario, this data could be sold to insurance companies who can integrate it into their models of assessing risk to price discriminate between individuals. In short, commercial producers benefit from the collection of excess data taken inside the home, in a way that the user may not expect and possibly be harmed by. The second group that can access private data from SHTs is law enforcement agencies. With a warrant, they may order companies to release any data relevant to an investigation. This is an extreme invasion of user privacy and overstep of power, as law enforcement is able to see essentially everything that happened in a home over a given period of time using data collected by SHTs. Legal tradition in the United States has long emphasized the sanctity of the home, but law enforcement having unfettered access to any type of data from SHTs jeopardizes the very principle underlying the Fourth Amendment.

To address these concerns, we propose a set of industry best practices for commercial producers of SHTs that gives users control over the type of data collected and shared by their SHTs so that they can decide what information leaves the home. We recognize that the appropriate

[1] Smart Home - United States: Statista Market Forecast.

implementation of such standards will vary between companies, products, and resources. This paper delineates three potential options that demonstrate a deliberate and effective effort towards enabling user control: (1) **generating individual privacy policies using machine learning**, (2) a mechanism that allows users to **opt-in by data type**, and (3) an **accessible privacy settings menu** with monthly data reports.

The first option recommends a machine learning algorithm that tailors privacy policies for each user based on data they label. This algorithm can be enabled following two specific standards: (1) SHT data must be available in a format easily understandable to users, and (2) each product must offer a trial period during which customers can label the data collected on them, and these labels must be used to determine what information to share with the company. Among the three options provided, the machine learning algorithm is by far the most comprehensive and most cognizant of individual perceptions of privacy, but it demands computational resources that small companies may not have.

Secondly, companies can implement opt-in by data type privacy policies in place of conventional opt-out options. For example, when setting up their device, users should be prompted to select privacy preferences for collecting and sharing images, videos, biometrics, voice recordings, location data, and time-stamped events. This option would maximize privacy by default, ensuring that otherwise oblivious users have protections that do not necessitate going out of their way to configure privacy features. While an opt-in policy does not allow for the same level of personalized control over SHT data as the machine learning algorithm above, it nevertheless meets the intent of the standard by allowing users to dictate what data is stored and shared.

Finally, the third option focuses on fostering privacy awareness in users by encouraging companies to distribute monthly data reports. These data reports should plainly show the type, quantity, and content of data collected by all devices in the home, in addition to information that can be inferred from a combination of data. The user should be able to navigate to an accessible privacy menu and either delete or refuse the collection of any of the above fields at any time. Such a system would motivate users to take a more active role in dictating privacy policies and ensure that they have both knowledge and control over what data is collected, stored, and shared.

Furthermore, we compare the three recommended policies and their limitations on both commercial producers of SHTs and their users, addressing system default settings in cases where consumers don't express their preference. Specifically, we recognize that requiring user choice at granular levels is unnecessary and allow for system defaults that are communicated transparently and heavily weigh user preferences when available. Ultimately, the time is ripe for companies to take more responsibility and actively promote user privacy. The quantity and sensitive nature of data collected by SHTs have the potential to inflict blatant violations of consumer privacy and undermine the sanctity of the home stressed by the Fourth Amendment. The adoption of industry standards of privacy-oriented design enabled by user control over the type of data that is stored and shared by SHTs will mitigate these harms.

1. Introduction and Background

6:30am. The alarm clock chimes earlier than usual. Derek Watkins, a recently married 35-year old software engineer, is normally awakened at 8am every morning in his home in Pacific Heights, an affluent suburb of San Francisco, by his Google Nest Hub. Today, his Nest smart clock scanned his schedule and adjusted the alarm to ensure enough time before a 7:30am presentation. He stops the alarm and strolls into his bathroom, where his Kohler system has preheated his shower with his preferred water temperature, flow rate, and duration. Afterwards, as he walks through the living room to prepare breakfast, Derek leaves a trail of motion-activated smart lights flickering off in the bedroom and on in the living room. Before he even sets foot in the kitchen, the aroma of freshly brewed Hazelnut coffee beans from his Keurig programmable coffee machine reaches him. After Derek grabs his coffee mug, the Keurig notices that its supply of Hazelnut beans is dangerously low. This information is sent back to the company, and within seconds, Derek gets two emails: an automated confirmation for next-day delivery on Hazelnut beans and another from CoffeeMate offering a personalized discount for Hazelnut creamer.

Ten minutes later, Derek heads for his garage. "Hey Google, remind Sarah to pick up my dry cleaning", is recorded by multiple Nest speakers throughout the house and immediately uploaded to the cloud, where it will be reviewed and transcribed by "language experts" for purposes of improving speech technology. Footage on his Nest Secure Alarm system completes a scan of Derek's face before rolling up the garage doors. A few minutes later, it shows a Tesla Model S slowly pulling out of the driveway and speeding away. The smart lock on the garage door clicks, logging a departure at 7:18am. All lights are off. The house is silent.

A month later, Sarah accuses Derek of cheating and files a divorce lawsuit. Her lawyers request a search warrant from the district court to access Derek's Google Nest records. During the course of the month Derek was suspected of cheating, data from his Nest smart clock showed that he frequently slept past 8am and arrived late at work. Motion sensors divulge that he came in through the backdoor past midnight on sporadic occasions. The Kohler system and Keurig machine confirms missed showers and stronger espressos scheduled earlier in the mornings after Derek's cheating allegations. Nest speaker records disclose auditory indications of fornication at times when Sarah claims to be out of the house. Finally, facial recognition from his Nest Secure Alarm and smart door lock reveal a series of unfamiliar female countenances.

1.1 - The Dangers of Smart Home Technologies

The fabricated story above warns of a chilling, yet entirely realistic, Orwellian future. The growing number of smart gadgets in the home capable of eavesdropping on and recording every aspect of life, from confidential conversations to shower habits, pose severe threats to individual privacy. These devices, also known as smart home technologies (SHTs), collect massive amounts of information ostensibly to make everyday processes more efficient. The data is also uploaded to the cloud where the manufacturers are able to use or sell this personal information in obscure and largely unregulated manners.

Worryingly, many users of SHTs are oblivious to the type of personal information that can be inferred from the data collected by their devices. Lowering the barrier to understanding and controlling SHT data collection is critical for ensuring that otherwise oblivious users have privacy protections by default with minimal burden to actively configure privacy features.

As the growing popularity of SHTs gives both corporations and law enforcement access to unprecedented amounts of deeply personal information, questions arise for both users, commercial producers, and law enforcement about user privacy regarding the external entities who create, manage, track, and regulate these devices: What data do SHTs collect? Where and how is the data stored? Who has ownership and access to the data? Before SHTs, such sensitive personal information would be difficult, if not impossible to infer. Even if such information is willfully handed over to a third party or legally accessed by law enforcement, the sheer amount of sensitive information will undoubtedly encroach upon personal liberties and reasonable expectations of privacy. In short, users must be able to learn about and choose the data that can be uploaded by their SHTs, as it is dangerous to allow the sheer amount of information that may be inferred through SHTs to remain unregulated.

While it can be argued that one way for consumers to avoid any potential privacy invasion is to simply avoid using SHTs, the disruptive paradigm of such a pervasive physically connected world will make opting out difficult for effective integration in society. Therefore, as such devices proliferate, manufacturers should disclose and offer clear ways for consumers to opt-out of features that may violate their privacy.

In the following paper, we analyze the harms inflicted by privacy concerns about SHTs from two perspectives: consumer protection and law enforcement. Following that, we recommend a new industry standard to the makers of SHTs that prioritizes user control by allowing them to determine what kind of data is collected by their SHTs and how that data is stored and shared.

1.2 - What are Smart Home Technologies?

