Being Smart About Smart Devices: Preserving Privacy in the Smart Home

by
Simran Shankardass
Sc.B. Brown University, 2024

Advisor: Timothy Edgar
Reader: Kristina Mallory

A thesis submitted in partial fulfillment of the requirements for Honors in the Department of Computer Science at Brown University

Providence, Rhode Island
May 2024

Abstract

Smart home devices are a subset of Internet of Things focused on optimizing the home. Their functioning relies on the collection, storage, and use of a vast amount of user data from within the home. As a result, they raise a large number of privacy concerns, which are inadequately addressed by smart device companies that downplay the severity of these risks or call them ‘inescapable’ in order to prioritize profit; and inadequately addressed by the US government, which has thus far provided no strong, holistic, and comprehensive law or regulation addressing consumer data privacy or smart homes. This paper thus provides a framework for making smart home devices more privacy preserving. This includes an explanation of how they work (using the popular voice assistant, smart thermostat, and smart doorbell as examples); categorization and analysis of the privacy concerns or risks thus created; a discussion of technical methods for improving privacy preservation in smart home devices, to address these concerns; and policy recommendations to incentivize their use. Before providing policy recommendations, I analyse the current policy landscape in the US regarding smart home devices and consumer data privacy. The output is the aforementioned framework that shows the feasibility of improving privacy preservation in smart home devices and describes how to address specific privacy concerns using a necessary combination of technical and policy methods.

Acknowledgements

I spent the better part of my time at Brown saying confidently that I would never do a thesis, it wasn't for me, I just wasn't that kind of girl… It turns out I am that kind of girl, and I would not have been able to become her without the help of many people. Thank you to everyone in general, but also specifically:

My thesis advisor, Timothy Edgar – you helped me write something I'm proud of. Your guidance along the way, not just on smart devices and privacy but on how to approach writing in general, was invaluable. I will always remember not to let the perfect be the enemy of the good.

My reader, Kristina Mallory – for encouraging me and my ideas, and for taking me on so last minute!

Professor Julia Netter – TAing your Responsible CS class and witnessing your passion for it while writing a thesis on data privacy has been uniquely motivating, especially during TA meetings when someone says "privacy" or "thermostat" and everyone looks at me like it's my domain. I wouldn't have known what I wanted to write my thesis on without CS 1952B. Though I never got to take the class, I was looking through its website wishing I could have when I had my lightbulb moment, so CS 1952B is where it all began.

Actually, it began when Grace decided to do a thesis, which gave me FOMO. Then Sophie decided to, which gave me more FOMO. (The F in 'friends' stands for FOMO.) Sophie and Grace, my best friends and biggest supporters at Brown – you give me maybe more confidence in myself than I should be allowed to have. I hope you're proud of me.

Cecilia – we did it! ☺ Thanks for joining me at the dining table every night to work on our respective theses, taking me on mental health walks to Wickenden, and attempting to breakdance on the living room floor to procrastinate work. Having you as a thesis buddy meant I never felt like I was doing it alone.

The Brown'sTones, Brown's freshest, flyyest, fiercest a cappella group – thank you for listening to my thesis updates three nights a week for two whole semesters. You're all the best stress relief I could ask for. Quinn – if you didn't already know, the countless Spicy Withs you've got me from Jo's after rehearsal have fueled so much of my late night work.

And of course, my family, who allowed me for a year to use the word 'thesis' as an excuse to get out of anything I didn't want to do, even when they knew I was really doing other things that weren't thesis. Mama – thank you for (mostly) always picking up my calls when I wanted to chat during my breaks, and also for telling me once when I was stuck, "I'm pretty sure I could have finished writing this already." Papa – thank you for being really, sincerely interested in what I was writing. It was amazing watching you put Amazon Echos in every room of the house at the same time. To my twin sister Saira – your reaction when I told you how many words I had written made it sink in for the first time that I had written a thesis. Thank you for making me feel like the smartest and most impressive person in the world. I needed it ☺

Finally, it is everything I have learned at Brown University that allowed me to complete this. I love this school and I have loved my time here more than I could ever explain. I feel so lucky to have been able to call it home.

P.S. Amazon, Google, etc., if you're listening (which you probably are), thanks for the privacy violations! They made a great thesis topic.

Contents

1. Introduction 6

2. Voice assistants 9

2.1. Types of virtual assistants 9
2.2. What makes a voice assistant smart? 10
2.3. Architecture 13
2.4. Voice assistants and privacy 14

3. Smart thermostats 15

3.1. Types of thermostats 15
3.2. What makes a smart thermostat smart? 16
3.3. Architecture 18
3.4. Smart thermostats and privacy 19

4. Smart doorbells 20

4.1. Types of smart doorbells 20
4.2. What makes a smart doorbell smart? 20
4.3. Architecture 23
4.4. Smart doorbells and privacy 24

5. Privacy concerns 25

5.1. Data collected 25
5.1.1. By voice assistants 25
5.1.2. By smart thermostats 26
5.1.3. By smart doorbells 27
5.2. Privacy concerns 27
5.2.1. Invasive inferences 28
5.2.2. Improper data access and use 29
5.2.2.1. Smart device companies 29
5.2.2.2. Third parties and advertising 29
5.2.2.3. Governments and surveillance 30
5.2.3. Security weaknesses 31
5.2.3.1. Database breaches 31
5.2.3.2. Device breaches 31
5.2.4. Violation of user rights 32
5.2.4.1. Data collection without user consent 32
5.2.4.2. User control over collected data 32

6. Technical methods of privacy preservation 33

6.1. Reducing data collection 33
6.1.1. Data minimization 33
6.1.2. Synthetic data 33
6.1.3. Federated learning 34

6.2. Improving data security **34**
6.2.1. Encryption 34
6.2.1.1. End-to-end encryption 34
6.2.1.2. Encryption in use 36
6.2.1.2.1. Memory enclave encryption 36
6.2.1.2.2. Homomorphic encryption 36
6.2.2. Anonymization and pseudonymization 37
6.2.3. Differential privacy 38
6.2.4. Cloud versus local storage 39
6.2.4.1. On-device storage and processing 39
6.2.4.2. Using a combination of databases, servers, or caches 40
6.3. Improving device security **41**
6.3.1. Offline processing 41

7. Policy landscape and recommendations **42**

7.1. Current US policy landscape **42**
7.1.1. Sector-specific federal laws 42
7.1.2. Federal Trade Commission 46
7.1.3. State laws 47
7.1.4. Proposed regulation 49
7.2. Recommendations **50**
7.2.1. Regulations 50
7.2.1.1. Data collection 50
7.2.1.2. Data access and use 51
7.2.1.3. Data security 52
7.2.1.4. Device security 53
7.2.1.5. User rights 53
7.2.2. Enforcement 54
7.2.2.1. Severity of penalties 54
7.2.2.2. Who can enforce laws? 54

8. Conclusion **55**

Works Cited **57**

1. Introduction

"We tell ourselves this story that our home is the thing that we can control – it's private, it's protected, it's our space." (Kode)

As of May 2023, over 500 million Alexa-enabled devices had been sold worldwide (Garfinkle). In 2024, in the US alone, almost 70 million households actively use smart home devices, and studies estimate that by 2028, this will increase to over 100 million ("US Smart Home Statistics"). Experts expect that, globally, there will be 22 billion connected smart devices by 2025 (Oracle). The Internet Age has given way to the Internet of Things Age; there is an Alexa in every room of my house.

The Internet of Things (IoT) is a network of devices that integrate the everyday with the internet – physical objects that use cutting-edge technologies in order to connect to and communicate with other devices and systems (Oracle). IoT devices are being adopted rapidly around the world, made economically attainable and technically feasible by coinciding technological innovations like vastly improved and miniaturized sensors and batteries, cheap and compact computer processing and data storage, low-cost wireless connectivity, big data analytics, and more (Porter and Heppelmann). IoT medical devices allow remote, real-time patient monitoring without the effort and cost of a doctor's visit (Meola); industrial IoT like smart manufacturing improves productivity in factories and enables a seamless supply chain in what is being called Industry 4.0, the fourth Industrial Revolution (Harsh); entire cities are looking to become 'smart' by interconnecting their energy, logistics, and communication grids (Mason). Many applications seem endlessly important and grand. Yet one of the most significant ones lies a little closer to home – more specifically, within it.

Smart home devices are consumer facing IoT that promise to make life easier, more convenient, and more seamless. The smart home is equipped with gadgets that communicate with each other and with the outside world so that its owner can run it from anywhere, and, more importantly, so that it can run itself. The smart bed detects when its occupant rises at 8 AM and notifies the connected smart coffee machine, which starts making a cappuccino with the last of the milk; the smart fridge adds milk to the shopping list that it sends in the online grocery order every week; by the time the coffee is ready, the smart thermostat has adjusted the kitchen temperature so that it is warm enough to make breakfast in, knowing that the homeowner comes into the kitchen around 8:15 each morning. Easy, convenient, seamless.

But the sentiment that "smart people avoid living in 'smart' homes" is echoed by many who hesitate to connect their residences to the Internet and thus to the world at large (Magee). The benefits provided by devices like smart beds, fridges, and thermostats depend almost entirely on the collection and use of a vast amount of data that comes from directly inside the individual's home, and which can therefore be extremely sensitive. For example, in 2018, Google Home (Google's artificial intelligence speaker) and Chromecast (its streaming device) were found to reveal a user's physical location to within 10 meters (Burgess, "Google Home's data leak"). The security camera company Ring, which provides video and Wi-Fi-connected home security cameras and doorbells, was charged by the Federal Trade Commission (FTC) with violating user privacy by allowing employees unrestricted access to video recordings taken by Ring products, including thousands of videos in "intimate spaces" like bathrooms and bedrooms (Federal Trade Commission, "FTC Says Ring Employees Illegally Surveilled"). There is no question that when it comes to using smart home devices, the violation or resulting lack of

privacy is a growing concern. Yet IoT companies repeatedly try to convince people that it is not, or when they cannot, that giving up this privacy is the only way to enjoy the benefits of IoT. There is "no single federal legal framework in the US that limits what data is collected from your smart home devices or how it is used" (Tuohy). Current research agrees that privacy threats exist but does not present a holistic, comprehensive way to tackle them; discussions and proposals within government and policy circles and by groups like the FTC are plenty, but there have been no clear or well-defined outcomes to ensure that smart home devices preserve privacy. Given their increasing adoption rates and the nature of the data they collect, such clarity is urgently needed.

The intention of this thesis is to show how smart home devices can be made more privacy-preserving by providing a comprehensive framework for the same. It addresses both technical and policy considerations, neither of which are sufficient on their own, and both of which technologists and non-technologists must have a sound understanding of in order to better protect privacy. In sections 2, 3, and 4, I use the smart voice assistant, smart thermostat, and smart doorbell as examples to explain the way that smart home devices actually work, providing an overview of the key features that make them 'smart' and their architecture. In section 5, I establish the extent of the privacy concerns by categorizing and detailing key privacy risks associated with the devices and their potential impacts; I categorize these threats such that I can provide technical and policy recommendations that directly correspond to them. In section 6, I discuss technical methods for privacy preservation which directly correspond to the threat categories. Finally, in section 7, I assess the current US legal and political landscape regarding consumer privacy and give policy recommendations to be used alongside the technical recommendations. I focus on the US because it is lagging in privacy protection, and to limit the scope of the paper.

I chose voice assistants, smart thermostats, and smart doorbells as examples because, in addition to being amongst the most popular smart home devices, they each raise a distinct privacy concern due to their specific features. Voice assistants listen for and understand natural language commands given by a user in order to complete tasks, and in doing so collect voice recordings from within the user's home. Smart thermostats allow a user to adjust their home's temperature remotely using a computer or phone application and can make fine-grained adjustments to the temperature on their own. To achieve this, they collect data on home occupants' movements within the home and temperature preferences – this can be surprisingly sensitive. Smart doorbells notify owners when a visitor is at the door and use video cameras to show owners who it is. Therefore these collect video data from the vicinity of the user's home. All of these devices have the potential to drastically change users' lives (compared to, for example, smart coffee machines). Voice assistants streamline day-to-day tasks, work as a central point of management for other devices, and even offer accessibility for people who cannot use traditional technology by allowing them to use their voices; smart thermostats help individuals consume energy responsibly, decreasing their power costs and carbon footprints in light of the looming environmental crisis; and smart doorbells improve home safety by allowing users to check who is outside their door before opening it, as well as by alerting them to suspicious or abnormal events outside their homes. It's therefore harder to say that the privacy risks they create outweigh the benefits, making them good examples of smart home devices with which to create such a framework. I intend for the framework to be applicable to smart home devices in general.

Finally, I use the terms IoT, smart devices, and smart home devices more or less interchangeably moving forward to refer to smart home devices. I also use users and consumers, interchangeably to mean the people using smart home devices.

2. Voice assistants

The origins of voice assistants go back much further than the introduction of Siri and Alexa to the beginnings of speech recognition technology. The first voice-activated toy (Radio Rex, a wooden dog that emerged from its "house" when its name was called) was released in 1922 (Markowitz). Bell Labs created Audrey, an Automatic Digit Recognition machine, in 1952 (Moskvitch); IBM's Shoebox calculator, launched in 1961, used primitive digital speech recognition to recognize 16 words and 10 digits (Mutchler); at MIT in the 1960s, Joseph Weizenbaum developed ELIZA, the first Natural Language Processing (NLP) computer program or first chatbot, which used pattern matching and substitution to simulate conversation and "understanding"(Ireland; Weizenbaum); and Carnegie Mellon completed Harpy, which could recognize 1000 words and understand sentences, in the 1970s (Mutchler).

Strides towards virtual assistants using voice technology were made after the 1980s, when Hidden Markov Models (HMMs) were developed and used not just to search for sound patterns but, based on past observations, predict the likelihood that unknown sounds were words ("Short History of Speech Recognition"). With much more advanced speech recognition and NLP made possible, Apple launched its virtual assistant, Siri, in 2011, and the "era of voice assistants" followed with Amazon Alexa, Google Assistant, Microsoft Cortana, and more (Mutchler). In 2015, Amazon Echo hit the market – the first smart speaker, a physical audio device equipped with Alexa to be used in the home. Now, virtual assistants, specifically voice-enabled ones, have become commonplace and show no signs of dying out. In 2024, there are at least 4.2 billion digital voice assistants being used worldwide (Thormundsson); some predict that by the end of 2024, there will be more voice assistants than humans (Stelitano).

2.1. Types of virtual assistants

A virtual or digital assistant is a software agent or application that completes tasks for a user based on its understanding of user input (Yasar). It is important that we distinguish between types of virtual assistants to understand what kind is used in a smart speaker:

- *Traditional chatbots*: Chatbots use text interfaces to simulate conversation with users (Yasar). Using NLP and keyword recognition, chatbots attempt to understand a user's question and provide a response (Owen). Importantly, traditional chatbots are rule-based – that is, they answer questions based on predefined conversational trees (O'Neill). Think customer service chatbots providing scripted answers or suggestions to FAQs and common follow-up questions (this is a common use for chatbots). If the user continues to ask more specific questions, the bot eventually does not have a programmed answer and connects the user to a human customer service agent; it cannot adjust its answer based on context and does not learn from conversations to improve its answers.

- *Conversational AI chatbot*: Conversational AI chatbots simulate more dynamic, and human-like conversations with users in order to help them complete tasks (Yasar). Using NLP, machine learning (ML), and predictive analytics, they process inputs to understand user intent and learn from prior conversations to recognize patterns and provide more intelligent, context-specific recommendations (Ahmed). These might use text or voice interfaces, or both.

- *Voice assistants* or *voice-activated assistants*: Voice assistants also use artificial intelligence (AI) to understand user input and complete tasks, and continuously learn in order to improve interactions. They are distinguished from other types of virtual assistants by their ability to "converse" with users through smart speakers, using voice recognition, speech synthesis, and NLP to listen for and respond to voice commands when they are made (O'Neill; Ramos). Unlike chatbots, if voice assistants do not understand a question, they continue to converse with the user (ask additional questions, make clarifications, and so on) until they can come up with something. Voice assistants like Siri and Alexa are commonly used to find out the weather, play music, send emails and text manages, manage calendars, and more.