SHTs consist of sensors, monitors, appliances, and devices in the home that are networked together to enable automation and remote control of everyday processes. Collectively, such devices are part of the Internet of Things (IoT). Appliances that can be networked include lighting mechanisms, windows, garage doors, fridges, TVs, heating systems, and washing machines.[2] SHTs are designed to make life more convenient by saving time and effort through automating processes such as turning lights on and off or automatically adjusting room temperatures based on the user's daily movements. Other than convenience, SHTs can also enhance safety. In a pandemic, video doorbells can reveal visitors and allow the owner to talk without having to risk exposure to the coronavirus.

There are an estimated 27 billion IoT connected devices in 2017, growing at 12% every year to reach more than 125 billion devices by 2030. In the US alone, the household penetration of SHTs is expected to grow from 32% in 2020 to 57% in 2025.[3]

The most popular SHTs can be grouped into four categories: entertainment, home monitoring/security, smart speakers, and connected utilities (lighting, thermostats, etc.). Each category of SHTs gathers different types of data, including audio, video, biometrics, location, and usage patterns, among others. Entertainment products consist mostly of smart TVs and digital media adapters, such as the Samsung smart TV, Apple TV, and Roku devices. Home monitoring/security devices range from door and window sensors to door locks to IP cameras. Smart speakers, including Amazon's Alexa and Google Nest Hub, track and store all questions and commands received from users. These voice assistants often have control over other SHTs. The information monitored, collected, and stored by these devices includes everything from financial data to information on medical conditions, physical fitness, shopping routines, music preferences, browsing behavior, and much more.

The proliferation of SHTs, and consequently, data collection, introduces new privacy risks in the home. While some privacy risks posed by video and audio enabled devices are obvious, many users are simply unaware of the potential privacy harm from devices that do not necessarily record audio or video. The data collected by these devices, such as light bulbs and thermostats, is often subject to algorithms that infer more sensitive information, including sleep patterns and home occupancy. Access to any combination of this data can make it possible to watch what someone is doing from anywhere in the world, reveal when someone leaves the house, and show their movements between rooms.

[2] R.J. Robles, T. Kim. Applications, systems and methods in smart home technology: a review. International Journal of Advanced Science and Technology.

[3] Statista.

When analyzing the issue of SHTs and the potential resulting harms, there are several competing interests that must be reconciled - those of the users, commercial developers of SHTs, and law enforcement organizations. Users cite the convenience and connectedness that SHTs can provide as primary reasons for their adoption of such technologies. However, they are also concerned about who might be seeing the data that the devices can collect, because of the details it could reveal about their personal home lives.[4] Companies that produce SHTs care about providing high-level functionality and features in their devices, while also making a profit. Their developers say that the data they collect is used to improve the quality of their service. What is left unsaid is that the data can also be sold to advertisers, insurance companies, and even potential employers for purposes that consumers never intended their data to be used for.

Finally, law enforcement agencies care about being able to access data from SHTs as a way to collect evidence during an investigation. For instance, in the case of *Arkansas v. Bates* (2015)*,* prosecutors ordered Amazon to turn over audio recordings from an Echo device that was potentially related to a murder. In turn, Amazon tried to refuse the request, arguing that they wanted to protect the privacy rights of their customers from the government, especially when the data being sought may include expressive content protected by the First Amendment.[5] In this instance, Amazon dropped its argument after the defendant authorized it to release the recordings. Nevertheless, this case is an example of what can happen when law enforcement, SHT producers, and user values all conflict at once. In the following sections, we will further analyze the harms that can result when commercial and law enforcement interests are prioritized over those of the users.

1.3 - Major Violations of Consumer Privacy

A cursory analysis of current industry practices reveal abysmal privacy practices. Commercial producers do not explicitly disclose data collection and management practices to users. They quietly harvest, analyze, and sell personal data to third parties such as advertisers as a source of revenue. For example, Amazon employs thousands of workers to actively listen and transcribe audio from its home devices, a practice not documented in its privacy policy.[6] The company has also admitted to storing some audio transcripts indefinitely. Similarly, Apple has been reported to use external contractors to review recordings from Siri, its voice-activated assistant that can be easily triggered on accident by similar words. These contractors have access to Siri recordings of

[4] Serena Zheng, Noah Apthorpe, Marshini Chetty, and Nick Feamster. 2018. User Perceptions of Smart Home IoT Privacy. Proc. ACM Hum.-Comput.

[5] Stanford, J. (2017). Memorandum of Law in Support of Amazon's Motion to Quash Search Warrant.

[6] Jacobson, A., Gold, J., Hodge, N., & Widmer, L. (2019, September 27).

"confidential medical information, drug deals, and recordings of couples having sex".[7] Although Google allegedly only keeps copies of voice clips when devices are directly activated, there have been a plethora of instances in which Nest devices have been triggered by background noise. For example, a written transcript of a user's private conversation to a friend said: "If you ever get booked down to my house for some reason the key safe for the back door is 0783".[8] Facebook and Microsoft partake in similar practices.

Large technology companies are constantly improving their SHTs, which necessitate even more data collection or advanced inferences. Recently, Amazon filed a patent application for an algorithm that would let its Echo devices identify statements of interest such as "I'm craving chocolate" to build profiles on anyone in the room and target them for related advertising. A network of such SHTs would be able to build a comprehensive chart of a family or individual's patterns, monitoring everything from screen time to hygiene habits to travel schedules. Similarly, Google also filed a patent in 2018 that would establish an interconnected smart home system that can detect the status and activities of persons in the household via audio or visual cues.[9] Thus far, even without cameras, Google Home devices can use audio data to determine from regular household noises when different individuals arrive home, when they usually eat dinner, whether they cook or order out, how often they clean, etc. By adding a camera to this product, Google will be able to see what the user cooks, what brands of kitchenware are used, and how often the user stocks certain items in the fridge. All this information can be appropriated by third parties (such as insurance agencies and employers) and used to make inferences about an individual or family's health conditions, income status, and more.

Continuing with this example, it is important to understand what Google can do with this data. By watching and listening both inside and outside the home, Google amasses an incredible amount of information on device owners and anyone else inside the home at any time, including non-consenting individuals. In a 2016 patent ("Privacy-aware personalized content for the smart home"), Google gives an example: "a client device may recognize a tee-shirt on a floor of the user's closet and recognize the face on the tee-shirt to be that of Will Smith. In addition, the client device may determine from browser search history that the user has searched for Will Smith recently. Accordingly, the client device may use the object data and the search history in combination to provide a movie recommendation that displays, "You seem to like Will Smith. His new movie is playing in a theatre near you."[10] While none of these examples are blatant violations of the law, they undeniably curtail individuals' reasonable expectations of privacy and encroach upon the sanctity of the home.

[7] Hern, A. (2019, July 26). *Apple contractors 'regularly hear confidential details' on Siri recordings*. The Guardian.

[8] *How Google and Amazon are 'spying' on you*. Consumer Watchdog.

[9] Podracky, J. (2018, December 5). The Next Phase of Smart Home Tech: Ethical Implications of Google's New Patent. Data Science W231 Behind the Data Humans and Values.

[10] Zomet, A., & Urbach, S. R. (2019, October 22). Privacy-aware personalized content for the smart home.

While Amazon and Google's privacy policies allow owners of Echo or Nest devices to delete their voice recordings or opt out of some data collection, the default settings store all information and the process of deleting records is complicated. Additionally, consumers only have access to basic information they explicitly share with the companies, such as name, address, and payment options, but not information discretely collected or inferred.[11] While requiring users to specify privacy settings for every type of data on every SHT device they own would impose an excessive burden upon them, it is clear that current SHT defaults are insufficient to protect privacy. It is outside the scope of our paper to explore the details of an appropriate default privacy setting on top of our recommended individualized privacy policies below; however, we urge future studies to account for the users who are not compelled to create their own.

1.4 - Law Enforcement and the Sanctity of the Home

When SHTs are able to collect and store information about every detail of a person's home life, it becomes possible for law enforcement to access that information when collecting evidence. Because of these devices, new data exists that would have been previously inaccessible to law enforcement, such as information on which room someone is in and everything they said in a certain room. Some data should never leave the home. The legal history in the United States has shown a strong preference for protecting privacy in the home, but it has often been left to the courts to decide how new technologies fit into the existing privacy framework. As shown by the cases below, the Court has tried to keep law enforcement from using technology to invade the home, but only after harms have already been done.

In their classic 1890 article titled "The Right to Privacy", Supreme Court Justices Warren and Brandeis express their concern for the ways in which then-recent inventions such as instantaneous photographs and newspaper gossip were "invad[ing] the sacred precincts of private and domestic life."[12] They foresaw that society and technology were making it possible that "what is whispered in the closet shall be proclaimed from the house-tops"[13] (as cited in Warren and Brandeis, 1890). Their article demonstrates that early in the legal history of the United States, there was concern that the progression of various technologies could be too intrusive. The Justices believed that individuals should be protected from the potential damage to reputation, trust, and dignity that can result from such intrusions. It is almost hard to comprehend the extent to which current SHTs have realized Brandeis and Warren's fears.

The Supreme Court has held that the Fourth Amendment protection applies when a person has a reasonable expectation of privacy in the place or item searched or seized, as outlined in *Katz v*

[11] Amazon. (2011). Amazon.com Privacy Notice. Amazon.

[12] Warren, & Brandeis. (1890, December 15). The Right to Privacy. Harvard Law Review.

[13] Luke 12:3 New International Version

United States.[14] Individuals hold more than a reasonable expectation of privacy in the data generated by their SHTs because of the quality and quantity of information held by the devices. Additionally, the Court has found the home to be an area that deserves unique protections. Therefore, the reasonable expectation of privacy can definitively be applied to data collected in the home. Subsequently, any data that is generated by SHTs deserves to be protected by a warrant requirement before becoming accessed by law enforcement. Anything less would inevitably enable the mass surveillance of the private lives of SHTs users.

In *Riley v. California*, the Supreme Court held that law enforcement's potential intrusion of privacy is no longer limited only to the physical realm, due to the enhanced capabilities of cell phones. Before technologies such as smartphones, law enforcement could only obtain limited data about an individual, let alone information from within their home. Such limited information gathering "constituted only a narrow intrusion on privacy".[15] However, like smartphones, SHTs have the capacity to collect and store massive amounts of information. Especially when paired with other user information such as home address or calendar details, it gives a complete record of user preferences, behavior, and surrounding situations. The data collected from SHTs also differ from physical evidence in both quantity and quality. Browsing history or behavioral patterns stored by SHTs can reveal the user's private interests or concerns. Such information would not only give law enforcement access to a user's physical records, but it would also open a window into their mind, a place even more sacrosanct than the home and one in which no warrant was intended to give access to. This is in direct violation of the Court's interpretation of the Fourth Amendment as the central aim of the Framers to "place obstacles in the way of a too permeating police surveillance" and as one that "seeks to secure the privacies of life against arbitrary power".[16]

Similarly, in *Kyllo v. United States*, the Supreme Court ruled that police use of a thermal imaging device to gather evidence about the defendant's activities in his home violated his right to a reasonable expectation of privacy. In their reasoning, they relied on the fact that the government used a device not generally available to the public to "explore details of the home that would previously have been unknowable without physical intrusion".[17] Similar logic can be applied to the data collected by SHTs - both reveal details of the home. Allowing access to this information without user consent is a clear invasion of the sanctity that the Court in *Kyllo* tried so hard to protect.

While the Courts would likely uphold their emphasis on the privacy of the home if presented with a case involving law enforcement's access to SHTs, SHT producers should not wait for such cases to be brought to court before taking action. In the time it takes for a case involving

[14] Katz v. United States. (1967). Oyez.
[15] Riley v. California. (2014).
[16] Carpenter v. United States. (2018). Oyez.
[17] Scalia, A. & Supreme Court Of The United States. (2000) U.S. Reports: Kyllo v. United States, 533 U.S. 27

SHTs to reach the Supreme Court, millions of users face the risk of unjust data access by law enforcement. Our recommendation takes proactive steps to give users greater control over their information, empowering them to protect themselves rather than relying on the justice of the courts.

2. Recommendation

To address the unnecessary amount of invasive, private data generated by SHTs leaving the home and becoming accessible to commercial producers and law enforcement, we recommend a new industry standard for the makers of SHTs. **Our standard requires that users are able to control what types of data are collected and stored by their devices and what data is shared back with the company.** If the fictional Derek Watkins had been protected by this privacy standard, he would have been able to prevent some of his revealing, personal data from being needlessly shared in his divorce case. If he had access to a data report showing the type, quantity, content, and inferences made from the data collected by his SHTs, he would have been able to understand the implications of sharing this data. Derek could have specified that he did not want his motion sensor to record every time he came into the house or that his Keurig should not store data on when he schedules earlier and stronger espressos instead of lattes. Our proposed standard would give users the control they deserve over the details of their home life. SHT companies can meet this standard by changing the way their device is built, their data management practices, or any other number of methods, as long as they follow the guiding principle of greater user control. The details of how the standard will be implemented is left to each company in order to allow them to innovate and decide what is most effective for their devices.

Below, we will provide a few examples of how the proposed standard could be met, as well as an end-to-end example showing how we envision an example system to look and work. Our primary recommendation, the machine learning algorithm detailed in option one, encompasses two steps: first it specifies how manufacturers should format and display data in an intuitive, user-friendly way; and second it allows users to select what data they are comfortable with sharing in a manner generalizable to all data collected by the SHT. We provide additional recommendations for further flexibility, along with a discussion on the benefits of each.

2.1 - Option 1: Generating Individual Privacy Policies using Machine Learning

An option to create a customer-specific privacy policy is to use a supervised machine learning technique to classify all this data into either “private” or “shareable” according to user preferences. However, the amount of information collected from a smart home system can amass well over 200 MB of data every day.[18] Compounded by how long SHTs are run, this tremendous amount of data is constantly increasing. Due to the large amount of data produced, it is difficult for a single consumer to manually label all the information generated by their SHTs. We propose a framework for a machine learning algorithm that chooses the best data to present to users for labeling and that would provide the most meaningful labels for the manufacturer to sort the customer's data into two categories: upload to the cloud or keep private.

We are proposing two specific standards for this algorithm, directed to SHT manufacturers. First, the data produced by these technologies must be available in a format easily understandable to customers. We define easily understandable as conforming to Nissenbaum's theory of Contextual Integrity (CI).[19] Second, customers must be offered a trial period where their SHTs will collect data at the full, intended capacity, where the trial period is determined by the manufacturer. At the end of this trial period, customers must be provided with some way, such as a phone application, where they can review the data collected by their SHT. During this review process, they will be able to grade several data points, which we will call events, and manually state which events they are comfortable with sharing and uploading to the manufacturer, and which data points they would rather keep private. It must be made clear to the customer how various data points impact functionality. Depending on the SHT, it must also be made clear to the customer if any functions or uses are impacted by privatizing specific events. We will explore different possible solutions that could have allowed Derek to have better control and understanding of his data.

2.1.1 - End-to-End Customer Example

First, we can envision how this framework would work at a high level for Derek. Derek purchases his new Keurig programmable coffee machine on January 21, 2020, and he immediately sets it up to produce a latte a half hour after he wakes up every day around 8:30am. For a month, the machine records which coffee is produced with the timestamp. Occasionally, Derek opts for a stronger espresso the mornings after his midnight escapades.

On February 21, 2020, Derek's Keurig app (**Image 1**) sends an alert requesting him to label a set of data selected by our framework to be most indicative of his preferences. After opening his app, he will be led to a screen providing him with several options. Should he click "Label Data," he will be taken to the additional page shown, which will display information about the device as well as the data to label. Afterwards, he may choose to Review Previous sessions, where he may alter any labels or inspect all the events from the trial period, or he may select Label Data to

[18] Higginbotham, S. (2014, July 29). How much data can one smart home generate? About 1 GB a week.
[19] Nissenbaum, H. (2004). Privacy as Contextual Integrity. Washington Law Review.

re-label a new set. He may also select User Settings, or of course, Schedule Coffee for the next day. An interactive prototype[20] is available.