There are other ways to segment virtual assistants – for example, task-based (focusing on specific work, such as email management) versus predictive (predicting user needs based on historical data and offering help unprompted). However, for the purposes of this thesis, the distinction between text-based and voice-activated virtual assistants is more significant. Some virtual assistants might accept both text and voice inputs, but I focus on those that primarily accept voice inputs.

Furthermore, while voice assistants can be integrated into phones, laptops, and more, I use 'voice assistants' from here on to refer to smart speaker devices that are equipped with the assistant technology – for example, the Alexa-enabled Echo speaker. I also use 'smart speaker' interchangeably.

2.2. What makes a voice assistant smart?

Key features and technologies

A Voice User Interface (VUI) enables voice assistants to realistically "converse" with users – listen to their commands, understand them, and respond appropriately (Ramotion). VUIs use the following:

- *Wake word detection*: "Wake words" are words or phrases that activate a voice assistant – for example, "Alexa", "Hey Siri", and "Okay Google". Audio in the vicinity of the speaker, such as people talking, cars honking, TV shows, and other background noise, is processed (cleaned and made sense of using *signal processing*) until the speaker detects the wake word (Gonfalonieri; "How Alexa works"). Once it is detected, the voice assistant turns on, records the audio that follows, and sends it to the cloud or server where it is converted to text using speech recognition software (Gonfalonieri). Wake word detection is important because it prevents the voice assistant from completing tasks that the user has not asked it to do, based on something the user has said to someone else in the room; the voice assistant only begins listening for commands after it detects the wake word.

- *Automatic Speech Recognition (ASR)*: Simply put, ASR is *speech-to-text (STT)* technology that enables voice assistants to recognize human voice. ASR algorithms analyze the speech recording and translate spoken commands into textual representations ("What is an AI voice assistant?"). A combination of two trained models is used to

determine exactly what is being said. The acoustic model, trained from speech databases, is used to break down the speech recording into small recognizable phonemes, which are distinct sounds or groups of sounds in the English language (Garbar; Jiřík) – for example, phonemes that sound like "tell", "me", "the", and "whether". The linguistic model then predicts actual sentences that might have been said (Garbar). "Tell me the whether" and "Tell me the weather" sound the same, but only one is a meaningful sentence – it is the job of the linguistic model to help the voice assistant identify it. Additionally, ASR distinguishes between accents, pitches, tones, and other aspects of a user's voice in order to analyze who is speaking, what the context is, and therefore what the command might be (Jiřík).

- *Natural Language Processing (NLP)*: NLP uses artificial intelligence and computational linguistics to enable machines to "analyze, understand, alter, or generate natural language", or the way that humans talk to each other, and thus communicate with humans in forms including speech and text (Gonfalonieri). After ASR converts speech to text, NLP algorithms (specifically *Natural Language Understanding (NLU)*, a subset of NLP) are used to understand the intent behind the command by analyzing sentence structure, keywords, grammar, context, and more, transforming the natural language command into something machine-readable ("What is an AI voice assistant?"). Identifying keywords is of particular importance because it allows the voice assistant to carry out actions relevant to specific tasks. For example, if a user tells Alexa to "set the thermostat to cool", Alexa must know to engage the smart thermostat API that controls the thermostat (Fingas); if a user tells Alexa to play a song, Alexa must know to open the music application. Finally, *Natural Language Generation (NLG)* outputs a natural language response for the voice assistant to give to the user. The NLG response is typically textual and must be converted to speech (Gatt and Krahmer).

- *Internet and cloud connection:* Once the voice assistant understands the command or question, it accesses an external knowledge base comprising cloud databases, APIs, the web, and more to carry out actions and find information. This is done with the help of the dialog manager, which accepts the machine-readable input from NLP, interacts with the external resources, and produces a machine-readable response for NLG ("Dialogue Manager").

- *Speech synthesis:* This is *text-to-speech (TTS)* technology that converts the textual response to speech with which the voice assistant will "answer" (Stelitano). In other words, it reads aloud the text.

Additionally, smart speakers have the following key features or abilities:

- *Learning:* A key feature of any smart device is continuous learning. Learning from past interactions, recorded audio, and other historical data helps voice assistants understand content and intent better and provide more accurate and personalized responses. For example, over time, a user's Alexa might learn that when they say "Queen", they are usually referring to the band led by Freddie Mercury and not the Queen of England.

- *Integration with other IoT devices:* Smart speakers can be integrated or connected with other IoT devices, specifically smart home devices, in order to create seamless experiences for the user. For example, a user might connect their smart speaker to their smart thermostat and thus change the temperature by giving voice commands, or tell the smart speaker to unlock the door, which it will do via its connection to the smart lock.

Figure 1 below depicts the flow of data and the use of different technologies in order to process the input and produce a response.

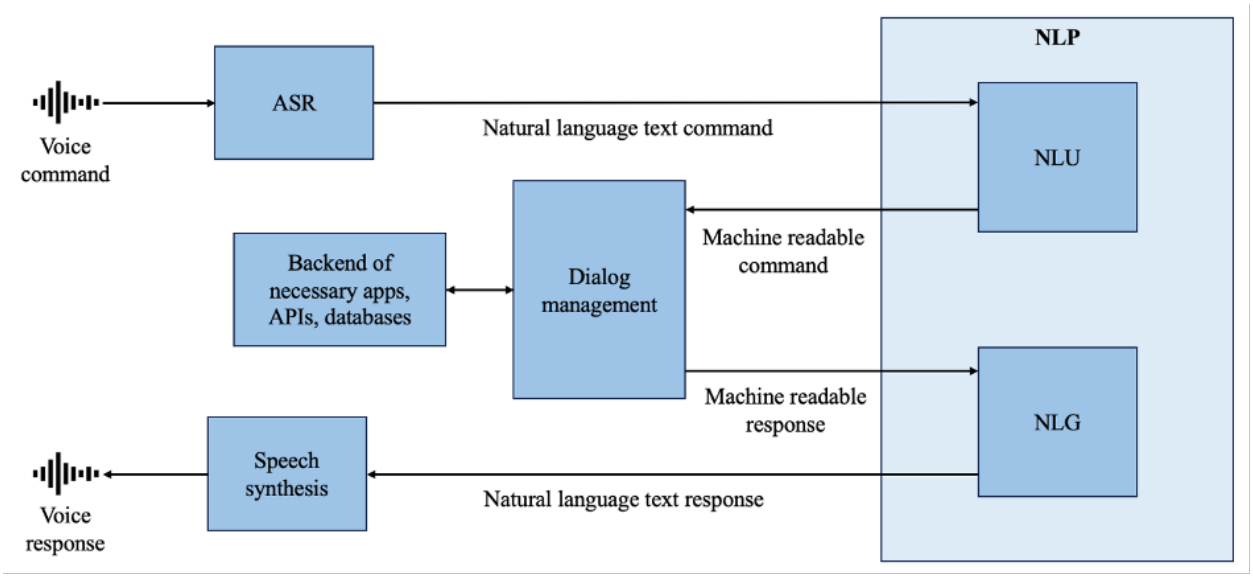

Figure 1: Data flow and technologies used by voice assistants. Arrows indicate the direction of data flow.

2.3. Architecture

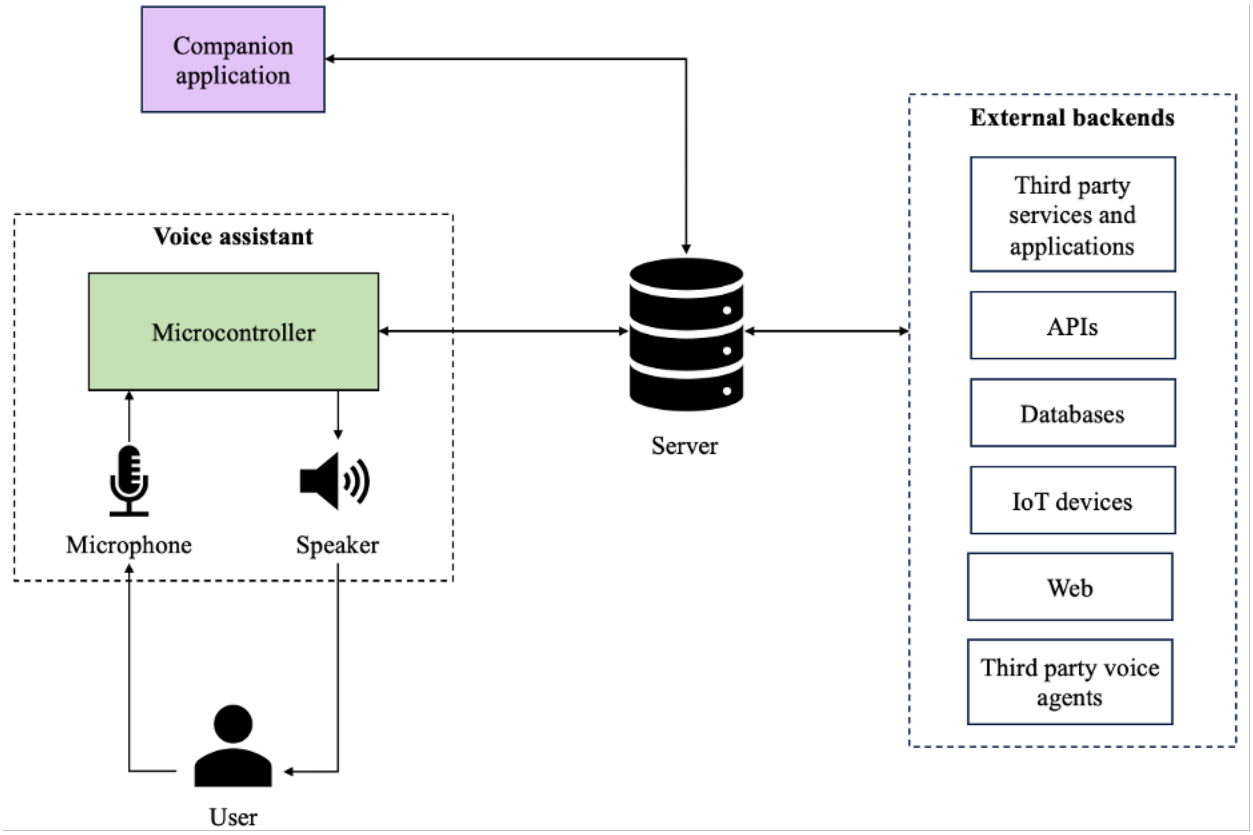

Figure 2: Architecture diagram of a voice assistant. Arrows indicate the direction of data flow.

As seen in Figure 2, the smart speaker device itself consists of a *microcontroller* (which contains the device's local memory and a *Central Processing Unit (CPU)* that processes data and executes program instructions), along with a built-in *speaker* and *microphone*. The *user interface* typically comprises this built-in *speaker* and *microphone*, though more advanced voice assistants might also have a *screen* display. When the user speaks to the voice assistant, the microphone records the sound, which is processed and cleaned locally. The recording is then sent to a *remote* or *cloud server* where ASR and NLP are used to translate it to text and understand it; the recording is also stored in the cloud. (A local server would likely not have the computing power to support ASR and NLP technologies). The server then activates software extensions and accesses databases, APIs, and the web as necessary – for example, opening specific applications, checking the news or the weather, or controlling connected smart home devices – in order to fulfil the action identified from the user command (Thormundsson). NLP and speech synthesis are used to generate a spoken natural language response, which is sent back to the local device. The speaker plays this to the user.

It is important to note that when it comes to voice assistants, most of the work occurs in cloud servers, not locally. The device has Wi-Fi, Bluetooth, and local storage; however, interpreting sounds with machine learning algorithms is computationally heavy, so it typically done on remote servers (Gonfalonieri). The response is then sent back to the local device.

A voice assistant may also have an additional user interface in the form of a *companion mobile or web application*. This is used for set-up and device support, but the user can often use

this to communicate with the voice assistant as well. For example, a smartphone can be connected to Amazon Alexa to play music, and in this case the volume can be adjusted by speaking to Alexa or by pressing the volume buttons on the phone. Additionally, a companion app might be able to display responses sent by the server (Bolton et al.).

2.4. Voice assistants and privacy

Since voice assistants work by understanding humans' spoken commands, they are always listening – first for their wake word, and then for the complex sentences that follow it. As a result, the most significant privacy concern relating to voice assistants is how much they hear and understand when users are not speaking to them, as well as what they do with the recordings they take and who has access to them. Additionally, does access to the smart speaker mean someone can spy on users within their homes? There are concerns around potential espionage if hackers access peoples' voice assistants, as well as the potential for hackers to control other devices and interfere with users' lives in other ways if their voice assistants are IoT-integrated or connected to banking, shopping, health, or subscription applications.

I will discuss these privacy concerns in more detail in section 5.

3. Smart thermostats

In 1620, Cornelis Drebbel invented a mercury thermostat to regulate the temperature of a chicken incubator (Tierie). Two centuries later, in the 1830s, Andrew Ure invented modern thermostatic control with the bimetallic thermostat, which contained metal that would expand as temperatures increased and cut off the energy supply to bring them back down ("Early History of Comfort Heating"). Then, in 1886, Albert Butz patented the analogic electric thermostat, setting the stage for the digital electric thermostats that are widespread today ("Thermostat").

3.1. Types of thermostats

A thermostat is a temperature-regulating device used to set the temperature in a room or set of rooms. Thermostats commonly seen in homes today include:

- *Basic thermostats*: These allow the user to set the temperature for a room or system. They detect when the temperature in a room deviates from the setpoint temperature and direct the heating, ventilation, and cooling (HVAC) system, which includes heaters and air conditioners, to heat or cool the room accordingly to return to the desired temperature.

- *Programmable thermostats*: These allow users to set not just a current temperature but a temperature schedule for the HVAC system to follow. Programmable thermostats control HVAC systems based on a setback schedule, where the room is kept at a setpoint temperature when occupants are active within it and a lower, energy-efficient setback temperature at other times (e.g., when the occupants are out of the room, out of the house, or asleep) (Lu et al. 212). However, these thermostats do not react to occupancy – if the user sets a schedule saying to lower the temperature between 9 am and 5 pm, when they expect to be out of the house, and then comes home at 3 pm, the thermostat will not increase the temperature unless the user manually resets it. In other words, the setback schedule programmed into the thermostat is static and cannot adapt to dynamic occupancy patterns (Lu et al. 213).

- *Wi-Fi-enabled thermostats*: These thermostats are connected to Wi-Fi; users can connect to them remotely through a user interface like a mobile application, and adjust the temperature and other thermostat settings from any Wi-Fi-connected location.

- *Reactive thermostats*: Unlike programmable thermostats, reactive thermostats are occupancy-detecting. They use various kinds of sensors (e.g., motion, door, key card access) to control the HVAC system based on whether the house is occupied or not (Lu et al. 213). While they can save energy when used, the amount of energy actually saved is "limited by their inability to respond quickly" to changes in occupancy (Lu et al. 213).

- *Smart thermostats*: In addition to the functionality listed above, smart thermostats act on their own to heat and cool different parts of the house in the most energy-efficient way. Not only are they occupancy-detecting, meaning that they adjust temperatures based on whether a room is occupied or not (for example, they might override a set schedule if a room becomes empty), but *learning* thermostats. This means that they use data about

residents' occupancy patterns and other behaviour within the home to create a model that is used to predict what temperature to heat a room to at a given time.