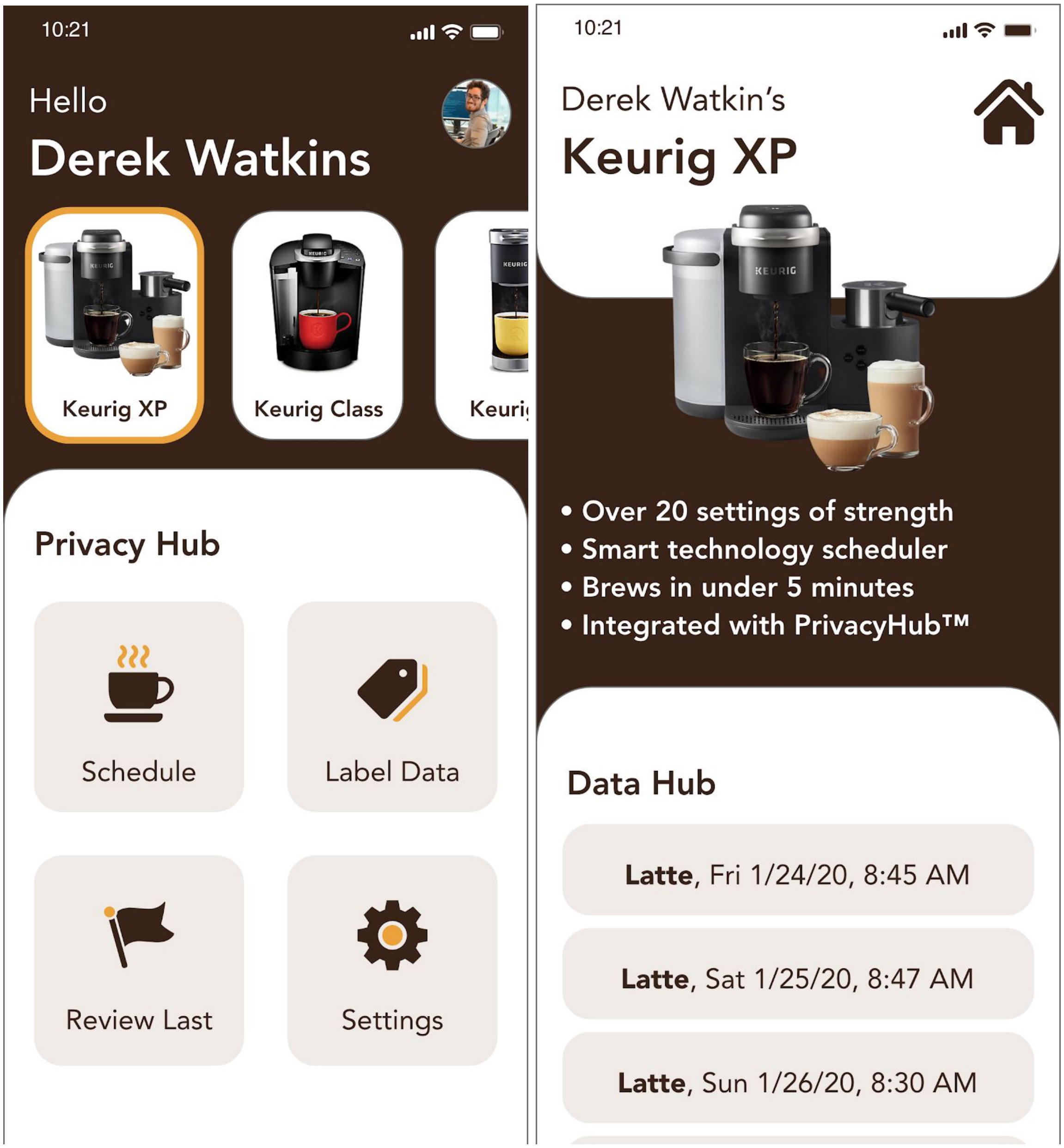

10:21
Hello
Derek Watkins
Keurig XP
Keurig Class
Keuri
Privacy Hub
Schedule
Label Data
Review Last
Settings
10:21
Derek Watkin's
Keurig XP
KEURIG
• Over 20 settings of strength
• Smart technology scheduler
• Brews in under 5 minutes
• Integrated with PrivacyHub™
Data Hub
Latte, Fri 1/24/20, 8:45 AM
Latte, Sat 1/25/20, 8:47 AM
Latte, Sun 1/26/20, 8:30 AM

Image 1: Example User Interfaces For Privacy and Data Hubs

In the backend, the classifier trained by the manufacturer is given the labelled data. The classifier recognizes Derek does not like his Wednesdays or espresso days recorded, and moving forward the machine does not share those events with the manufacturer.

[20] https://xd.adobe.com/view/999f5f57-a53d-4542-ba7f-4601df8aaaa1-cccc/

In the situation outlined above, Derek would avoid being incriminated by his coffee machine as the early-morning espresso days would not be shared and saved by Keurig. Additionally, Derek can also specify against the collection of his favorite coffee flavor to prevent excessive advertisements from CoffeeMate. Applying the same recommendation to the other SHTs in Derek's house would also allow him to modify his privacy policies to best suit his needs.

Our recommendation is to first format the data that Derek sees in the DataHub (**Image 1**) in a way that makes it easier for the customer to understand what information is collected and the implications of sharing the data with the manufacturer relative to their own home situations; specifically we provide techniques for determining if the information allows for profiling, tracking, and identification. Second, the data provided to Derek for labelling should be selected using pool-based active learning, which will ensure that Derek will only label the most important information rather than all the information created by his SHTs. We will explore the technical aspects of these two sub-recommendations.

2.1.2 - Data Formatting and Nissenbaum's Theory of Contextual Integrity

Contextual Integrity (CI) asserts that information flows govern every aspect of our lives, and that privacy is provided and protected by appropriate information flows. When applied to SHTs and the unregulated sharing of private information, this is all the more apparent. Our framework for a machine learning algorithm aims to hone in on each customer's definition of an appropriate information flow, which is "what information about persons is appropriate, or fitting, to reveal in a particular context" (Nissenbaum 120). Specific parameters, which define if an information flow is appropriate, are defined by data subject, sender, recipient, information type, and transmission principle.

On the technical side, our proposed industry standard for data which aligns with these principles is to provide the data in a format as defined by CI. Bugeja et al.[21] where, "using the CI as an overarching framework, an IoT-based [SHT] S can formally be described as a tuple (H, N, U, L, D, P) where H: house, N: nodes, U: users, L: links, D: data, and P: policy." We incorporate these tuples to define each event, where instead of S being a single technology, it will be an event recorded by the technology. Accordingly, each of the parameters in this tuple can be further broken down into specific categories and additional parameters which will be used in processing and selecting in the algorithm. We will go into some specific examples to illustrate these below.

[21] Bugeja, J., Jacobson, A., &amp; Davidsson, P. (2020). A Privacy-Centered System Model for Smart Connected Homes. IEEE.

H, or the house, may be defined by the specific zone or area of the house where the SHT is located. This may not be applicable for certain technologies which are located in only one zone or if the zones are not labeled by the technology, but by narrowing this parameter as much as possible where applicable for each separate event, customers will be able to select and create more specific privacy policies.

N, or nodes, are defined by the type of physical components the SHT includes. There are three main categories SHTs fall into: connected device, mobile device, and backend. We define functions for each to be incorporated into the parameters. Connected devices and mobile devices may "implement several core capabilities ⊆ {connectivity, sensing, actuating, interaction, storage, processing, gateway, programming, remote-admin…" Backend capabilities fall into one of {edge, cloud}, where edge indicates a technology performing storage and processing within H (home) and cloud indicates that the processing for this event occurs outside of H.

Users are given by U, and are defined by a list of all the data subjects, data controllers, and data users involved with the event S. Data subjects, or d_s, are the customers in question, most likely a human user who may be personally identified by the data collected by the SHT. Data collectors, or d_c, are those involved with collecting the data, and will usually be the manufacturer of the technology. The data user, or d_u, are those who are able to access and use the data collected by the SHT if the data were not privatized. It is critical to provide this data set to the customers to allow them to understand who is handling each event and data point collected by their SHT.

Links, or L, ordinarily would indicate the inputs and outputs of each information flow created by the SHT. Here, we will instead use it to indicate the time the event was registered in the format [YYYY]:[MM]:[DD]:[HH]:[MM]:[SS].

For data, we may incorporate the parameters given by U should the manufacturer find it to be more flexible. Additional data that our proposal requires is d_i, or data item, d_p, or processing purpose, and d_t or retention time. The data item is the bulk of the information for each event S. This may include what N detected or collected, for example. The processing purpose informs why the d_i was collected. The retention time describes the condition for storing this data and not privatizing it, such as how long the data will be stored.

The policy P is defined by Bugeja et al. as "a set of tuples (l, d_{pi}, s, r, c) where l: link group identifier, d_{pi}: data permissions, s: sender ∈ (N∪U), r: recipient ∈ (N∪U), and c: condition for transmission specifying when the data is transferred to the recipient(s)." P primarily defines how information will be transferred from d_s to d_u.

Correct compliance with these parameters automatically provide metrics for identification, localization and tracking, and profiling (equations provided in Bugeja et. al.). These parameters must be computerized and provided to the customer as additional information during labelling, and we provide examples for this in the UI in **Image 2**.

By providing these parameters, we can standardize the events in a way digestible and able to be used for training by our proposed machine learning algorithm. This proposed events format also enables the customer to understand what data points are collected by their SHTs, as well as the implications and uses for each.

2.1.3 - Pool-Based Active Learning With Support Vector Machines

After purchasing a SHT and allowing the trial period to pass, the customer will be prompted to "label" a set of events, or data points, as "shareable" or "private." These events are formatted as described above, as a tuple (H, N, U, L, D, P). Using the events collected over the trial period and the associated labels provided by the customer, we can effectively separate the data points going forward according to the customer's unique and informed preferences.

However, the question remains in how to balance the massive quantities of data accumulated through the trial period with accommodating the customer and creating a quick labelling period. We propose using support vector machines to implement pool-based active learning.

Active learning allows learning algorithms to choose the data they want to learn from in order to perform better with less data. This type of learning reduces the algorithm's need for large quantities of data, which is precisely the type of learning we propose should be implemented for fast and confident training. Pool-based active learning allows the learning algorithms to select from a massive pool of unlabeled but registered data.

Support vector machines (SVMs) are a type of supervised learning that act as policies, which may be paralleled as hyperplanes, separating data based on given parameters, which create the feature space that are parallel as graphs. Here, the policy would separate events that are "shareable" and events that are "private", and the graph is a multi-dimensional one with the various parameters as the axes. SVMs use events lying close to the policy to inform it if the policy needs to be adjusted in each round of learning. The customer will operate as the oracle labelling the data. An example showing how support vector machines work can be found in **Figure 1**.

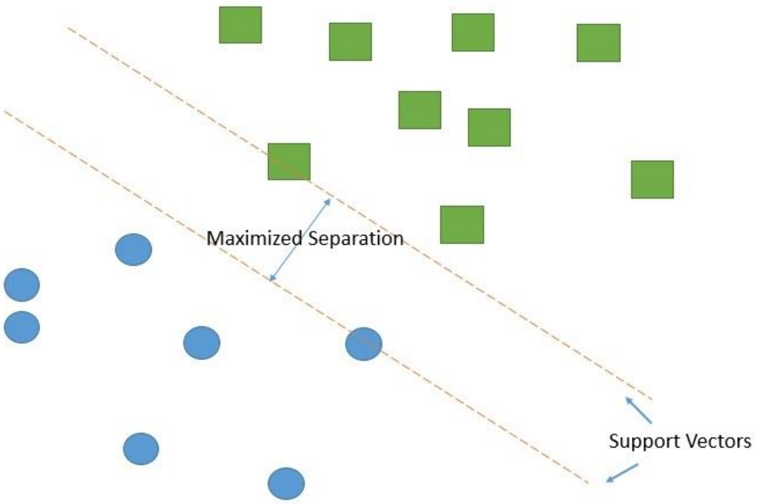


Figure 1. SVMs use events lying close to the policy to inform how the vectors should be adjusted, aiming to effectively separate clusters of data..

Using SVMs with pool-based active learning allows the learning algorithm to find the best queries from the available pool of data, which are events recorded during the trial period. Tong et al. described the specifics behind creating a function which selects the next event to be labeled.[22]

Moving back into Derek's point of view, he will be presented with an event which is formatted in accordance with our proposal. Derek will also be given information on how this event will be used, if it can be used to identify him, and other relevant information to aid his decision. Then, he will select if this specific event is "shareable" or "private." This decision will be sent back to the SVM, which will then use pool-based active learning to select the next event that will be most informative for the final policy. This new event will be presented to Derek, thereby restarting the process. The manufacturer may test the number of such events necessary for their specific SHT to create an accurate and effective policy that may be appropriately individualized to the customer. We leave an area of innovation open to the manufacturer in creating effective rules based on the labelled data provided.

2.1.4 - Using the Labels and Example Data with User Interface

Our proposed pool-based active learning algorithm produces an optimal pool of labeled data, which may then be used to train a classifier. Upon implementation and selection, the customer

[22] Tong, S., & Koller, D. (1998). Support Vector Machine Active Learning with Applications to Text Classification. *Proceedings of the Seventeenth International Conference on Machine Learning (ICML-00)*

will have completed their trial period, and the SHT will continue functioning, except privatizing the data in accordance with this classifier.

Bringing back Derek, we will now dive into an example of an event formatted as described in Section 2.1.2. Take Derek's set of ADT-manufactured motion activated smart lights set up around the house in the following rooms: Master Bedroom, Bathroom, Living Room, and Kitchen. Each of these smart lights will produce an event each millisecond: motion detected (turn light on) or motion not detected (keep light off). If motion is detected, the event will be shared with the manufacturer and the light will turn on. If motion is not detected, the event will not be shared and the light will remain off accordingly. We provide an example of parameterized data in accordance with the above guidelines in **Table 1**. The parameter d_{pi} under P will be the label provided by the customer at the end of the trial period and noted at TBA for now, but the default will be "shareable" for any calculations. Additional information on identification, localization and tracking, and profiling will be provided to the customer based on these parameters.

H	**N**	**U {ds, dc, du}**	**L**	**D {di, dp, dt}**	**P {l, dpi, s, r, c}**
Master Bedroom	Backend {edge}	{Derek, ADT, ADT}	2020:01:22:08:30:20	{off, light, 0ms}	{Derek's home, TBA, Derek, ADT, di = on}
Bathroom	Backend {edge}	{Derek, ADT, ADT}	2020:01:22:08:45:40	{on, light, 8.64e7ms}	{Derek's home, TBA, Derek, ADT, di = on}
Living Room	Backend {edge}	{Derek, ADT, ADT}	2020:01:22:08:46:10	{off, light, 0ms}	{Derek's home, TBA, Derek, ADT, di = on}
Living Room	Backend {edge}	{Derek, ADT, ADT}	2020:01:22:08:47:15	{on, light, 8.64e7ms}	{Derek's home, TBA, Derek, ADT, di = on}

Table 1: Data Collected from Derek's Smart Lights

As we see in the table, H tells the customers which zone in the house the SHT device (given by d_p) is located in. The column for N provides information about what type of a device d_p is, and U provides basic information about the user and manufacturer's relationship. L gives the timestamp when each event was recorded, and D gives the bulk of the information of what the event was, what type of device recorded it, and how long the data will be stored for. Finally, P tells Derek additional information about the data. Within the user interface for the SHT in the labeling step, we recommend adding information on whether the customer can be identified, tracked, or profiled based on this data and provide examples for this in **Image 2**, but these rows are the baseline for our recommendation.