3.2. What makes a smart thermostat smart?

Key features and technologies

Smart thermostats make faster, more precise, and more intentional adjustments to a home's HVAC system compared to manual and programmable thermostats, allowing for higher energy savings without sacrificing comfort. In order to do so, they use the following:

- *Motion detection*: Smart thermostats use *sensor technology* to detect environmental variables such as temperature, humidity, light, carbon dioxide level, occupancy, sound, and more – data which guides decisions about how to adjust the temperature in order to optimize energy usage and comfort within the home. Other than temperature sensors, which are necessary for all kinds of thermostats so that they can tell whether or not the room is at the desired temperature, smart thermostats commonly use *motion sensors*. Motion detection provides smart thermostats with real-time occupancy data, which allow them to adjust the temperature to a comfortable level when the house is occupied but prioritize energy saving when it is empty ("Motion Sensor Thermostats"). Motion sensors can also help smart thermostats make more fine-grained decisions about heating and cooling. Remote sensors can be used in different rooms, which might run warm or cold due to insulation, natural light, or size differences, so thermostats know to adjust the temperature in the specific room the user is in rather than the house in general. Furthermore, users can configure sensors such that they only detect certain types of motion or motion in specific areas (Vigderman and Turner); a motion sensor might be configured so that it only detects motion when someone comes a certain distance into a room, thus only spurring the smart thermostat to count a room as occupied when a person is fully inside it and not just in the doorway or passing by.

  In section 3.3, I discuss the hardware of a smart thermostat; there, I will touch on the different kinds of motion sensors.

- *Deep setbacks*: A setback is a temperature that the thermostat allows the house to drift to while it is unoccupied in order to save energy (e.g., by not cooling an empty house on a hot day or not heating an empty house on a cold day). A deep setback allows greater energy savings by enabling smart thermostats to let the temperature drift far from the setpoint temperature when they predict that the occupant will not return soon (Lu et al. 213).

- *Smart scheduling*: Smart scheduling is used to start adjusting the temperature at the best possible time after a setback. Smart thermostats employ *preheating*/*precooling systems* that use current sensor data, historical occupancy data, and a categorization of the efficiency and capacity of the HVAC system to decide when to begin heating the home (for example, to preheat it or begin heating it only once the occupants arrive) (Lu et al. 212). This involves predicting both when occupants will return and how long it will take for the house to reach the desired temperature. The optimal preheat/precool time will minimize occupant discomfort at home along with energy usage. Based on occupants'

predicted return, the thermostat will slowly preheat/precool the house with higher efficiency but lower capacity; if occupants arrive earlier than predicted, the thermostat can direct the HVAC system to heat the house at a higher capacity, forgoing efficiency, to quickly make up the difference (Lu et al. 214). For example, the Google Nest smart thermostat has Time-to-Temp and Early-On features which are used for smart scheduling (Baterna).

- *Geofencing*: Geofencing is setting an invisible perimeter around the home that determines when a smart thermostat user is "home" and when they're "away" (Grant). This is distinct from in-house motion detection; rather than detecting motion using a sensor, geofencing uses GPS signals to determine where a user is with respect to the geofence. When a user crosses the geofence, the thermostat switches from energy- and cost-saving mode (e.g., using a deep setback) to user comfort mode (e.g., turning up the heat on a cold or turning up the air conditioning on a hot day). This is useful because the thermostat can adjust the temperature before the user actually arrives, minimizing the time that they spend in a too hot or cold house. Of course, geofencing requires the use of Wi-Fi or mobile data and location services, typically on the user's smartphone, whose location is used to detect where the user is with respect to the geofence (Grant). As a result, it is more accurate (with an accuracy of 100 to 200 meters) in urban areas with better cell service (Wixted).

- *Wi-Fi connection and remote access*: Wi-Fi connectivity enables users to access and interact with their smart thermostat remotely, using a phone or web application. They can set and check temperatures from anywhere; some smart thermostats also provide alerts or notifications, insights on energy consumption, and suggestions to change the temperature based on analyzed data or energy use insights.

- *Demand-response*: Remote accessibility enables smart thermostats to be used for *grid* balancing through demand-response programs. Demand-response is "balancing the demand on power grids by encouraging customers to shift electricity demand to times when electricity is more plentiful or other demand is lower" (Bertoli). Utility companies offer incentives to customers to enroll in such programs; in exchange for reduced rates or other incentives, they can remotely access an enrolled customer's thermostat, instructing it to change temperatures before and after "peak events" (times when electricity demand is highest) in order to reduce the load on the energy grid (Rhode Island Energy).

- *Learning* and *data analysis*: While smart thermostats allow users to manually set temperatures and temperature schedules, they also continuously learn in order to be able to control temperature more efficiently and automatically, i.e., keep occupants comfortable while using as little energy as possible (Baterna). Smart thermostats collect data about occupants, including but not limited to users' temperature preferences (at different times of day, in different parts of the house, etc.), when they leave and enter the home, and their movement patterns within it. This data is stored, analyzed, and used to train machine learning models which are then used to predict and set desired temperatures without the user having to manually program the thermostat with them.

- *Integration with other IoT devices:* Smart thermostats can be integrated or connected with other IoT devices, specifically smart home devices. For example, a user might connect their smart thermostat to their smart speaker or smart display and control the temperature through voice commands. As another example, they might also connect their smart thermostat to their smart lighting and use lighting as an indicator of room occupancy.

3.3. Architecture

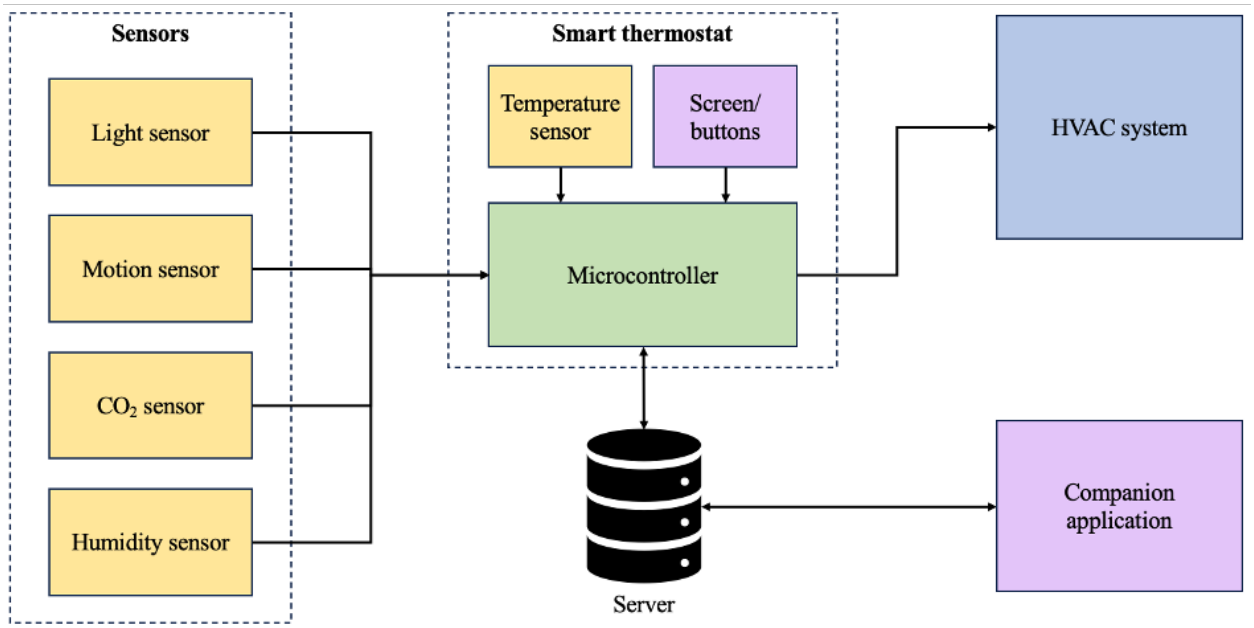

Figure 3: Architecture diagram of a smart thermostat. Arrows indicate the direction of data flow.

Figure 3 shows the architecture of a typical smart thermostat system.

The thermostat itself consists of a *microcontroller*, which contains a *CPU*, the device's local memory, and timers. These enable the thermostat to process data inputs that it receives, as well as instructions from hardware and software programs that must be carried out for the desired actions and functions of the thermostat to take place. Thermostats also usually have a built-in rather than externally connected *temperature sensor*, which tells the thermostat the current temperature so that it knows if it is at the target or if changes need to be made Additionally, a smart thermostat often also has a built-in screen or buttons, and might have a built-in USB port.

There are typically two *user interfaces* through which the user can manually interact with the thermostat. One is the aforementioned *screen* or buttons or dials on the external part of the smart thermostat, which the user can use at the very least to adjust the temperature and, in more advanced models, to complete complex tasks like setting temperature schedules and adjusting settings; the thermostat might also have a *USB port* through which preset schedules can be uploaded. The second is a *mobile or web application* through which the user can remotely access and interact with the thermostat. The user interfaces send user input to the thermostat, and receive and display output (notifications, temperature, heating or cooling status, etc.) to the user.

The thermostat also receives data inputs from one or more external *sensors* – these can include CO_2 sensors, humidity sensors, motion sensors, and/or others, depending on the design of the thermostat. (Figure 3 is not exhaustive.) *Motion sensors*, as mentioned ins section 3.2, send the thermostat data that it uses to determine occupancy. The most commonly used motion sensors are Passive Infrared (PIR) sensors, which detect motion by detecting infrared energy that humans emit as heat (Vigderman and Turner). Microwave sensors are also common – they emit microwaves and record when they are reflected back by objects in the room to figure out the distance of objects, and then record changes in those distances as "motion" (Vigderman and Turner). Dual sensors use both PIR and microwave technology in order to reduce false positives – both sensors must detect motion before it is considered motion (Vigderman and Turner). (Other kinds of motion sensors include ultrasonic, vibration, area reflective PIR, and tomographic sensors, but since they are less popular, I will not discuss them in detail.)

The microcontroller of the smart thermostat interfaces with the *HVAC system* in order to control the temperature. If the room temperature is different than the desired temperature (which might be set by the user using the user interface or predicted by the thermostat using machine learning), then the microcontroller directs the HVAC system to adjust the heating or cooling of the house to arrive at the target.

Finally, the smart thermostat stores collected data in a database or *server*. The server sends back the data to the thermostat as necessary. Similarly, the smart thermostat application sends data to and receives it from the server. The server might be local or remote using a Wi-Fi connection, or a combination of both; I will discuss the trade-offs between these options later on.

3.4. Smart thermostats and privacy

Privacy concerns related to smart thermostats largely stem from the fact that in order to be 'smart', smart thermostats collect and store data from directly within the home. This data may seem innocuous, but the insights that can be drawn from it are extremely sensitive – for example, when a house is empty and when it isn't, when specific rooms of a house are empty, and how a person moves within their home.

I will discuss these privacy concerns in more detail in section 5.

4. Smart doorbells

Smart home security systems allow users to monitor their home security remotely, and usually in real-time. Homeowners can use a mix of various IoT products to achieve this, including but not limited to security cameras (both indoor and outdoor), smart locks, smart lighting, car and house alarms, and smart doorbells. Smart doorbells in particular are among the most popular smart home devices (Molla). The global smart doorbell market is estimated to grow from 3.2 billion US dollars in 2023 to 13.7 billion US dollars in 2033 ("Global Smart Doorbell Market Size"); in 2024, smart doorbell sales are estimated to grow by 27 percent (Molla).

4.1. Types of smart doorbells

A smart doorbell, or video doorbell, is a doorbell which notifies a homeowner when someone is at their door. The doorbell is activated when pressed by the visitor, or may detect when a visitor arrives at the door and then chime. There are two main types of smart doorbells:

- *Wired smart doorbells*: These connect to the home's electrical system wiring, just like a traditional doorbell ("How Does a Doorbell Work?"). As a result, they are provided with a continuous power supply and don't require batteries, which need to be replaced often – unless the electrical system fails, the doorbell will be operational for a long time (Bitner).

- *Wireless smart doorbells*: These are battery-powered and don't need wiring (Bitner). As a result, they can be placed anywhere outside the door that the owner wants without any need to take the home's wiring configuration into account – this is useful because users can place the video doorbell such that it captures the view they desire. However, wireless smart doorbells need frequent battery replacements or recharging; when the batteries die, the doorbells are "inoperable" (Bitner).

For the purposes of this thesis, I will not distinguish between wired and wireless smart doorbells; the features and functionality I am focusing on are common to both types.

4.2. What makes a smart doorbell smart?

Key features and technologies

Smart doorbells provide enhanced security to homes with the following features, used in conjunction with each other:

- *Live video and video recording*: Smart doorbells are equipped with high-definition cameras that show who or what is at the door when activated, enabling users to learn who the visitor is without having to open the door. Cameras are activated either when someone rings the doorbell or when the system detects someone approaching, depending on how advanced the system is (I discuss motion detection next in this section). They provide both live feeds and record videos, both of which can be viewed using the doorbell companion application that users can access on their phones. Ring doorbells allow users to access a live feed at any time they wish through the app, regardless of whether the doorbell is rung or a visitor is detected (Trethewey). Smart doorbells might

also be programmable such that they record videos at specific times chosen by the user, such as at night (*night vision* allows for clear videos to be taken after dark) or when the user is on vacation (de Looper).

- *Motion detection*: Motion sensors enable smart doorbells to detect when someone or something approaches the door, even if the doorbell is not pressed. When the sensors detect motion, the doorbell cameras can begin recording and send alerts to the user on the doorbell companion application. For more advanced alerts, Ring allows users to set "motion zones", which are zones of interest to the homeowner (de Looper). Users might want their doorbell camera to begin recording or send an alert only when motion is detected within a certain distance from their house or in a certain location – for example, within the yard or driveway but not on the sidewalk.

  As discussed in section 3.3, PIR sensors are the most popular motion detectors, followed by microwave and dual sensors. Outdoors, PIR sensors are more error-prone, due to wind, bugs and animals, passing cars, and other events that are easy to falsely detect as motion that the user needs to be alerted to (Iyer). A "multimodal sensing approach" – using PIR sensors along with temperature, light, and/or humidity sensors – can reduce false positives (Iyer).

- *Two-way audio*: Advanced smart doorbells have built-in speakers and microphones through which users can communicate with visitors in real-time; visitors can communicate back, making the audio two-way. This is useful to speak with mailmen, delivery people, and unknown visitors without having to open the door; additionally, if a known visitor arrives and no one is home, the user can communicate this to them through the speaker.

- *Cloud storage*: Camera recordings (and audio) must be stored in order to be accessed at later times. As a result, many smart doorbells provide cloud storage; recordings are continually uploaded to the cloud (using a Wi-Fi connection) and can be downloaded when necessary (Trethewey). This is a crucial feature, as it enables users to review past events at their doors whenever they desire and provide a database to examine in cases of theft, harassment, stalking, and more. Ring even provides a feature as part of the Protect subscription which allows users to review an "event history timeline" of footage (de Looper).

- *Wi-Fi connection and remote access:* Wi-Fi connectivity enables users to access and interact with their smart doorbell remotely, using the doorbell's companion phone or web application. They can view live feeds and videos recorded from their front door no matter where they are; receive alerts (in the form of messages, photos, and suggestions to view the live feed) and when there is a visitor; and use the phone's microphone and speaker to communicate with them. Wi-Fi connectivity also enables the use of the cloud.

- *Video analytics*: For smart doorbells, the data recorded and learned from is primarily photographic and videographic. Video analytics is the "process of monitoring video streams in real time" and learning from them in order to identify trends and patterns in the observed environment (Thakkar and Ukani, "IoT-Based Smart Doorbell" 221). It can

be used for differentiating between people, animals, and objects like delivery packages; anomaly and "unusual motion" detection (detecting abnormal or suspicious activities near the front door, such as intruders); and sending alerts when specific activities occur that the user cares about (Crosling). Advanced smart doorbells perform *facial recognition* and even *liveness detection*; in conjunction with video analytics, these technologies help smart doorbells provide users with more intelligent notifications and less false alarms (Crosling).

Facial recognition involves detecting a face and analyzing the features to determine if it is familiar. First, the face is detected and photographed or recorded, often using a deep neural network model (Thakkar and Ukani, "Proficient and Economical Approach" 70). Then a trained deep learning model is used for liveness detection – distinguishing between real faces and "fake" faces, for example, a face printed on a mask or displayed on a screen. This can be active or passive; active liveness detection requires the visitor to perform some activity or somehow engage with the doorbell system, while passive liveness detection uses deep learning algorithms to detect liveness based on characteristics of the scene recorded (Thakkar and Ukani, "Proficient and Economical Approach" 73). If the face is fake, then the user might be notified of a spoofing attempt. If it is real, the photograph or recording is passed to the facial recognition model. This works by asking users to 'tag' people who are recorded at the door. The model is pre-trained and continues to learn from the data passed to it. As people approach it in the future, the user can be told not only that someone is at the door but whether it is a stranger or familiar face, and, if familiar, who exactly it is (Gibbs).

In general, smart home devices continuously learn. Learning from recorded videos, motion detection events, and other historical data helps smart doorbells to respond more accurately over time.

- *Integration with other IoT devices*: Smart doorbells can be integrated or connected with other IoT devices, specifically smart home devices, in order to streamline more actions for the user. For example, a user might connect their smart doorbell to their smart lock and thus unlock the door remotely upon receiving a notification from their smart doorbell and checking who is at the door; they might even have the door unlock automatically if the smart doorbell's facial recognition identifies the visitor. They might also use their smart doorbell to give voice commands to the smart lock while they are at home rather than using the companion application.

4.3. Architecture

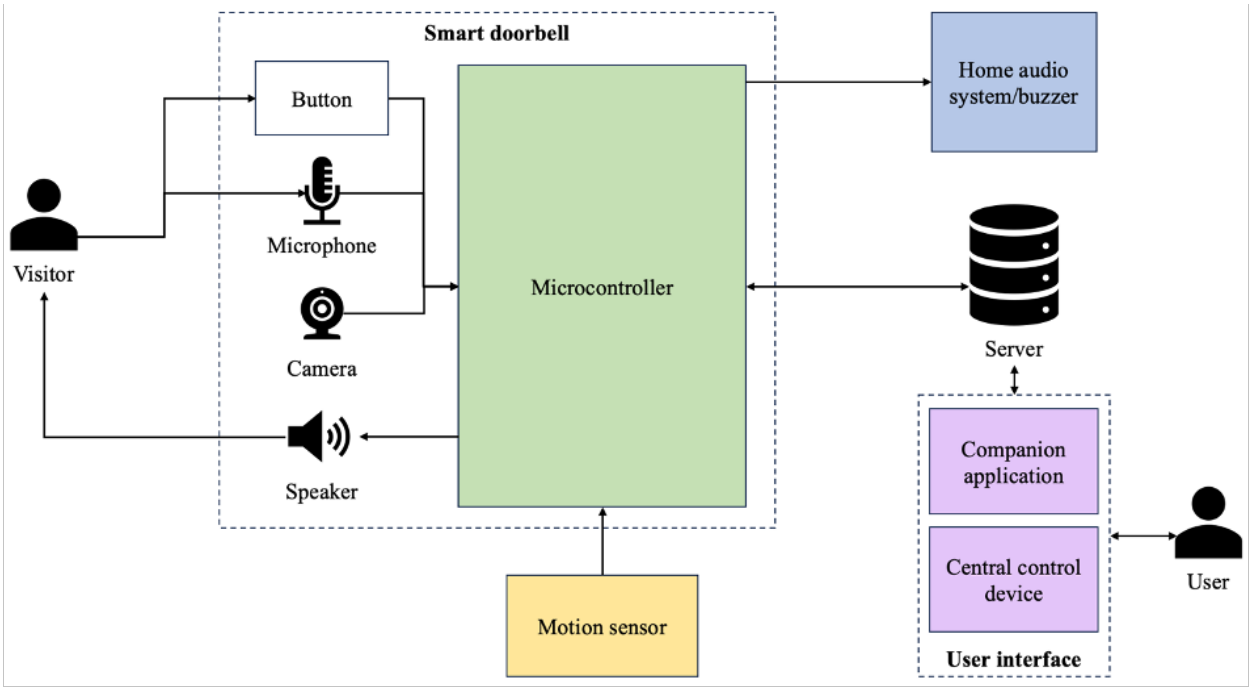

Figure 4: Architecture of a smart doorbell. Arrows indicate direction of data flow.

Figure 4 shows the architecture of a typical smart doorbell system.

The smart doorbell itself consists of a *microcontroller* with the device's local memory and a *CPU* that processes data inputs and executes software and hardware program instructions. It also contains the physical doorbell *button* for the visitor to press. Additionally, the doorbell has a built-in *speaker* and *microphone* for audio and a built-in *camera* for photos and video recording. When the button is pressed or *motion sensor* detects motion, this data is sent to the smart doorbell, which activates the camera to begin recording the event. When the button is pressed, this also causes the *buzzer* to make the doorbell sound that alert the homeowners.

Photos and video recordings, along with their metadata, are sent to a *server* that the system uses to handle actions and data processes, including the video analytics processes described in section 4.2 (face detection, liveness detection, facial recognition). Another possible architecture choice is to have the central doorbell unit itself handle this without an intermediate server; however, using a server permits the use of complex and computationally heavy algorithms that the doorbell unit cannot support (Thakkar and Ukani, "Proficient and Economical Approach" 70). Therefore, my architecture diagram shows a server where data is stored and analyzed. Servers can be local or cloud-based, but cloud-based servers are popular with smart doorbells because of the volume of data that needs to be stored. I will discuss the trade-offs between these options later on.

The main *user interface* that enables the user to interact with the doorbell is the doorbell's companion phone or web application. Once the server has processed the data sent to it, it sends the necessary notifications to the application for the user to see. The user can also use this to view live feeds and recordings downloaded from the server; communicate with the user by speaking into their phone's microphone, which will send the audio to the server, which will send it to the doorbell speaker to be played to the visitor; and tag detected faces to support facial

recognition. Data entered by the user will be sent back to the server as an update. Additionally, the doorbell might have a second user interface in the form of a central control device that can be placed within the home – this might have a display screen of some sort for homeowners to view video, as well as speakers and a microphone for audio communication.

Last, the smart doorbell might be integrated with other IoT devices – for example, smart speakers, keyless smart locks, smart alarms, or broader smart home security systems. In this case, the server interacts with those devices.

4.4. Smart doorbells and privacy

Privacy concerns related to smart doorbells largely stem from the fact that use cameras and speakers extensively. Not only are there concerns about these being hacked and used to spy on homeowners as well as those in their immediate surroundings, but also concerns about the amount of video and audio recordings collected and stored by these systems, how they are used, who has access to them, and more. These are especially significant since the doorbell cameras are on the outside of the house, and thus record people and passersby who have not consented to being recorded.

I will discuss these privacy concerns in more detail in section 5.

5. Privacy concerns

In this section, I will discuss the privacy concerns associated with voice assistants, smart thermostats, and doorbells (most of which apply to smart home devices in general). I will first detail the kinds of data collected by each type of device. I will then lay out the privacy concerns that subsequently arise from the collection and use of this data.

5.1. Data collected

Sections 2, 3, and 4 show how voice assistants, smart thermostats, and smart doorbells function – by collecting data from their users and surrounding environments, storing it, analyzing it, and then using that analysis to act. This is true of smart home devices and IoT in general. They are data generators – to work as they promise to, they need to collect enormous amounts of data. The International Data Corporation (IDC) estimates that by 2025, IoT devices will generate almost 80 zettabytes of data a year (one zettabyte is equal to one trillion gigabytes) (ARO).

Smart home devices collect data from a number of sources, namely: inputs from user interactions, cameras, microphones, and other sensors; data from device setup; and data from connected accounts, applications, and external devices. In this section, I detail the specific kinds of data collected by each of the three devices I focus on.

5.1.1. By voice assistants

As suggested by the name, voice assistants primarily collect voice data. They store voice recordings of what users say to them and transcripts of the recordings, and glean information from how users interact with them, what they ask them to do, and which third-party applications, services, and devices they interact with as a result.

Examples of data that can be collected from user interaction include, but are not limited to: content streaming activity, such as the kind of music the user likes, the shows that they watch, and the podcasts that they listen to; purchase activity, including groceries, clothes, applications, medicine, and anything else the user shops for using the voice assistant; browsing habits, based on search requests given to the voice assistant; and the content of emails, text messages, and other communications sent and received with the voice assistant, including voice conversations (“Amazon Echo Studio”). In short, anything that is done using the device is recorded and analyzed. Additionally, voice assistants might collect biometric information, such a voice profile, in order to “recognize” users who speak to them (“Amazon Echo Studio”). The metadata of these recordings is also collected – for example, the date and time the recording was made, the location, and, if the voice assistant does distinguish between users, which user spoke.

Data is also generated simply from setting up a device – this includes actions like creating an account, purchasing the subscription if necessary, and connecting the device to Wi-Fi. This data includes: identifying information, such as name, contact information, and age, which might be entered during account creation; location data; payment information; and device and network information, including IP addresses, Wi-Fi credentials, usage and crash data, device settings, and device names. For example, if you have multiple connected voice assistants in different rooms, their names – ‘Bathroom Alexa’, ‘Kitchen Alexa’, etc. – will be stored (Clauser).

Furthermore, data is collected from third-party accounts, applications, and devices that are connected to the voice assistant in order for it to carry out commands. When a Google account is connected to a Google Nest voice assistant (Google requires that you have a Google account to use its assistant), the voice assistant gains access to all information associated with that account – Google search history, Google Drive usage, games downloaded from Google Play, and more. The names, phone numbers, and addresses of people in a connected Contacts application will be collected; meetings and events from a connected Calendar app; and playlist names from a connected Spotify. If smart home devices are connected to the system, as they often are so that the user can run the home using the voice assistant, data is harvested from them too – the house layout created by a robot vacuum, the usual coffee order from a smart coffee machine, and temperature set points from a smart thermostat.

5.1.2. By smart thermostats

For smart thermostats, the primary sources of data collection are sensors within the house (both built-in and connected) and user inputs to the thermostat. Various kinds of sensors are used to collect environmental data, including internal temperatures, ambient light measurements, humidity levels, carbon dioxide levels, smoke levels, and detected motion ("Google Nest Learning Thermostat"). Motion sensors specifically collect data on occupancy and sleep patterns: when the user leaves the house and when the user returns; when the house is occupied and when it is empty; when the user is asleep and when they are awake ("Google Nest Learning Thermostat"). Motion sensors in different rooms can collect movement data from within the home, for example, how much time is spent in each room, when certain rooms are used, and when they are empty. Door and window sensors can collect data on when doors and windows are opened and closed.

Smart thermostats also collect data on user's temperature preferences from user interactions with the thermostat device itself and with its companion application. Every adjustment to the temperature set point and set temperature schedule is recorded, along with metadata like the timestamp and location of the adjustment. This data is used by the thermostat to learn and store what temperature the user likes at different times of day, different times of year, and in different rooms of the house. Other direct adjustments, such as changes to settings, are also recorded. Additionally, if the thermostat has a microphone or is smart speaker enabled, it also collects voice data. Google Nest thermostats even have code that prompts the user to input information about their home (such as the layout) which Google says will help energy providers to "generate energy more efficiently" (Hernandez et al. 2).

Similar to voice assistants, data is generated from setting up the thermostat and sensor devices. This data includes: account data and set up information, like name, contact information, and address; system and software logs, including device settings, HVAC settings, and wiring configurations (Hernandez et al. 2); device usage statistics; application data, such as login times and locations; Wi-Fi network, credentials, and IP addresses; device placement; and device names.

Since smart thermostats can also be connected to other online accounts, they can also collect the associated account information, as mentioned in section 5.1.1. If a user enables geofencing, the thermostat can collect not only location data about the home, but the user's location when they travel away from it, since geofencing uses GPS signals to determine where the user is and judge whether they are 'home' or 'away'. Then the thermostat can collect

geolocation data about the user whenever their phone is connected to the internet. Additionally, smart thermostats collect information on energy usage and bills from the HVAC system and connected utility accounts.

It is worth noting that despite the fact that the user does not actively engage with a smart thermostat to the extent that it engages with a voice assistant, the smart thermostat still collects extensive, potentially sensitive data about the user.

5.1.3. By smart doorbells

Smart doorbells collect data from cameras, microphones, sensors, and doorbell interactions, and, like voice assistants and smart thermostats, data from device setup and connected accounts, applications, and devices. (I won't go into excessive detail about data from the latter two sources, since it is similar for all smart devices – broadly, this includes things like identifying details, device and network data, and location data.)

Smart doorbells collect data on doorbell activity, as well data recorded by their cameras and microphones, along with the corresponding metadata. This includes each time a doorbell was pressed, images taken, videos recorded, and audio picked up (Burgess, "Data Amazon's Ring Cameras Collect"). They may also collect data from external sensors, including motion data, temperature, and ambient light ("Amazon Ring Video Doorbell"). The primary data collected is video data – this includes videos of people and animals who approach the door or ring the bell, and videos of people in the vicinity of the house if they are within range of the motion detector and thus activate the camera. Ring doorbells, for example, can detect movement "up to 155 degrees horizontally" and can potentially record people walking down the street who do not otherwise interact with the doorbell or house; some Ring doorbells can record audio from up to 20 feet away, enabling them to record conversations of passersby (Burgess, "Data Amazon's Ring Cameras Collect").

Additional data is collected from the companion application of the doorbell and connected accounts or devices. For smart doorbells specifically, this might include geolocation information if geofencing is enabled ("Amazon Ring Video Doorbell"); data generated from facial recognition, including names that the user inputs to tag detected faces, with their corresponding images; underlying face prints of image, for doorbells that have biometric features ("Eufy Video Doorbells"); and, as in the case of Ring, actions as minute as zooming in on footage in a recorded video (Burgess, "Data Amazon's Ring Cameras Collect"). Ring also has a Neighbours application that can be used by Ring users and law enforcement agencies to share suspicious activity and crime alerts and videos – everything uploaded to this site, including videos personally recorded by users and not by doorbell cameras, captions, comments and interactions on other users' posts, and post metadata, is recorded (Burgess, "Data Amazon's Ring Cameras Collect"). Another smart doorbell provider, Arlo, has users create an "emergency response profile" which can contain sensitive information such as age, sex, and gender expression, that the smart doorbell then stores as associated with the user ("Arlo Video Doorbell").

5.2. Privacy concerns

The reliance on vast amounts of data is the root of most smart home privacy concerns. Humans intuitively consider the home to be 'private' – yet by inviting smart devices, with their

various sensing technologies and internet connections, into their lives, individuals expose their innermost lives to, essentially, the world. Worries include, but are not limited to: what does the data reveal, and is it something that the user doesn't want anyone to know? Who has access to the data? How is the data being used? Can it be used against the person it was collected from, or in a way that causes them harm? Additionally, can the devices themselves be compromised and used to spy on residents?

Some of these privacy concerns arise due to the efforts of companies to use collected data in order to generate insights on user behaviour and generate profits. Others arise due to weak security in data transmission, data storage, and the devices themselves. Since they relate to all three smart devices, I will analyse them together.

5.2.1. Invasive inferences

Data generation does not end at data collection. In general, the amount of data collected is concerning not just because it is sensitive itself but because it can be used to reveal even more sensitive information. Those with access to collected data can combine it with other information from third parties and analyze it to make "inferences" about users which reveal information about users that they do not expect or want anyone else to know – a clear violation of privacy. Since the data collected from smart devices comes from within the home, and since it is so vast, inferences drawn from it can be particularly revealing and personally identifying.

Here are some examples of inferences that can be made by each smart home device discussed:

Voice assistants might infer the following from voice commands: height and weight; dietary restrictions and lifestyle choices like veganism, from food orders and recipes searched; protected classifications like gender, race, religion, and sexual orientation; health information, perhaps from medical correspondence and online searches; financial status; personality traits; and more. For example, a user might search for prenatal vitamins and first trimester tips, enabling the voice assistant to infer that the user is pregnant; if the user later searches for wine, the voice assistant might infer that they are no longer pregnant ("Amazon Echo Studio"). Metadata also helps to create inferences. Timestamps of commands to smart devices can be used to determine when a user locks the door, turns off the lights, and runs the robot vacuum ("Google Nest Mini"). Timestamps combined with device names and voice commands can be used to infer daily schedules: if a user plays music in the bathroom every morning on a device labelled 'Bathroom Alexa' and in the evening on 'Kitchen Alexa' after searching for a salmon recipe, Amazon might deduce when they shower and cook each day.