We will now consider how Derek will label the data collected by his Keurig, and how to integrate the information into the UIs displayed in **Image 1**. As discussed previously, the data will be formatted in a clear manner. We provide examples of the parametrized data in **Table 2**.

H	**N**	**U {ds, dc, du}**	**L**	**D {di, dp, dt}**	**P {l, dpi, s, r, c}**
Kitchen	Backend {edge}	{Derek, Keurig, Keurig}	2020:01:22 :08:30:20	{latte, coffee, 0ms}	{Derek's home, TBA, Derek, Keurig, di = on}
Kitchen	Backend {edge}	{Derek, Keurig, Keurig}	2020:01:23 :08:16:40	{cappuccino, coffee, 8.64e7ms}	{Derek's home, TBA, Derek, Keurig, di = on}
Kitchen	Backend {edge}	{Derek, Keurig, Keurig}	2020:01:24 :08:45:10	{latte, coffee, 0ms}	{Derek's home, TBA, Derek, Keurig, di = on}
Kitchen	Backend {edge}	{Derek, Keurig, Keurig}	2020:01:25 :08:47:15	{latte, coffee, 8.64e7ms}	{Derek's home, TBA, Derek, Keurig, di = on}
Kitchen	Backend {edge}	{Derek, Keurig, Keurig}	2020:01:26 :08:30:25	{latte, coffee, 8.64e7ms}	{Derek's home, TBA, Derek, Keurig, di = on}
Kitchen	Backend {edge}	{Derek, Keurig, Keurig}	2020:01:27 :07:30:20	{espresso, coffee, 8.64e7ms}	{Derek's home, TBA, Derek, Keurig, di = on}
Kitchen	Backend {edge}	{Derek, Keurig, Keurig}	2020:01:28 :08:41:55	{latte, coffee, 8.64e7ms}	{Derek's home, TBA, Derek, Keurig, di = on}
Kitchen	Backend {edge}	{Derek, Keurig, Keurig}	2020:01:29 :07:25:30	{espresso, coffee, 8.64e7ms}	{Derek's home, TBA, Derek, Keurig, di = on}

Table 2: Data Collected from Derek's Coffee Machine

We can dive into the data collected from the Keurig coffee machine. Derek purchases the machine on January 21, and programs it that night to produce coffee the next day, which is when the trial period starts. Keurig sets the trial period to be a full month, so on February 21 Derek is prompted by the machine to label the given data, with one full week given in **Table 2**. Derek is interested in privatizing all instances where he programs the machine to produce espresso before 8 AM, as well as all Wednesdays. In the above data, this would privatize January 22, January 27, and January 29, meaning d_{pi} would be set to "private." However, having Derek label all such instances for each day between January 21 and February 21 would be tedious, so pool-based active learning using SVMs would be employed in the backend to only provide the labels where the most information may be extracted, reducing the workload expected of Derek.

The data in **Table 2** will be displayed in a user-friendly manner, and we provide example user interfaces (UIs) below for the data itself, providing a template of our vision for how the labeling process may be laid out for the customer. The beginning home screen was displayed in **Image 1** and discussed in Section 2.1.1. After clicking “Label Data” and beginning a labelling session, Derek will be prompted by the manufacturer’s interface to provide labels for the data collected from the trial period. He is presented by the UI displaying the data, with an example in **Image 2**. He sees the data selected by pool-based active learning and views the individual events as formatted using Nissenbaum’s Theory of Contextual Integrity.

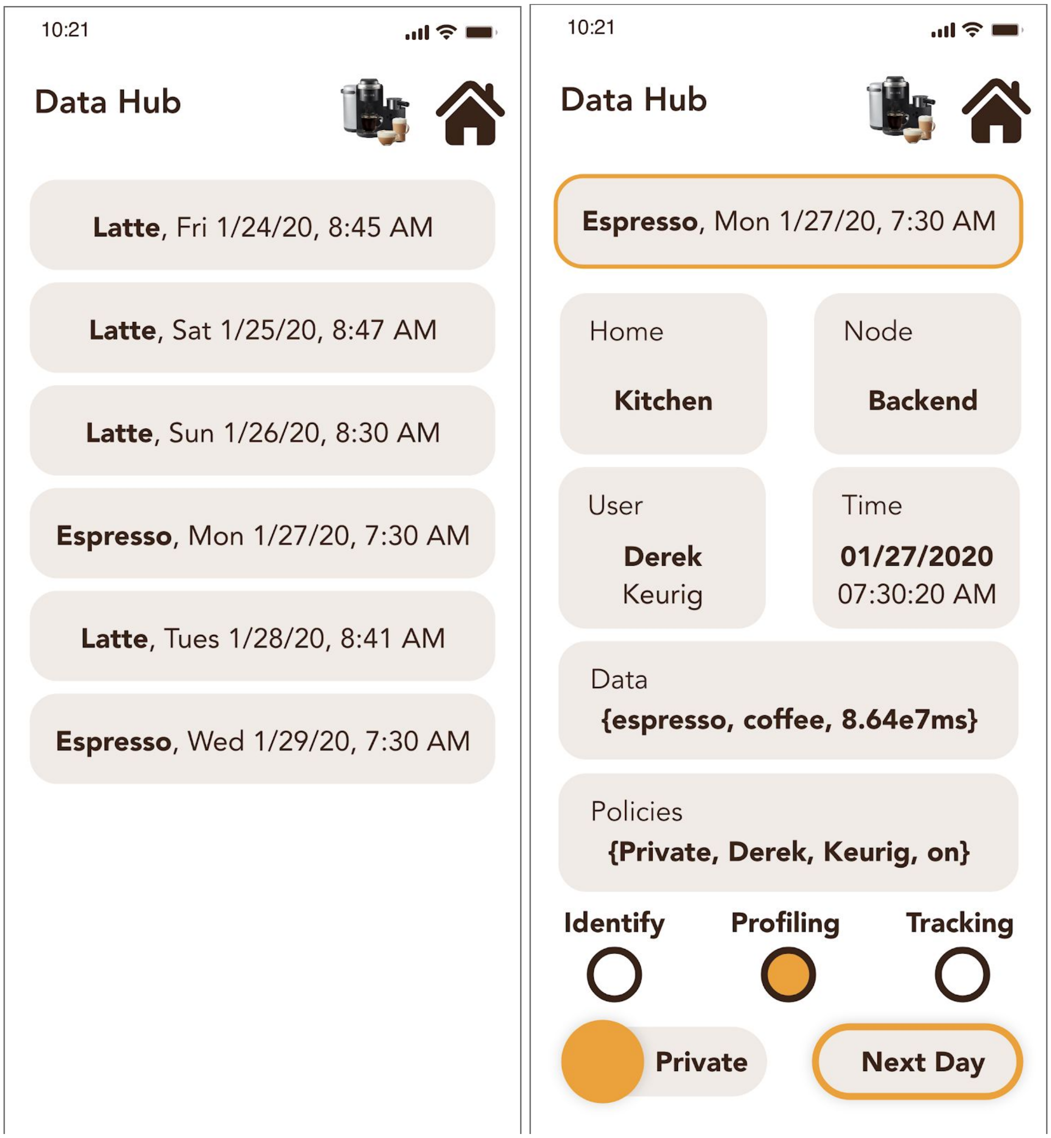
Once he selects one event to start the labelling, the app will then run through several additional events to label, chosen by pool-based active learning. The data available will be laid out in an educational and open manner, with an option to share the event or privatize the event.

As Derek views the data collected, he realizes that the machine has recorded his espresso outliers. He also decides that he does not want his Wednesday coffee habits recorded. Using the app, Derek sets the events from January 22, January 27, and January 29 to private.

Upon completing all the labels, Derek will be taken back to the home screen displayed in **Image 1**. From here, he may view the previous session, start a new labelling session, visit his settings, or schedule his coffee. All data collected by the SHT from this point will be separated into privatize or shared according to the manufacturer’s classifier that is trained on the data labelling by the most recent session.