Smart thermostats might infer the layout of the house using precise names and locations of sensors and thermostats within different rooms. Data from motion sensors in the house can be used to put together movement patterns and even sleeping patterns; for example, if motion is detected as odd intervals throughout the night, it might be inferred that someone in the house has trouble sleeping. If a user enables geofencing, their GPS coordinates are recorded even out of the house (while their phone is connected to the internet); if they are at the same location every weekday from nine to five, this might be inferred to be their workplace. Temperature preferences combined with external data on weather can reveal if a user runs hot or cold, and utility bills and energy usage data might reveal financial status.

Smart doorbells significantly can make inferences about people outside the home. For example, using recorded videos and facial recognition data, a smart doorbell might infer that a

specific neighbour goes on an hour-long run each day around 5 pm. If a user puts up Christmas or Hanukkah decorations in their yard every December, this might be used to infer their religion. Additionally, inferences might be made from recorded conversations had with visitors, similar to inferences made by voice assistants.

(The examples here are not exhaustive. Many of the further privacy concerns that arise hinge on how these inferences can be used, so inferences will be referred to throughout.)

5.2.2. Improper data access and use

5.2.2.1. Smart device companies

Smart device companies say that they collect and store data to improve the devices and services offered to users. Indeed, the ability of smart devices to learn from user behaviour is a key feature. The more data collected and the more understood about a user's behaviour, the more accurately the smart device can make predictions. For example, Ecobee uses data from its smart thermostat to make more intelligent decisions about energy usage in order to reduce energy costs ("Ecobee Smart Thermostat"). Similarly, a Ring doorbell might learn from stored videos to perform better facial recognition and more accurately identify visitors to the user. Amazon Echo might learn the nuances of a user's communication from past recordings to provide more correct and relevant answers.

However, smart device companies do not just use this data to improve their models, but to generate profit from user behaviour through targeted marketing and by personalizing products to get users to spend more time using them. A smart doorbell company might infer from a user's constant checking of live feeds that they are paranoid and try to sell them many more security products. Going back to the pregnancy inference example, a Google voice assistant might infer that a user is pregnant from a question about first trimesters and that they are no longer pregnant from a command to buy wine, and further infer from the timing between the requests that the user has had a miscarriage. The user might then see Google search autocompletions or suggestions for dealing with loss or getting pregnant again. If Google is wrong, this can be devastating, but even if Google is right, many users find this level of predictive power invasive and creepy.

There are further concerns about how smart device companies manage data. What access do employees have to data? In 2019, it was revealed that Apple employees reviewing Siri recordings for quality control regularly heard confidential information, including business dealings, doctor and patient discussions, and sexual encounters, accompanied by identifying user data (Hern, "Apple overhauls Siri"). Therefore, even if smart device companies don't share your data with others, their own use of it provides plenty of cause for worry.

5.2.2.2. Third parties and advertising

Another privacy concern is data being sold to or shared with third-party companies (such as advertisers, service providers, and business partners). Most smart device companies claim not to sell personal data to third parties, but do provide them with access to it for targeted advertising and tracking purposes. Advertisers use this data to determine when and what ads to send in order to increase the likelihood of a sale – for example, by determining a user's schedule from their smart thermostat or voice assistant data, they can send specifically-timed ads that are more

impactful at certain times of day or depending on whether a user is at work, school, or home ("Google Nest Learning Thermostat"). Another example is an advertiser determining that a user runs cold from their thermostat data and targeting them with ads for jackets, hot drinks, and space heaters.

Shared data might also be used in other equally invasive ways. For example, insurance companies might use this data, which they would not otherwise have access to, to determine insurance rates for specific users.

Finally, there is the risk of third-party companies identifying users from data. Data is often shared with the defense that it is de-identified (Bose's privacy policy says that they may use or share de-identified data "without limitation") and therefore not a risk ("Bose Smart Speakers"). However, de-identified data is different from anonymous data because it is personal data, and, especially when it includes location information, can be re-identified without too much difficulty ("Bose Smart Speakers").

5.2.2.3. Governments and surveillance

Another privacy concern is the possibility of these vast collections of data being used for law enforcement purposes. In 2022, Ohio Ring user Michael Larkin was phoned by local police and asked for few hours' worth of footage from his front door camera to help with an investigation on a neighbour, and he provided it; the police then asked for the whole day's footage; a week later, Ring itself instructed Larkin to send footage from all 21 of his cameras to the police, who had received and delivered a warrant to Ring (Ng). Larkin had a say in the first two cases, but in the last he did not.

Smart doorbell companies, Ring in particular, have historically cooperated with law enforcement agencies, providing video surveillance footage for use in investigation; police departments can even request video footage from users directly on the Ring Neighbours app (Ng).

When law enforcement agencies provide warrants, companies do not need user consent to provide this data, but there are concerns about governments and police being given access to user data without user consent even when they don't have warrants. Amazon admitted that in 2022, it gave Ring footage to police without a warrant or consent at least 11 times (Biddle). Additionally, as the number of warrants requested increase, there are concerns about users losing control of their data.

Most importantly, occasional law enforcement access can easily turn into general and constant surveillance. If governments are given access to a continual stream of data about citizens, they can use it to gain insight into private activities and potentially prosecute or discriminate against individuals based on things they do with an expectation of privacy. For example, government officials can use video footage to determine if pregnant women carry their fetus to term, or location data to keep track users who go to Planned Parenthood – in a country with divided views and laws on abortion, this could be used to harass, flag, or even arrest individuals with certain beliefs ("Amazon Ring Video Doorbell"). Online searches could be used to reveal political, social, and religious beliefs, based on which individuals might face discrimination or ostracization. Combined with location, conversation, and identifying data, the government might even be able to infer things like the location of organizing meetings for political protests and the people involved, and shut them down or prosecute them. Overall, the

number of sensors, cameras, and microphones that are now near constantly recording data might be setting the stage for a future surveillance state.

5.2.3. Security weaknesses

5.2.3.1. Database breaches

Due to the amount of data collected and the computational power needed to interpret it, smart devices typically store and process data in cloud databases or servers (Lynskey). These data stores become immediate targets for hacking and data breaches that will result in tons of user data falling into the hands of bad actors.

This unauthorized access can and has been achieved in a number of ways. If databases are weakly protected or unencrypted, hackers might be able to break into them; the FTC filed a complaint against Ring doorbells in May 2023, accusing them of failing to keep user data secure by storing video recordings unencrypted, among other things ("Amazon Ring Video Doorbell"). A failure to implement access rights to data can enable employees and third-party contractors to access all data even if it isn't necessary for their job, which can additionally lead to data leaks, such as when an Amazon employee leaked customer email addresses in 2020 ("Amazon Echo Studio"). In 2022, Apple even gave up data, including addresses and phone numbers, to a phishing scam where hackers pretending to be law enforcement officials sent "emergency data requests" ("Apple Homepod"). Finally, there is the concern that third parties who have been given access to data may not implement the same data protections as smart device companies themselves; in 2021, third-party company GetHealth leaked Apple Healthkit data, including identifying information and medical data, after storing it in an unencrypted, non-password protected database ("Apple Homepod").

Bad actors, whether they are hackers or company employees, can cause great harm to individuals using their data. Hackers can use voice data to create imitations of users' voices (Zahn); thieves can use data about house layouts and occupancy patterns from smart thermostats to determine when to break into a user's home ("Google Nest Learning Thermostat"); footage from smart doorbells, combined with facial recognition data and location data, can be used to identify individuals and potentially stalk, harass, or blackmail them.

5.2.3.2. Device breaches

In addition to data being compromised, there are concerns about smart devices themselves, with the large number of sensors, cameras, and microphones that they use, being compromised and used to spy on individuals.

Smart devices can be hacked via unsecured or weakly secured Wi-Fi and Bluetooth connections. USB ports can also be exploited to install malicious firmware, or used as backdoors into the device (Hernandez et al. 5). Such security bugs enable hackers to access cameras and microphones and even control them, and thus watch and listen to individuals through their devices. Additionally, hackers can use compromised but not very sensitive smart devices, which might be less protected, as gateways to connected systems or devices. This is called lateral movement and has happened before – in 2019, a hacker got into a couple's smart thermostat and used it as a backdoor to access the smart security cameras on the same network (H. Peterson).

5.2.4. Violation of user rights

5.2.4.1. Data collection without user consent

While smart device users know that they are agreeing to some amount of data collection, there are concerns that smart device companies collect additional data without user consent, whether intentionally or unintentionally.

With voice assistants (and other smart speaker enabled devices), a specific worry is that they are always listening. Technically, voice assistants do not record every word said in their vicinity. They "selectively listen" for their wake word and only begin recording after detecting it, when users are intentionally communicating (Clauser). However, accidental activation, when a sound is misinterpreted as a wake word, is not uncommon. A Bloomberg analysis of Alexa transcripts showed that Alexa "woke up accidentally" over 10% of the time, and Apple's HomePod is also guilty of this (Lynskey; Denham and Greene). As a result, conversations not had with the voice assistant can be recorded, stored, and analyzed relatively often, amounting to a huge violation of privacy. Additionally, most voice assistants cannot filter different voices (Clauser). This means that after the wake word is detected (or falsely detected), anyone in the vicinity is subject to being recorded, including children and guests who did not consent to this.

Smart doorbells are not even subtle about violating consent. Since they have cameras, microphones, and sensors that point outwards, they regularly record the actions and conversations of passersby who have not consented to it and who are not even attempting to engage with the doorbell – especially if the cameras and microphones have long ranges. In 2021, a UK woman won a court case against her neighbour where she accused him of infringing on her privacy by setting up Ring cameras that pointed at her home, whereby he could "see" and "listen" to her ("Amazon Ring Video Doorbell").

5.2.4.2. User control over collected data

Finally, privacy concerns arise when users do not have adequate control over their data. Namely, if users cannot easily opt out of (and opt into) data collection, data use for training, and data sharing, or if they cannot easily review their collected data and have it deleted, then they do not have control of their data and cannot make choices to protect their privacy.

Not being able to delete data is especially concerning. Some companies allow users to easily delete collected data; others make it difficult but ultimately provide a means to do it; others still might delete certain data like video footage or an audio recording of a purchase, but still keep data gleaned from the interactions, such as what was purchased. Amazon's Ring privacy policy states that while people can delete recorded videos using their account, Ring might still retain them ("Deleted Content and Ring Protect Recordings may be stored by Ring in order to comply with certain legal obligations and are not retrievable without a valid court order") (Burgess, "Data Amazon's Ring Cameras Collect"). There is also the question of whether data that has been shared with third parties is subject to deletion requests.

Without the ability to completely delete collected data, users that buy smart home devices are essentially signing away their right to privacy. Any information about them that is collected is potentially out in the world for good and might be used against them at any point in the future.

6. Technical methods of privacy preservation

Having categorized and established the major privacy concerns that arise from the use of smart home devices, I will now discuss suggestions for technical methods to counter them. Broadly, I will look at methods to simply collect less data overall; better secure data to reduce the risks of data sharing, processing, and prevent unauthorized access; and to improve device security. (I break down the technical suggestions this way to remain in keeping with the categorization of privacy concerns as far as possible, and thus build a framework where the problems and solutions are easy to connect to each other. All of the suggestions themselves are not necessarily exclusive to one kind of risk, but I point out all risks addressed by each technical suggestion.)

6.1. Reducing data collection

As previously mentioned, the root of most privacy concerns relating to smart home devices is that they generate and use so much data. It is undeniable that this data enables them to function in ways that are useful to the user – to ask for smart devices that don't collect data at all is unreasonable and defeats the purpose. However, if the amount of data collected can be reduced some amount without compromising too much on the smart device's accuracy, this will greatly reduce the risk of privacy violations (here, data that is 'collected' refers to data that is sent to company servers, not data kept on the local device with the user). To that end, here are some suggestions that might enable this.

6.1.1. Data minimization

Data minimization is the practice of only collecting data according to need. This might sound like an obvious method, but that is it because it is – only collecting data when it is absolutely necessary is obviously a way to reduce the amount of data collected. Certainly some smart device functions might require data collection, but some also might not. Siri, for example, uses as little data as possible when delivering results ("Improving Siri"). If a user asks a question about a sports event, Siri can provide it using only general location data, but will use more precise location data for a question that requires it, such as the distance to the nearest football stadium ("Improving Siri"). In short, some actions simply don't require a huge amount of data collection, and only data that's actually essential to completing them should be used. As an example, a voice assistant on a laptop with a 'speech' feature should not have to collect all the content in a user's text messages if the user asks it to read them aloud; it can simply tell the laptop to read out the texts using its 'speech' feature.

6.1.2. Synthetic data

With the advent of generative AI and therefore tons of AI-produced information and content, using synthetic data sets has become a possible identity-protecting option for smart device companies. AI models can be trained on real, identifiable information to generate fake data points for synthetic datasets that are "statistically identical" to the real ones but have no information relating to real people (Hern, "'Anonymised' data").

Of course, this requires some amount of real, identifiable information to be collected and processed – however, once the synthetic data is generated, the real information can be disposed of. Overall, less data needs to be collected.

6.1.3. Federated learning

One of the biggest challenges with making privacy-preserving smart home devices is that in order to provide accurate, personalized services that users want, they need to learn from a lot of data – and not just data from one user's device. Therefore companies need to collect a lot of data from different users and train models on all of it, and this makes users uneasy. The issue is then how to personalize a smart home device without "hoovering up your data" (Hao). A solution for this is federated learning.

Federated learning is a privacy-preserving machine learning technique that allows smart devices to learn from data without collecting it to transfer to external servers (Hao). It involves training different copies of machine learning models with local data on all users' devices, and then sending the trained models rather than the data to the cloud or central server, where they are combined into the "master model" that the smart home device uses to make predictions (Hao). Therefore the smart device can get smarter without the company keeping the user's data.

Apple uses federated learning to improve Siri and provide complex, necessary functionality. For example, Siri can now distinguish between voices so that it only wakes up when the owner of the phone says, "Hey Siri" and not when anyone says it – this is important because a user saying "Hey Siri" should not activate all other iPhones nearby (Hao). This kind of voice detection seems like it would require a lot of voice data, but federated learning enables it without requiring Apple to collect and keep users' voice data. Similarly, Google uses federated learning to refine detection of "Hey Google" and reduce accidental awakening – voice data on the local device is used to adjust a local model, and a summary of that model is sent to Google's servers and combined with summaries from other users' devices to improve Google Assistant functionality for everyone without putting their data at risk.

6.2. Improving data security

6.2.1. Encryption

A standard way to protect user privacy is by encrypting user data. *Encryption* is the process of changing plaintext or usable data into an unreadable format using a cryptographic algorithm and key, in order to prevent anyone except the intended recipient from reading the data ("Encryption"). *Decryption* with the key is the only way to restore the data to its original state ("Encryption"). *Symmetric encryption* uses the same key to encrypt and decrypt the data; *asymmetric* encryption uses two different keys, a public key to encrypt data and a private key to decrypt data.

6.2.1.1. End-to-end encryption

Many companies encrypt data at rest (while it is stored or otherwise not in use) and in transit (while it is transmitted over networks to third parties or between devices or systems). If there is a data breach or network eavesdropping, the compromised data at rest or in transit is then

unusable. However, encryption at rest and encryption in transit work in their specific contexts only – that is, encryption algorithms are employed by the database or server for data at rest and by networks (using secure communication methods like SSL and TLS) for data in transit (Ertl).

Additionally, servers typically use Advanced Encryption Standard (AES) algorithms which use symmetric encryption. The data is encrypted with the symmetric key, which then must be sent to the user so that they can decrypt it. Since the key is being transmitted as well, the channels through which the data is sent must be secure – if a bad actor accesses the key, they can decrypt the data. Importantly, users must also *trust* the third party managing the keys and encryption in that context – even if it has strong security, there is always the risk of the third party leaking the key or having to provide it to law enforcement or other authorities upon production of a warrant. However, since the sender cannot guarantee the security and trustworthiness of all the different networks and servers that the data might travel through, this puts the data at risk.

Therefore smart devices can benefit from using *end-to-end (E2E) encryption* instead. This encrypts the data independent of the technologies used in the server or for transmission, and then transmits the fully encrypted data to the recipient (Ertl). Specifically, it uses asymmetric encryption (Ertl). This involves encrypting the data before transmission with the recipient's public key, which they can share with anyone who wishes to communicate with them – in this case the smart device company. The recipient does not share their private key, so only they can decrypt the data. Even if the data is intercepted in transit, it is encrypted and thus unusable. Additionally, data is encrypted once (at the start) and decrypted once (at the end), rather than at multiple points along the transmission (Ertl). Therefore E2E encryption is much better at preventing unauthorized users from accessing data.

Furthermore, since network or Internet Service Providers (ISPs) do not provide the encryption, they never have access to unencrypted data; if a law enforcement agency requests it, they can only be given the encrypted and therefore useless version (Hassel). There is also no benefit to their leaking it.

E2E encryption does not come without trade-offs or risks of its own. First, E2E encryption only means that encryption keys are generated and managed on the device rather than by a third party – the key can still be compromised if the device is hacked or otherwise compromised, and then the E2E encrypted data is no longer protected. Therefore E2E encryption requires that the device have strong security of its own. However, E2E encryption does reduce the issue of trust, because only the local device itself needs to be trusted and users know what device that is. Second, the asymmetric encryption that it uses is slower and more resource-heavy than symmetric encryption, therefore it is less efficient for sending larger amounts of data such as those managed by smart devices. However, this can be mitigated by using asymmetric and symmetric encryption together. For example, E2E or asymmetric encryption might be used to establish a secure, trusted channel for exchanging a symmetric encryption key, often called a 'session key', that was used to encrypt data. In other words, the key for a symmetric encryption algorithm can be asymmetrically encrypted, reducing the risk of transferring it to the recipient of the data – then the larger amounts of data can be transferred using symmetric encryption without worrying about the symmetric key being compromised in the process. After the communication session ends, the session key is discarded – a new one is generated for a new session.

6.2.1.2. Encryption in use

The types of encryptions discussed in section 6.3.1.1 cover data at rest and data in transit. However, there is an additional state of digital data that they do not cover – data in use. Data in use is data that is being actively accessed or otherwise used (updated, input, processed) by a device or application; it is typically stored in the local memory of the device, in unencrypted form so that it can be processed and computations can be run on it (Velimirovic). This makes it a target for hackers. Furthermore, it comprises not only sensitive and personally identifiable information but decryption keys for encrypted data at rest – if data in use is compromised then those keys are compromised, following which data at rest is compromised. It is clear that data in use cannot be ignored when taking measures to safeguard privacy.

I discuss both memory enclaves and homomorphic encryption as means of protecting data in use.

6.2.1.2.1. Memory enclave encryption

Traditionally, data in use has been protected using encrypted memory enclaves. With this technique, all data processing and computation on data is done within a private region of memory called an *enclave* or *trusted execution environment (TEE)* (Rjaibi). Access control is implemented within the enclave, so data within it cannot be read or modified by any process outside it at any point; the CPU only decrypts data (for processing) which is within the enclave (Rjaibi). The enclave memory itself is encrypted either with hardware-based encryption, whereby specialized hardware components provide built-in enclaves, or with software-based encryption, whereby encryption algorithms encoded in the software encrypt and decrypt memory as necessary (Velimirovic; Rjaibi). Thus, the decrypted data within cannot be accessed even if the enclave is – the enclave itself is secured (Rjaibi).

Memory enclave encryption therefore protects data by isolating it within a secure region. However, the data itself still needs to be decrypted to be processed.

6.2.1.2.2. Homomorphic encryption

Homomorphic encryption provides a different, cryptographic approach to securing data in use. Specifically, it allows computations to be performed directly on encrypted data. The data never needs to be decrypted, thus does not become a target for a data breach. (It is worth noting that homomorphic encryption is still an emerging technology and may not be as immediately applicable as some of the other technical methods mentioned for reasons which I examine at the end of this section – however, I include it in the overall discussion because it so directly addresses concerns relevant to smart home devices.)

Let the unencrypted data to be processed be x; the device would like to perform computations on it to get the output $f(x)$. The encrypted form of x is $E(x)$. Homomorphic encryption enables the device to process $E(x)$ and get $E(f(x))$, which can be decrypted to $f(x)$, the desired output. This computation is done without ever knowing what x is. To rephrase this, computations on encrypted data are done in such a way that, once the output/data is decrypted, the output is the same as if the computations were done on plaintext data (Velimirovic).

Of the many benefits of homomorphic encryption, two are especially relevant to smart home devices. First, it makes it much safer for smart device companies to use cloud computing

and servers, which are often needed for smart device data because of its sheer volume and computational demands; smart device companies that use homomorphic encryption can enjoy the computational power of the cloud without compromising on safety. Second, it reduces the risk of sharing data with third-party service providers or companies, which smart device companies also often make use of. Having a third-party process the data, for example, run a machine learning model on it, looks something like this: the decrypted data is encrypted by the smart device company with its public key and sent to the third-party. The third-party then performs their computation on the encrypted data, receives an encrypted result, and sends it back to the smart device company. The smart device company can then decrypt the result with their private key – as mentioned previously, this decrypted result is the same as the result that would have been gained from processing the unencrypted data.

There are, of course, always limitations. Fully Homomorphic Encryption (FHE), allowing all kinds of computations to be performed, has extremely high computational demands, making it slow and requiring huge amounts of storage or memory because it can increase the size of processed data; and some complex operations cannot be efficiently performed on encrypted data (Velimirovic). This means that perhaps for the moment, homomorphic encryption is not practically applicable to smart home devices. However, there are less demanding kinds of homomorphic encryption that companies might be able to start with – Partially Homomorphic Encryption (PHE) allows certain selected computations and Somewhat Homomorphic Encryption (SHE) allows a limited number of computations below a certain complexity threshold. Using either of these may still provide a certain amount of protection, and a certain amount of protection is better than none.

Additionally, at the current pace of innovation, it is not preposterous to think that breakthroughs in technology might speed up homomorphic encryption enough that it will be able to be applied to smart home devices in a few years. French start-up Ravel Technologies said in 2022 that it had made FHE "scalable" and "successfully overcome FHE's biggest challenges" of slowness (Pasternack); CEO Rand Hindi of Zama, a cryptography company, said that hardware accelerators will speed up FHE by at least 1000 times by 2025. With the amount of money being invested in advancing homomorphic encryption, it might become practical to use sooner than expected. Furthermore, researchers from Leipzig University in Germany demonstrated that FHE is both feasible and practical as a solution for privacy concerns in smart mobility (Hannemann and Buchmann 9); smart mobility is an IoT approach to making urban transport greener and more efficient, and is on the smart city rather than smart home scale, but the feasibility of FHE in an IoT context is a good sign nevertheless.

6.2.2. Anonymization and pseudonymization

Anonymization and pseudonymization are two of the most commonly used techniques to keep data private. The goal is to transform the data such that it is not associated with the user, or at least make it difficult for anyone to be identified from the data. This reduces the risk of privacy violations if there is a data breach, and when sharing data with third parties for advertising or other services.

Anonymous data is "information that does not relate to an identified or identifiable natural person" or "personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable" ("Recital 26"). This process of irreversibly changing data so that it does not link back to a person is *anonymization*. It is done by removing direct and indirect

personally identifying attributes from the information, including names, SSNs, and birth dates. The trade-off, of course, is a loss of information that might cause the anonymized data to be less accurate or useful for some purposes.

Pseudonymization is "the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject (individual person who can be identified) without the use of additional information", with this additional stored separately and securely ("Art. 4"). Pseudonymization does not remove all identifying data, but de-identifies it or reduces the likelihood that data can be linked to the identity of an individual (van Schendel). Data points such as name, SSN, or birth date are replaced with random characters or codes ("Differences"). For example, Apple's Siri uses a random identifier (in its case, an alphanumeric string associated with one Siri device) to track device data instead of using identifying details from users' Apple accounts. After six months, the data is disassociated from the random identifier (Hern, "'Anonymised' data").

It is important to note that pseudonymization is not the same as anonymization. Anonymized data is meant to be irreversibly unidentifiable, while pseudonymized or de-identified data is only unidentifiable without the mentioned additional information. This means that pseudonymized data can be re-identified.

However, neither anonymization nor pseudonymization are strong enough measures of privacy protection. It has been shown even anonymized or aggregated data can be de-anonymized in a number of ways. For example, "anonymous" medical billing data from Australia's health department was re-identified by cross-referencing "mundane" data like the birth years of older mothers or mothers with many children (Hern, "'Anonymised' data"). A team of researchers from London and Belgium built a model that indicated that a dataset with 15 demographic attributes "would render 99.98% of people in Massachusetts unique" (Hern, "'Anonymised' data"). Latanya Sweeney shockingly showed that 87% of the US population can be personally identified with just zip code, birth date, and gender. By "linking" this data, from an "anonymized" healthcare database, with data from public voter records, Sweeney found the Governor of Massachusetts' personal health record ("Differential Privacy").

Therefore it is clear that auxiliary or additional data from external sources can be 'linked' with apparently anonymized or pseudonymized data and used to de-identify it. There are countless such sources; smart device companies can hardly predict what additional data someone might be combining with the data that they share themselves, and what it might reveal. Therefore, relying on the above methods to make data private is not enough. Other methods (discussed independently in detail) do a more meaningful job of "anonymizing" data: homomorphic encryption, which allows the processing but not reading of encrypted data, which therefore cannot be understood to re-identify; differential privacy; and synthetic data, which is simply not associated with any real individual.

6.2.3. Differential privacy

In section 6.3.2, I discussed how data from multiple sources can be "linked" together or combined to de-anonymize it and identify specific individuals, making it risky to maintain large databases or look for patterns in large amounts of data. Differential privacy aims to counter that, addressing "the paradox of learning nothing about an individual while learning useful information about a population" (Dwork and Roth). In other words, it enables the accessor of the

data to find patterns and statistics of interest in the dataset without learning being able to tell if any individual's affected them (Dwork and Roth).

This is achieved by introducing some amount of noise to the dataset, which "deliberately fuzzes every individual data point in a way that averages out across the dataset" (Hern, "'Anonymised' data"; Hao). Noise here refers to small, arbitrary changes which do not change the overall patterns of the data or reduce its accuracy – for example, the population count of a census data set must not change. Once the dataset is thus adjusted, it can be stored, used to train machine learning models, or shared with any third parties with much less risk of identification (Hao). The data in the dataset has been made "technically incorrect", and therefore difficult to reverse engineer to get original data (Hern, "'Anonymised' data").

A concern with differential privacy is finding the balance between inserting noise into a dataset to protect privacy and ensuring that the data is useful. If less noise is introduced, the data is less private but more accurate; if more noise is introduced, the data is more private but less accurate. However, this trade-off is not unique to differential privacy – attempting to make a dataset more private via the others means discussed usually also leads to a decrease in accuracy one way or another. Yet differential privacy offsets this concern by providing truly anonymized data that can't be used to identify individuals even if it is supplemented with data from other sources. By doing so, it allows large databases to be stored and analyzed without worry – for smart device companies, which gather so much data, this is ideal. In fact, it works better on large datasets or databases because, the more individuals represented, the more the effect of any one on the whole is reduced ("Differential Privacy").

Apple and Google have been using differential privacy for some time now; for example, Google collects user location data to provide useful aggregate statistics like place busyness in Maps or energy usage data from Nest thermostats to provide energy reports, but adds noise to the data so it cannot be used to identify individuals. It is certainly worthy of consideration by smart device companies in general.

6.2.4. Cloud versus local storage

The volume of data collected by smart devices, and the heavy computational demands of processing it, mean that most smart device companies use cloud servers for storage and processing rather than local servers. Cloud storage and processing is a cause for concern with many users of smart home devices. While they are more scalable than local ones, and accessible from anywhere via the internet, they also present greater security risks that put user privacy on the line. Cloud or remote servers are popular targets for hackers due the amount of data stored in one place and the internet connection, and have been subject to data breach after data breach; data on them can be accessed and used by company employees or third parties; and users' have less access and control over their own data. Therefore, I include below some suggestions that reduce or minimize cloud storage in order to protect privacy better.

6.2.4.1. On-device storage and processing

An obvious way to protect privacy is to not use the cloud at all. Instead, data can be stored and processed on the device itself. The benefits of this are clear. Even for cloud servers that have strong security features, storing data on them still exposes it to a potentially large number of external people, including smart device company employees and cloud service

providers. This increases the risk of data leaks and other misuse of user information. Cloud servers also become targets for hackers and other malicious agents. Using local storage, on the other hand, keeps the data in the user's control. Additionally, on-device processing is actually faster than processing using the cloud ("Eufy Video Doorbells").

There are, of course, trade-offs associated with using local servers. They are less scalable, which can be an issue with the amount of data collected by smart devices. They may also not be able to support more complex actions or features, like facial recognition and liveness detection in smart doorbells. Not using cloud servers also means giving up remote access - there are security features to not having your data on a device or server connected to the internet, since Wi-Fi and other networks cannot be exploited to access data, but remote access is a key feature of smart devices and essential for some of them. For example, smart doorbells that don't send alerts and video recordings to the user when the user isn't home but has a visitor are losing a large part of their functionality.

This doesn't mean that smart device companies have shunned the idea of on-device processing completely. In 2021, Apple made changes to Siri to address privacy concerns and improve its performance (Yang). Apple decided that it no longer upload Siri audio recordings to Apple servers and would instead process users' requests and commands on their Apple device itself with on-device speech recognition, thus addressing "one of the biggest privacy concerns for voice assistants, which is unwanted audio recording" (Yang). Also, not having audio recordings on Apple servers means that in the event of an Apple data hack or leak, audio recordings will not be at risk. The video doorbell company Eufy takes it a step further – it has no cloud subscription and stores and analyses all camera footage locally ("Eufy Video Doorbells"). This kind of processing is especially useful for such sensitive data.

Ultimately, the choice to use only local storage and on-device processing depends on the company and user's needs. However, smart device companies can offer users the choice, laying out the trade-offs, rather than choosing for them.

6.2.4.2. Using a combination of databases, servers, or caches

A storage and processing solution that might protect privacy more while still providing users with some level of the promised functionality of a smart home device is one that splits storage and processing depending on the sensitivity of the data. Data can be split or moved between local and remote servers depending on how identifying it is and what the computational demands are, based on what it's used for. Additionally, different databases can be used to spread data out rather than having it stored in one place that becomes a target for hackers.

As an example, revealing data such as real-time occupancy or location data, which if accessed by a bad actor can be harmful to a user, can be stored in a separate local database for some amount of time before it is transmitted to a cloud server to be used for machine learning or other functions; otherwise, computation that requires sensitive data as an input can be done locally, while less sensitive data can be processed in the cloud in order to reduce the computational demands on the local server.

6.3. Improving device security

6.3.1. Offline processing

I have mentioned on-device processing previously but bring it up again as a way to make smart home devices themselves more secure – another privacy concern to address. Some functions of smart home devices can be achieved without an internet connection, like voice assistants launching applications, changing music volumes, or setting alarms. This means that internet security weaknesses cannot be exploited as a means of getting access to a user's smart device and spying on them. Often a user's smart device might have adequate security, such as strong passwords and multi-factor authentication, but be connected to a poorly protected Wi-Fi network.