The user interfaces shown in **Image 1** and **Image 2** display several options for the application or other technology designed by the manufacturer. We mockup a prototype for Keurig, creating a privacy and data hub where the customer may schedule their coffee, label their data, or modify user settings. The customer may also modify past session’s data. The interface for labelling data clearly lays out all the data relevant to the customer. An interactive prototype[23] is available as well. This interface is designed to be sleek and minimal, with a focus on informing the customer smoothly and transparently.

[23] https://xd.adobe.com/view/999f5f57-a53d-4542-ba7f-4601df8aaaa1-cccc/



Latte, Fri 1/24/20, 8:45 AM

Latte, Sat 1/25/20, 8:47 AM

Latte, Sun 1/26/20, 8:30 AM

Espresso, Mon 1/27/20, 7:30 AM

Latte, Tues 1/28/20, 8:41 AM

Espresso, Wed 1/29/20, 7:30 AM

Espresso, Mon 1/27/20, 7:30 AM

Home **Kitchen**

Node **Backend**

User **Derek** Keurig

Time **01/27/2020** 07:30:20 AM

Data **{espresso, coffee, 8.64e7ms}**

Policies **{Private, Derek, Keurig, on}**

Identify ○ Profiling ● Tracking ○

Image 2: Example User Interfaces For Labelling

However, this method is hands-on, with some innovation left to the manufacturer in developing a classifier that can use the labels selected by our recommendation, as well as a greater amount of engagement required by the customer to take the initiative to label all the data. We offer simpler options to continue allowing customized privacy policy based on user choices.

2.2 - Option 2: Opt-In by Data Type

Another way that our proposed industry standard could be fulfilled is by allowing users to specify privacy settings based on the general type of data being collected. The default settings would maximize privacy by not sharing any data with the company that is not absolutely necessary for functionality, and the users would have to actively opt-in for each type of data that they want to be collected, stored, and shared. For example, when users set up their device for the first time they could be prompted to select a privacy option for data types such as images/videos, biometric information like fingerprints and facial recognition, voice recordings, location information, and time-stamped events from devices like motion sensors, among other potential data types. Users would also receive monthly data reports detailing how their data is being collected and used so that they know if they want to modify their opt-in preferences.

If Derek had been under this kind of privacy policy, he would have had to opt-in for any of his location information or other time-stamped events to be stored and shared. He also would have had to actively allow the images of the strangers from his front door camera to be collected. This would have given him the awareness he needed to have more control over his privacy and data. This option does not allow for the same level of specific, agile, and personalized control over the data as in the machine learning algorithm above. However, it still meets the intent of the standard, which is to let users dictate what data from their SHTs is stored and shared, while being potentially easier to implement and simpler for users to operate. We provide an example implementation under the Settings option in our interactive prototype[24].

2.3 - Option 3: Accessible Privacy Settings

The final example that we will provide to demonstrate various methods of meeting our proposed standard of user control over data is providing monthly data reports and an accessible privacy settings menu that users can navigate to and change at any time. It is important that the privacy settings display clear, comprehensible information about what kind of data is being collected, stored, and shared. If it is too vague, users might be misled as to what is actually happening with their data. Some makers of SHTs are currently working with various implementations of this option.

For example, Amazon's Alexa offers a Privacy Settings page that can be accessed through the app at any time. It allows users to delete voice recordings, manage third parties that have access to personal information, see the status of other SHTs it is connected to, among a few other options.[25] While this is a good start to giving users more control, it still fails to make obvious to users what is happening with their data early in the process of owning their device. In order to find out any information about the status of their data privacy, users must actively seek out the privacy settings. The settings even obscure some information with misleading descriptions. For instance, it displays a switch labeled "Help Improve Amazon Services and Develop New Features"[26] that users can toggle. What the description doesn't make clear is that enabling this option allows Amazon employees to listen to voice recordings from your Alexa.

[24] https://xd.adobe.com/view/999f5f57-a53d-4542-ba7f-4601df8aaaa1-cccc/

[25] Amazon. *Amazon.com Alexa Privacy Settings*.

[26] Cipriani, J. (2019, August 06). Stop Amazon employees from listening to your Alexa recordings.

This system could be improved upon by writing the privacy settings in clear, understandable ways and by delivering monthly data reports to users that show what kind of data was collected from their devices that month and what was done with the data. For example, the report would specify whether any data was sold or shared to third parties. It could also alert users to any access requests from law enforcement. There are many forms that the reports could take, but the overall goal is to achieve greater transparency and trust between the company and user. The report would also make it clear that to change any of the current data processes, the user could go into the privacy settings and update their preferences. Receiving these reports would inform users about how their data is being used and motivate them to take a more active role in controlling that process. For instance, a report could have alerted Derek to the fact that his SHTs were keeping track of which days he showered and drank coffee. If he found this invasive, he could have used the privacy settings menu to keep this information from being stored. This combination of reports and an accessible privacy settings menu would be an effective way of ensuring that users have more control over what data is collected, stored, and shared.

3. Discussion

3.1 - Comparison of Recommended Individual Privacy Policy Implementations

The machine learning technique outlined in Section 2.1 provides immense potential to generate incredibly specific policies. By using the data generated by the customer and allowing them to select which events they would like to share or keep private, the user is educated about what is collected by their SHTs in addition to being able to choose their level of privacy with the manufacturer. Our use of Nissenbaum's Theory of Contextual Integrity as a baseline for data formatting highlights the data's sensitivity, information flows, and identifiability in a manner clearly laid out for the customer. We reduce the load on the customer and avoid the need for large amounts of data to be labelled using pool-based active learning. This combination keeps the user experience positive while collecting the necessary data to generate a highly individualized privacy policy. By having customers label their own data, we further emphasize that they are able to see and label exactly what data their devices are creating, bringing the problem of data privacy closer to the customer.

We recognize that classifiers have some measure of inaccuracy and that customers may not wish to label so much data at the end of the trial periods. Additionally, implementing the machine learning algorithm may impose unrealistic computational constraints on smaller companies. Therefore, our recommendation also includes standards for opting-in based on data type as well as enforcing accessible privacy settings. While these policies may not be as specific and individualized, they encourage transparency with the customer and allow them to have a measure of control over their data.

3.2 - Addressing Potential Counterarguments

We anticipate that our recommendation will be met with objections from several of the stakeholders involved in this issue. In any situation dealing with conflicting interests and groups of people, this is to be expected. The two primary stakeholders that would likely object to our recommended industry standard of user control of data are commercial producers of SHT and law enforcement officials. There is also a concern that too much burden is placed on users who might not have the time, skill, or interest in crafting their privacy policies. We will demonstrate that our recommendation takes their concerns into account and is in the best interest of society as a whole.

The companies that produce and operate SHTs can assert that an industry standard focused on user control would render their devices ineffective and unable to perform intended capabilities. While it is true that not being able to collect some types of data would likely limit aspects of device functionality or convenience, it is important that makers of SHTs find a balance between functionality and privacy concerns. A 2019 survey of Americans found that 81% of people believe that the potential risks of data collection by companies outweighs its benefits.[27] With this increasing user desire for privacy and control, it is beneficial for companies to adapt their standards in order to maintain trust and transparency with the user, making them more willing to engage with the device.

Additionally, SHT producers may claim that our proposed standard is an unacceptable burden on them and that they don't have the people, time, or money to follow it. Again, this is a valid concern. However, our recommendation is purposefully left broad in how it would be implemented, leaving companies with the opportunity to innovate and find the most appropriate and effective method for their product. The increased need for options for user control of data would also open up a space for new services and companies that could specialize in privacy-centered technology, leading to more innovation and growth in the industry while striking a greater balance with user control.

The second significant stakeholder group who would likely object to our recommendation is law enforcement officials. As argued in other 4th Amendment cases related to new technologies, such as *Kyllo v. United States*[28] and *Riley v. California*[29], they perceive access to data as crucial

[27] Auxier, B., Rainie, L., Anderson, M., Perrin, A., Kumar, M., & Turner, E. (2020, August 17). Americans and Privacy: Concerned, Confused and Feeling Lack of Control Over Their Personal Information.