Of course, certain requests and actions require the internet – for example, asking questions to Siri or Alexa that they don't have the answer for locally. However, if smart devices can use this without compromising much of their functionality, it might be worth it. Additionally, smart device companies could provide options to users to switch to no-internet mode versus internet mode (though this would not reduce the risks associated with internet connections when they are using internet-enabled functionality); when they disconnect from the internet, the device will still provide some functions rather than being rendered useless. This could be a useful option for users who only want a basic level of functionality from a smart home device – a device smart enough to get a B but not necessarily an A+. Apple has enabled this with Siri by processing as many functions on device as possible (it is clear from the number of times Apple has come up in this section that it is on the more privacy-preserving end of the smart home spectrum).

7. Policy landscape and recommendations

In section 5, I discussed the privacy concerns that arise relating to smart devices, and in section 6, I laid out and discussed technical suggestions that might be implemented to counter them. However, the reason that these technical suggestions may not currently be implemented by some smart device companies is hardly because they do not know about them. While there are certainly trade-offs that come with some technical choices (slower processing time, a need for more storage, less accurate predictions, etc.), this is not the only thing stopping smart device companies from making their products more privacy-preserving. The fact is that while they do not *have* to, smart device companies are unlikely to prioritize privacy over profit – this has been shown time and again by the track records of most major technology companies that provide smart devices. This is where policy and regulations relating to user privacy come into play. Companies must be incentivized or otherwise forced to uphold privacy protections by laws and held accountable if they do not.

Additionally, some privacy concerns are not necessarily or only technically addressable. For example, ensuring that users have rights over their data is not technically difficult, and moreover users not being able to control their data is not a natural consequence of smart device technology the way that learning about users is. This is more a matter of the government enshrining specific rights in law and then enforcing them.

Thus, comprehensive regulations that consider the growing popularity and danger of smart home devices are a necessity. This kind of formal privacy protection in the US is lackluster at best. Therefore, in this section I analyze the current state of US privacy protection regulations and, based on my analysis and the privacy concerns previously discussed, make policy recommendations to be used in tandem with the technical recommendations provided. In order to limit the scope of the paper, I focus on the US – to expand on it, I would also discuss international privacy protection regulations like Europe's General Data Protection Regulation (GDPR) in order to compare 'weak' and 'strong' privacy protections.

7.1. Current US policy landscape

7.1.1. Sector-specific federal laws

There is no single, comprehensive federal legal framework in the United States that regulates the collection and use of consumer data (Tuohy). In 2022, the US came as close to it as they ever have been when the American Data Privacy and Protection Act (ADPPA) was proposed. The crux of the act was data minimization, which would have marked a shift from the current standard of consent-based privacy (think constant and irritating pop-ups asking users to accept or decline cookies) by simply telling companies not to collect more data than reasonably needed, and providing a list of 17 permitted reasons why they might need data, including user authentication and fraud prevention (Edelman). While the act would still have allowed targeted advertising, which is ultimately the "economic driver" for most invasive data collection, it would have placed stricter restrictions on it than any existing US law that would have drastically reduced its invasiveness, and the restrictions significantly would have applied to operations across the US (Edelman). However, the law was not passed by the Congress at the time, due to concerns from states like California that it would "preempt" state privacy laws (Edelman). As a result, data is largely unregulated on a federal level.

This is not to say there are no laws regarding data management in general – a number of individual states have their own data privacy laws, and in recent years the rate of adoption of such laws has increased. However, the lack of a federal law or framework means that companies that collect data, including smart device companies, only need to protect data or otherwise follow data privacy laws for states that have their own laws – the data of a California resident is treated differently from the data of an Illinois resident, because California has its own privacy law and Illinois does not. As a result, users of smart devices are not granted equal protection or rights. In addition, the majority of US states do not have their own data privacy laws, meaning that companies can collect and use much data from residents of those states' in almost any way they desire without notifying residents (Klosowski). A federal law would ensure that users are not dependent on their state of residence for privacy.

This is also not to say that no federal laws governing data management exist. A number of federal laws cover particular sets of data, such as medical data, data pertaining to children, or data used by specific entities, and regulate within these sets ("Consumer Data Privacy Laws"). Yet these laws mislead citizens into believing that much more of their data is protected than it really is. Here is a non-exhaustive list of such laws, with brief descriptions of what they cover and how:

- *Health Insurance Portability and Accountability Act of 1996 (HIPAA):* This prevents the disclosure by "covered entities" of patients' sensitive health information without their knowledge and consent; it also requires covered entities to provide secure communication channels for discussing sensitive health information (CDC). Covered entities include healthcare providers like doctors, pharmacies, and hospitals, and health plans like health insurance (CDC). However, this is not at its core a privacy act – even the 'P' in HIPAA does not stand for 'privacy'. HIPAA specifically regulates information collected when a person seeks health care, that is, during their communications with covered entities (Morrison). Therefore, contrary to popular belief, all health data is not provided with federal privacy protection. Data collected from smart devices like FitBits or smart health monitoring devices, web searches on Alexa for medicines, GPS data pinpointing someone's location to an ER – it seems, in fact, a majority of the health data collected or inferred by smart devices – are not covered by HIPAA. Additionally, since HIPAA only applies to the aforementioned covered entities, smart device companies are not bound by HIPAA in any way except when interacting with the covered entities due to requirements on the other side.

- *Family Educational Rights and Privacy Act of 1974 (FERPA)*: This protects the privacy of students by limiting access to information from student education records and allowing students to inspect and amend their records (access is controlled by parents for students under 18 and by students once they are either 18 or begin postsecondary education, whichever comes first) (U.S. Department of Education). Schools cannot disclose information from report cards, transcripts, disciplinary records, and more without consent except in specific outlined circumstances. However, FERPA does not prevent the disclosure, without consent, of "directory" information, such a student's name, contact details, attendance dates, age, and awards; it also does not apply to education-related information that is derived from any source other than educational records, even if that

information is in the educational records ("FERPA"). Therefore, any data obtained by schools through inference or observation does not fall under FERPA.

- *US Privacy Act of 1974*: This governs the "collection, maintenance, use, and sharing" of individuals' information kept in systems of record by all federal agencies (Office of Private and Civil Liberties, "Privacy Act"). Namely, this gives individuals the right to request and view their records, request corrections in their records, and be protected against invasions of their privacy resulting from federal agencies having their information (U.S. Office of Special Counsel). Federal agencies cannot share information without consent, except in the case of certain exceptions. However, there are a number of exceptions that undermine the promised protection of the act. The "routine use" exception states that a federal agency may disclose information "for a routine use as defined in subsection (a)(7) of this section"; subsection (a)(7) defines "routine use" as "the use of such record for a purpose which is compatible with the purpose for which it was collected" (U.S. Office of Special Counsel). Not only is this vague, but leaves room for federal agencies to distort the original purpose of collection such that they have new 'compatible' purposes for sharing data that are not really the same as the original purpose. Additionally, the act narrowly defines what kind of information is protected. "Systems of record" refer to groupings of information under the control of any federal agency, retrievable by a personal identifier such as the individual's name, SSN, or biometric data (U.S. Office of Special Counsel). The key word here is 'retrievable' – any information in a database which is not specifically *retrieved* using an individual's personal identifiers is not subject to the law, even if it contains personal identifiers or other sensitive information, and there are likely many such databases that do not fit the official definition of a system of record. In a lawsuit on the matter, it was ruled that the disclosure of information "acquired from non-record sources – such as observation, office emails, discussions with co-workers and the 'rumor' mill – does not violate the Privacy Act… even if the information disclosed is also contained in agency records" (Office of Privacy and Civil Liberties, "Overview").

- *Electronic Communications Privacy Act of 1986 (ECPA)*: This protects wire and electronic communications (including email and phone conversations) while they are being made, while in transit, and while stored on computers (it was originally created to limit government wiretapping on electronic communications) (Bureau of Justice Assistance). However, while the ECPA has been updated somewhat, it still does not protect electronic communication data of all ages and natures from law enforcement – older data stored on servers or the cloud, as well as data from search queries, is not subject to the ECPA. Thus this data is still susceptible to being used by governments for surveillance purposes.

- *Video Privacy Protection Act of 1988 (VPPA)*: This was created to prevent the "wrongful disclosure of video tape rental or sale records", including "personal information divulged and generated in exchange for receiving services from video tape service providers" ("18 U.S. Code"). It has been interpreted such that it is applicable to the Internet Age due to the definition of a "video tape service provider" – "any person, engaged in the business … of rental, sale, or delivery of prerecorded video cassette tapes or similar audio visual

materials" ("18 U.S. Code"). "Similar audio visual materials" has been interpreted by courts to mean online videos on websites and applications (not including livestreaming), and streaming services. However, the fact that the law hasn't been specifically updated leaves room for argument on the part of streaming services when they are accused of violating user privacy. Additionally, the VPPA makes an exception regarding sharing "subject matter" data, meaning vaguer data like video genre or category, if it is shared for "the exclusive use of marketing goods and services directly to the consumer" ("18 U.S. Code"; T. Peterson). Therefore not only does the law specify that data can be shared without consent for targeted advertising, but does *not* specify whether this advertising must be on the service provider's platform, or from third parties. If the data is shared with third parties for advertising, are they also subject to the VPPA, given that they are not video service providers themselves?

- C*hildren's Online Privacy Protection Act of 1998 (COPPA)*: This restricts companies', websites', and online services' data collection on children under 13 years of age (FTC, "COPPA"). Parental consent is required for the collection and use of a child's personal information. However, there are concerns around whether parental consent is the best way of protecting child privacy, and whether data collection on children should be restricted as a blanket rule (Miller); additionally, especially with the increasing pervasiveness of the internet and smart home devices, it seems that collateral damage might increase in cases where an individual's age is harder to verify or keep track of (for example, data collected by smart home devices that record surrounding audio and video).

- *Graham-Leach-Bliley Act of 1999 (GLBA)*: This requires consumer financial products (banks, investment services, etc.) to explain both how they share data and that customers have a right to opt out of this data sharing (Klosowski). It also attempts to put into law security requirements, stating that these companies must make sure that customer data is secure, protect against "anticipated threats or hazards to the security or integrity of such records", and protect against unauthorized access that might cause harm to the customer (Edelman). However, companies simply have to protect stored data and disclose how they use it – the law does not actually dictate or restrict the manner of using data, meaning customer data can still be used by the companies that collect it in ways that might invade their privacy.

Thus, it is clear that these subset-specific federal data laws still leave much to be desired. Their definitions of what kinds of data are covered and who is subject to the law are too narrow, allowing much data to go unprotected and many types of companies and third parties to escape their jurisdiction. While it makes sense that different kinds of data might require different levels of regulation, section 5 showed how revealing even innocuous data can be and the many ways in which it can be exploited against users – given this, *all* companies should have to protect *all* data to a baseline level, and users should have control over all of their data as well. Also, disparate laws are confusing – not just for the individual, who finds it difficult to understand their rights, but for companies to follow themselves.

Finally, many of these laws are outdated and do not take into consideration IoT age we currently live in – IoT and smart devices in fact go unmentioned. There is also no comprehensive federal IoT cybersecurity framework or regulation in the US. The 2020 Cybersecurity

Improvement Act sets some security standards for IoT devices, but applies only to the federal government and not the private sector ("IoT Cybersecurity"); thus the smart devices that most consumers use, including smart home devices, are not subject to this.

7.1.2. Federal Trade Commission

It is worth discussing the powers of the FTC and the Federal Trade Commission Act (FTC Act) in more detail – while still not as strong a true federal privacy law, the FTC Act grants the FTC the authority to enforce privacy and security regulations in the US, and it has used this authority consistently and flexibly as consumers have been subjected to increasing data collection.

The FTC Act concerns government or agency enforcement of companies in the commercial sector. Section 5 of the FTC Act prohibits "unfair or deceptive practices in or affecting commerce" ("Section 5" 1). Unfair practices refer to those that cause "substantial injury to consumers", can't be "reasonably avoided by consumers", and are not outweighed by any benefits provided; deceptive practices refer to those where a "representation, omission, or practice" is likely to mislead consumers, a consumer's interpretation of the "representation, omission, or practice" is reasonable, and the misleading is material ("Section 5" 1). This is the origin of the FTC's enforcement of privacy and security regulations – it has used its general regulatory authority to protect consumers against these 'unfair' and 'deceptive' practices by punishing companies that violate consumer privacy and requiring them to improve their privacy and security practices. The FTC can also enforce many sector-specific laws such as COPPA and GLBA, as mentioned in section 7.1.1 (FTC, "FTC 2023" 3).

The FTC has a strong track record and has brought numerous cases concerning data privacy and security, including cases involving smart home devices specifically. For example, it charged Ring with illegally surveilling customers even in private areas of their homes such as bathrooms, and not implementing adequate access or security measures to prevent unauthorized viewing of recorded videos, livestreams, and consumer account data; as a result, a federal court ordered Ring to delete any data products (data, models, etc.) derived from unlawfully taken videos, implement a robust privacy and security program including multi-factor authentication, and pay $5.8 million to affected consumers (FTC, "FTC 2023" 5). The FTC also accused Amazon of violating both COPPA and the FTC Act by "indefinitely retaining" Alexa recordings of children's voices, and failing to uphold its promise to delete voice and geolocation data upon request; Amazon is now required to delete inactive accounts and certain voice data of children, cannot use that data to train algorithms, has to provide a strong privacy program, and pay a $25 million penalty (FTC, "FTC 2023" 5). In general, orders to companies have required actions like the implementation of comprehensive privacy programs, disclosure of violations to consumers, monetary damages paid to consumers, and the deletion of any data or data products illegally, unfairly, or deceptively gained or held (FTC, "FTC 2023" 3).

As previously mentioned, the FTC Act is not as strong as a federal privacy law would be. The derivation of authority from Section 5's prohibition of 'unfair or deceptive' acts means that the FTC is often limited to punishing companies only when they violate something stated in their privacy policy – poor data security or invasive data practices that violate privacy but are not explicitly privacy violations and are not deemed 'unfair' are not covered. It also doesn't provide consumer with a private right of action to sue if a company is unfair or deceptive. However, it is

still significantly better than the existing sector-specific laws, because it applies generally to goods and services in commerce rather than to a limited definition of data types and companies.

In addition to enforcement, the FTC provides privacy and security guidance for IoT companies on how to approach design security; recognized security practices to use, such as encryption, and standard policies to follow; authentication and access control methods; secure data management, including data minimization and security reviews; how to monitor and address security risks; how to create a "culture of security", including vetting third-party service providers to ensure they meet the same standards; and how to transparently and creatively communicate with users about privacy and security (FTC, "Careful Connections"). These are only recommendations and hold no actual legal force, but given that the FTC has the level of regulatory power that it does and is constantly calling for them to be enacted as comprehensive legislation, companies might be more inclined to follow them. It also mandates some of its restrictions as consequences when taking action against companies violating Section 5.

Given the limitations of the FTC Act, it still seems that comprehensive federal legislation is needed – the FTC itself has been calling for this. However, the FTC and FTC Act currently provide the strongest privacy protection in the US.

7.1.3. State laws

As mentioned, several US states have enacted their own data privacy laws, applicable only to residents of those states. California was the first to do so in 2019 with the California Consumer Privacy Act (CCPA); currently, 15 states have enacted comprehensive data privacy laws – California, Colorado, Connecticut, Delaware, Indiana, Iowa, Kentucky, Montana, New Hampshire, New Jersey, Oregon, Tennessee, Texas, Utah, and Virginia (Pittman; Folks). Only the California, Colorado, Connecticut, Utah, and Virginia laws are effective as of today; the rest will respectively become effective between 2024 and 2026 (Pittman; Folks). This rate of adoption of state laws in the last few years suggests that many more states will introduce their own laws, especially if there continues to be no federal legislation.

Broadly, the issue with having many disparate laws across states is similar to the issue with data type-specific laws, especially when considering the fact that states may also have state-specific laws that cover only specific data types, like Illinois' Biometric Information Privacy Act (BIPA). Too many of these cause confusion for individuals, who may be unsure what their rights are and where they apply, and for companies, who must tailor their privacy policies to work in a number of different states according to what they require. A federal law might limit companies in ways they do not particularly like, but it provides a much clearer standard and makes enforcing data privacy laws less complicated as well.