[28] Scalia, A. & Supreme Court Of The United States. (2000) U.S. Reports: Kyllo v. United States, 533 U.S. 27

[29] Riley v. California. (2014). Oyez.

to their ability to enforce justice and protect national security. We do not wish to deny them legal, warranted access to appropriate data. However, as discussed in the background section, American legal tradition has long protected the sanctity and privacy of the home. Therefore, law enforcement should understand the difference between data from other types of technology and data from SHTs. Because our recommendation addresses companies giving users control over the actual process of data collection and sharing, it does not complicate the job of law enforcement. They do not have to worry, as long as they have a warrant, about overstepping privacy concerns by seeing data that users did not know was being collected because our recommendation aims to make sure that data does not exist in the first place.

Finally, there is a legitimate concern that our recommendation places too much of a burden on users to specify how SHTs collect and store their data. While some users will take full advantage of their new level of control, others might find the process of labeling data, adjusting their settings, or opting-in by data type to be time-consuming and not worth their effort. This is to be expected. People approach the trade-off to SHTs between privacy and convenience in different ways. For this reason, it is important that the default settings on SHTs still respect the sanctity of the home and users' personal lives. It is beyond the scope of this paper to explore the technical details of how this might work, but we recognize that such work is necessary to move SHTs towards a more privacy-oriented design.

4. Conclusion

This report analyzed the harms posed by SHTs from both consumer privacy and legal perspectives and proposed industry best practices with a focus on user control. Privacy policies for conventional consumer technologies cannot be translated directly to that of SHTs. The sheer quantity and invasive nature of the data collected, ranging from private conversations to shower patterns, makes users even more susceptible to privacy violations. Commercial stakeholders can use this data to make inferences about the lifestyle, habits, and characteristics of their users. Law enforcement agencies could access previously unfathomable amounts of personal data, potentially overstepping their power and jeopardizing the very principle underlying the Fourth Amendment.

In the second half of the paper, to address these harms, we delineated three possible options that fulfill our proposed privacy best practices that would allow users to dictate what type of data is collected and shared by their SHTs. First, we present a technical solution: a machine learning technique that infers privacy policies for each user based on data they label. Second, companies can implement an opt-in by data type system in which users can select the types of data (images, videos, biometrics, audio, location, time-stamped events, etc.) that can be collected and shared. Finally, the third option advocates for both user knowledge and control over their data by delivering automated monthly data reports and offering accessible privacy menus. By offering a comprehensive, yet flexible set of industry standards, our proposal allows corporations the freedom to innovate in the user-privacy ecosystem.

In future efforts, we hope to address the issues that arise from giving SHT users too many options in their privacy settings. From a technical perspective, the current machine learning algorithm may require users to label too much data before their individualized privacy policy can become effective. An area of innovation here would be to fine tune the neural network to place more weight on user-labeled data rather than commercial system defaults. Additionally, from both a commercial and user perspective, future work should aim to strike a more detailed balance between minimizing the burden of excessive choices and maximizing user control over their privacy.

References

Amazon. (2011). *Amazon.com Privacy Notice*. Amazon. https://www.amazon.com/gp/help/customer/display.html?nodeId=468496.

Amazon. *Amazon.com Alexa Privacy Setting*s. https://www.amazon.com/b/?node=19149164011

Auxier, B., Rainie, L., Anderson, M., Perrin, A., Kumar, M., & Turner, E. (2020, August 17). Americans and Privacy: Concerned, Confused and Feeling Lack of Control Over Their Personal Information. Retrieved November 17, 2020, from https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/

Bugeja, J., Jacobson, A., & Davidsson, P. (2020). A Privacy-Centered System Model for Smart Connected Homes. IEEE. Retrieved 2020, from https://ieeexplore-ieee-org.libproxy.mit.edu/document/9156246

California Consumer Privacy Act (CCPA). State of California - Department of Justice - Office of the Attorney General. (2020, July 20). https://oag.ca.gov/privacy/ccpa.

Carpenter v. United States. (2018). Oyez. Retrieved November 15, 2020, from https://www.oyez.org/cases/2017/16-402

Cipriani, J. (2019, August 06). Stop Amazon employees from listening to your Alexa recordings. Retrieved November 17, 2020, from https://www.cnet.com/how-to/stop-amazon-employees-from-listening-to-your-alexa-recordings/

Hern, A. (2019, July 26). *Apple contractors 'regularly hear confidential details' on Siri recordings*. The Guardian. https://www.theguardian.com/technology/2019/jul/26/apple-contractors-regularly-hear-confidential-details-on-siri-recordings.

Higginbotham, S. (2014, July 29). How much data can one smart home generate? About 1 GB a week. Retrieved November 20, 2020, from https://gigaom.com/2014/07/29/how-much-data-to-a-smart-home-generate-about-a-1-gb-a-week/

How Google and Amazon are 'spying' on you. Consumer Watchdog. https://www.consumerwatchdog.org/privacy-technology/how-google-and-amazon-are-spying-you.

Jacobson, A., Gold, J., Hodge, N., & Widmer, L. (2019, September 27). Home. Retrieved November 15, 2020, from http://www.rmmagazine.com/2019/10/01/smart-home-devices-and-privacy-risk/

Katz v. United States. (1967). Oyez. Retrieved November 15, 2020, from https://www.oyez.org/cases/1967/35

Komando, K. (2019, June 20). When smart devices watch you, what do they do with the data? Retrieved November 17, 2020, from https://www.usatoday.com/story/tech/columnist/2019/06/20/what-do-smart-devices-do-data-they-collect-you/1483051001/

Luke 12:3. New International Version.

Maslin, J. (2019, July 22). Smart Home Devices help Amazon and Google to more invade your privacy. Retrieved November 15, 2020, from https://blogs.ischool.berkeley.edu/w231/2019/07/22/smart-home-devices-help-amazon-and-google-to-more-invade-your-privacy/

Nissenbaum, H. (2004). Privacy as Contextual Integrity. Washington Law Review. Retrieved 2020, from https://crypto.stanford.edu/portia/papers/RevnissenbaumDTP31.pdf

Riley v. California. (2014). Oyez. Retrieved November 15, 2020, from https://www.oyez.org/cases/2013/13-132

R.J. Robles, T. Kim. Applications, systems and methods in smart home technology: a review Int. J. Adv. Sci. Technol., 15 (2010), pp. 37-48.

Podracky, J. (2018, December 5). *The Next Phase of Smart Home Tech: Ethical Implications of Google's New Patent*. Data Science W231 Behind the Data Humans and Values. https://blogs.ischool.berkeley.edu/w231/2018/12/04/the-new-patent/.

Smart Home - United States: Statista Market Forecast. Retrieved November 15, 2020, from https://www.statista.com/outlook/279/109/smart-home/united-states

Scalia, A. & Supreme Court Of The United States. (2000) U.S. Reports: Kyllo v. United States, 533 U.S. 27 .

Serena Zheng, Noah Apthorpe, Marshini Chetty, and Nick Feamster. 2018. User Perceptions of Smart Home IoT Privacy. Proc. ACM Hum.-Comput. Interact. 2, CSCW, Article 200 (November 2018), 20 pages. DOI:https://doi.org/10.1145/3274469

Stanford, J. (2017). Memorandum of Law in Support of Amazon's Motion to Quash Search Warrant. Retrieved from https://regmedia.co.uk/2017/02/23/alexa.pdf

Tong, S., & Koller, D. (1998). Support Vector Machine Active Learning with Applications to Text Classification. *Proceedings of the Seventeenth International Conference on Machine Learning (ICML-00).* https://ai.stanford.edu/~koller/Papers/Tong+Koller:ICML00.pdf

Warren, & Brandeis. (1890, December 15). The Right to Privacy. Harvard Law Review. Retrieved November 19, 2020, from http://groups.csail.mit.edu/mac/classes/6.805/articles/privacy/Privacy_brand_warr2.html

Zomet, A., & Urbach, S. R. (2019, October 22). Privacy-aware personalized content for the smart home.