Separately, existing state laws do not necessarily provide strong enough protections. I will not discuss all 15 in detail, but below I look into two examples – one considered strong and the other weak.

- *California Consumer Privacy Act (CCPA)* and *California Privacy Rights Act (CPRA)*: California is considered to have the strongest privacy protections in the US. It provides consumers with several key rights, including data access and transparency; data deletion; opt-out from the sharing and sale of data with third parties; and non-discrimination, that is, customers choosing to opt-out or exercise other outlined rights cannot be denied or

given lower quality products (California Consumer Privacy Act). Companies must also secure data adequately.

The acts certainly have many strengths. They apply to a broader set of data than simply personally identifiable data, thus providing greater protection. The CCPA also requires a global or universal opt-out that allows users to opt out of data sharing or targeted advertising in one click across devices, sites, or browsers (Klosowski); this is important because users often don't exercise their opt-out rights because companies make it time-consuming or annoying for them to do so. Also, the CPRA added a limited "private right of action", or the ability for an individual to sue a company in some cases of data breach, addressing a major limitation of the CCPA (California Consumer Privacy Act). Additionally, there is a specific connected devices section that imposes a "cybersecurity design requirement" on IoT, stating that connected devices must have "reasonable security features" (Dempsey and Carlin 413).

However, there are still gaps to be filled. A number of exemptions weaken the protections promised – for example, companies can deny a user's data deletion request if the data is "necessary" for certain functions, some of which are quite vague and leave room for exploitation (Viljoen). There could also be stronger provisions for data minimization. Also, the provision of opt-out rather than opt-in features still places the burden of privacy protection on the user.

- *Virginia Consumer Data Protection Act (VCDPA)*: The VCDPA provides consumers with the rights to data access, correction, deletion, portability, and appeal, and opt-out (from the sale of their data, targeted advertising, etc.); companies must also limit data collection and use and adequately secure data ("17 new privacy laws").

  However, this bill was written with "strong input" from Amazon and other industry lobbyists (Klosowski). This is contrary to its very purpose, and therefore it has a number of limitations that mean it does not really serve consumers as it should. Due to lobbyist interference, many of the protections provided by the bill are "business-model affirming", meaning they essentially allow companies that gather large amounts of data to carry on as normal (Klosowski). Not only is it an opt-out bill, putting the burden of privacy protection on the user, but users must opt-out individually from every site or entity that has their data; this is also true for data deletion (Cross). Consumers also have no private right of action, limiting their ability to react to violations. Also, the 'personal data' that the bill covers does not include de-identified data, which as discussed is relatively easy to reidentify.

Unfortunately, most states have passed laws more similar to Virginia's than California's. Common limitations across state laws are: no or poor provisions for data minimization; no private right of action; they prevent data from being "sold" but not shared, but do not make it clear that they are effectively the same thing; exclusion of de-identified or pseudonymized data from protection; exemption of some companies, like financial or health institutions; no universal opt-out. Many of these laws essentially allow companies to continue collecting and using data almost exactly the way they did before, as long as they notify consumers, who are helpless to do anything about it.

This is not to say that state laws are not useful. Even a few good ones might influence companies to apply stronger protections across the board, simply because it's less complicated.

However, the more likely scenario, given companies' profit-making motive, is that they will continue to violate consumer privacy where they can. Where state-specific data privacy laws are lacking, a federal framework could provide a baseline of privacy protection that states can add to or accept as is.

7.1.4. Proposed regulation

After two years of little to no action on the federal privacy front, the American Privacy Rights Act (APRA) was introduced on April 7th, 2024, and described by the senators who introduced it as "the best opportunity we've had in decades to establish a national data privacy and security standard that gives people the right to control their personal information" (Energy & Commerce). While there is no guarantee that this will be passed – indeed, the track record doesn't inspire optimism – its strengths and weaknesses are still worth discussing.

The draft APRA outlines privacy protections for the following broader goals: stronger data privacy rights and control for users; better enforcement; protection of civil rights; and data security. For the first, it mandates data minimization, requiring companies to only collect, store, process, and otherwise use data as necessary to provide their product and fulfil other listed purposes; user rights to access, correct, and delete data; opt out rights for data sharing and selling, and the right to opt out from having data used at all; opt out rights for targeted advertising; and opt *in* consent for the sharing of sensitive data and processing of some sensitive data. For the second, it provides consumers with the private right of action so that they can sue companies or other actors who violate their privacy rights; prevents companies from using mandatory arbitration, where users have waived their rights to sue in case of some violation, in the case of "substantial" privacy harm; and authorizes the FTC, states, *and* private consumers to sue. For the third, it mandates that companies do not use personal information to discriminate against consumers; companies cannot provide lower quality products or services, charge differently, or deny products to consumers who exercise the aforementioned rights; and users can opt out of the use of algorithms that use data to make decisions regarding their employment, insurance, credit, healthcare, housing, and more. For the fourth, it requires strong data security in order to mitigate hacking and other unauthorized access to data; notification of consumers when their data is shared with a "foreign adversary"; it also puts the responsibility on company executives to ensure that companies meet the requirements.

It is clear that the APRA is comprehensive. Its provisions certainly directly address many of the privacy concerns brought up in section 5 and fill in many of the gaps left by current state laws, mentioned in section 7.1.3. However, there are a few significant areas that I believe could still be improved.

First, while the APRA covers more data than any existing law, it could still cover more. The APRA defines sensitive data as including individuals' online activity, such as cross-site and social media tracking data, which is significant because of how much this data is used for targeted advertising (Quinlan). However, it is becoming easier and easier for seemingly innocuous data to generate sensitive and revealing insights about a person – "any data can be sensitive depending on how it used" (Trujillo). Any comprehensive privacy law should attempt to account for this. Additionally, de-identified data is not counted as personal data. This means that companies can take personal data protected by the APRA, de-identify it, and then use it freely. Not only is this risky given the ability to re-identify data, but something that might make

consumers uncomfortable if they opted out of having their data used for training models, sharing, etc.

Second, while the APRA including a private right of action is a large step in the right direction, it is too limited. The private right to action does not apply to all rights named. For example, private consumers cannot sue if the data minimization mandate is violated (Trujillo). Without enforcement, laws are easy to ignore.

Third, and very importantly, the APRA has no provisions or restrictions for data collection and use by government entities – federal, state, tribal, territorial, or local (Trujillo). Government contractors – anyone using the data to provide some service to the government - are completely exempt from the bill (Quinlan). Consumer data is not automatically safe from violation because the government is using it – there should be specific and strong protections in place to mitigate government violations and surveillance concerns as well.

7.2. Recommendations

Having analyzed the current landscape of privacy regulations in the US, I now make recommendations for a truly comprehensive law that actually addresses the privacy concerns associated with smart home devices and does not undermine them with myriad exceptions. These are recommendations for a federal framework that specifically covers IoT devices – I believe that their increasing adoption warrants a framework that focuses on them. However, some recommendations might apply to broader data privacy as well.

My recommendations come under two categories – regulations and enforcement.

7.2.1. Regulations

Privacy regulations must cover the following areas: data collection, data access and use, data security, device (or website, application, etc.) security, and user control.

7.2.1.1. Data collection

First, and very importantly, companies must be required to follow data minimization practices. The large amount of data collected and thus stored, becoming a target for hackers, is the one of the primary reasons privacy concerns exist – collecting less data directly addresses this. This might look like providing companies with a list of purposes for which they can collect data and prohibiting any other data collection, as the APRA proposes – for example, Ring doorbells might need to collect some amount of video data to train facial recognition models if users want facial recognition functionality but don't need users' social media handles. This could also look like providing a baseline list of purposes for which companies can collect data, but allowing consent-based collection of additional data beyond that.

Each option has its pros and cons. Limiting data collection according to a strict list means that consumers don't have to endlessly click "yes" or "no" on various pop-ups asking for consent to collect data or track users, which can be tiring, but smart device companies have less data to train their models on, making them less useful to the user; allowing additional consent-based data collection helps companies provide stronger products without making users who don't wish to share additional data uncomfortable, but could be more annoying as well. Having the default

be the minimum level of data collection, however, means that only users who choose to consent to additional data collection have to deal with consent forms and pop-ups.

Additionally, limited on-device processing, thus the collection or use of as little data as possible, might be mandated. This is in acknowledgement of the fact that some functions are too computationally heavy to be done locally; however, technology groups can determine a number of specific purposes where this is feasible. For example, asking a voice assistant to read messages does not require the message contents to be uploaded to the device's cloud servers. This may seem like a minute detail – however, where smart devices are concerned, any minimization of data collection is beneficial.

7.2.1.2. Data access and use

Rules must also be set for how companies use data after it is collected. Importantly, all data should be covered by the regulations recommended. Existing laws have extremely narrow scopes that do not account for the fact that those with access to data can combine it with external data to generate sensitive insights. Significantly, de-identified data should not be treated differently from personal data. The current landscape allows companies to de-identify data, say that privacy laws no longer apply to it, and use it as they wish, often without consent. But de-identification has been shown to be reversible; de-identified data must be treated as sensitive accordingly.

As a default, *smart device companies* should not be able to use user data for anything but necessary purposes for the functioning of the device, user authentication, and other essentials. Some users are even uncomfortable with the thought of their data being used to train models and algorithms for smart devices or other company products. Collected data should also not be used to build consumer profiles to be shared or used for advertising. Moreover, companies should not be able to combine collected data with external data from data brokers for this purpose. Smart device companies should also specifically not be able to use consumer data inferences for targeted advertising of their own or other products on the device that collected it or on other connected devices or applications.

When it comes to sharing data, a comprehensive framework covers both third parties and the government:

Third parties should not have access to any data without user consent or for a specific functionality of the smart device; personal data (the broader definition of it) should not be shared at all. This means companies cannot sell data to third parties, but also cannot share data with them except that data which is necessary for the functioning of the device as specified. The distinction between selling and sharing is important. Selling data refers to the exchange of data for money or, as defined by some states like California, anything else valuable – for example, providing personal information to an advertising company in exchange for advertising insights (Stroink-Skillrud). Sharing data refers to any provision of data to a third party, including situations where the third party is integrated for functionality purposes, as is often the case in smart home devices. Many privacy laws explicitly ban data *selling*, but don't limit or regulate sharing, though shared data can still be used by third parties in ways that violate privacy. Therefore, third-party companies who are given access to data must use it for only specified purposes necessary for the device and not for targeted advertising. (Something to note is that both for smart device companies and third parties, I do not say that targeted advertising should be completely banned, as this is unrealistic and as it is often helpful to users to receive ads

somewhat tailored to them. However, targeted advertising should only be based on first-party data collected by the company itself unless a user opts into it. For example, if a consumer buys a chair using Amazon Alexa, they might be given ads for chairs on Amazon on their laptop – this is targeted advertising based on first-party data collection. However, if the smart device data is shared with IKEA for some reason and the user suddenly receives tons of ads for chairs from IKEA, or if Amazon infers from combined data that the user has a back condition that they did not disclose and begins to recommend medicines, this is no longer based on first-party data and can begin to feel uncomfortable. This reduces the creepiness factor of targeted advertising, as consumers are less likely to be startled by what their device and other third parties know about them. I mention that a user can opt into it more specific or inference-based targeted advertising – this is to acknowledge that there are users who might want this service, not find it invasive, or weigh its benefits as greater than any perceived creepiness.) Finally, any third-party with whom data is shared should be subject to the same privacy regulations, and the specific privacy policy and standards of the smart device company that shared the data. This ensures that consumers' privacy is not less protected when data is shared.

Governments (state, federal, local, etc.) should similarly be subject to the laws of the federal framework when given access to collected data from smart devices. A growing concern about smart devices is that they are creating a surveillance culture. Therefore, governments should not have free or unregulated access to smart device data. A court order or warrant should be required for a company to share data with any government agency or contractor, and consumers must be notified when their data is shared unless this is shown by the agency in question to be contrary to the purpose for requesting the data, that is, if the user being notified hampers the investigation in some way. The only situation in which a warrant may not be required is that in which a user consents to sharing information with the government, either independently or in response to a request - this might be the case in situations where the user is the victim of a crime themselves and wants to provide data to have the case solved as quickly as possible.

Then, after the data is reviewed for whatever government purpose, it must be deleted or otherwise removed from government access. This is crucial to prevent data from being used against consumers in ways unrelated to the purpose for which it was originally collected, if not immediately then in the long run as context and political situations evolve.

7.2.1.3. Data security

In terms of data security, strong security requirements must be mandatory. While specific methods of securing data might be up to companies to choose, data should be secure at all times – when stored, when in transit, *and* when in use (as discussed in section 6.3.1.2). This is to minimize the amount of time when collected data is vulnerable to hackers or other malicious agents, with the acknowledgement that anonymization and pseudonymization alone are not protection enough.

To start with, encryption of all data, and end-to-end encryption where possible, should be mandatory – in storage, in transit, in use. Cloud servers, whether owned by the smart device company or third parties, should meet the specified security standards as well. I also mentioned that third parties should be subject to the same privacy rules, which includes data security. Some responsibility can be put on smart device companies to ensure that third parties they integrate with are capable of meeting the data security requirements. This would add a double layer of

protection, as smart device companies cannot simply pass the blame for a third-party violation of privacy to the third-party while benefiting from their relationship with it.

Finally, it should be mandatory for companies to implement strong access control to prevent data breach incidents or misuse of data from within.

7.2.1.4. Device security

Since I am making recommendations for smart device regulation specifically, I must mention security measures of the device itself and not just the data collected.

Smart devices must have sufficiently secure login methods (both on the device and on companion applications or websites). Drawing from the CCPA, a connected device should be equipped with a preprogrammed password unique to the device, and require a user to generate a new password or means of authentication when they set up the device to use for the first time (Dempsey and Carlin 13). Passwords should meet strength requirements (length, mix of alphabets and numerals, special characters) and when resetting a password, a user should not be able to use any previous passwords.

Smart devices should also require the user to set up multifactor authentication. For devices that can be accessed by third parties (say, smart thermostats accessed by utility companies for demand response programs), access limits or access control should be put into place so they only access what is necessary for their function.

Finally, smart device companies can implement automatic security patches as vulnerabilities are found and addressed, rather than relying on users to update their device, to ensure that users don't miss security updates (accompanied by notification of users).

7.2.1.5. User rights

A federal framework must crucially enshrine user rights to data privacy – specifically, user consent to data practices and user control over collected data.

First, all data collection, processing, sharing, and other use must be done with the explicit consent of the consumer, who must be sufficiently aware of what they are consenting to. Ideally, this is opt-*in* instead of opt-out consent. With opt in consent, the default is maximum protection for the consumer – they can choose to opt in to specific less protective data practices if they so desire. This transfers the burden of privacy protection from the user, where existing opt-out systems place it, to the smart device company. It reduces the risk of invasive practices falling through the cracks because a user forgot to opt out. It also reduces fatigue caused by having to opt out of hundreds of different practices on many different applications and devices, which often causes users to give up and simply allow companies to do what they want. Therefore even if opt-out is implemented, it should at least be universal opt-out, so users can decline things like targeted advertising or tracking in one click across browsers or devices; the opt-out should be lasting until the user actively opts back it.

Consent should also be verifiable, revocable (at any time), and informed – companies must be transparent in their privacy policies and notify users in case they are changed so they can update their decisions. Of course, companies should not be allowed to discriminate against users or withhold services due to their choices regarding the use of their data. Often consumers are told that if they opt out of data use, they will not get the full functionality of the device, or that if they

don't sign terms and conditions that effectively ignore their privacy, they can't use the device. These are all forms of coercion that make any "consent" meaningless

Second, users should remain owners of their data even after it is collected. This means they should have ultimate control over it and be able to easily view, update, and delete it. Furthermore, corrections or deletions requested from the smart device company should be applied to third parties with the data as well.

7.2.2. Enforcement

Laws without enforcement are no better than suggestions, and companies will likely not take them.

7.2.2.1. Severity of penalties

Punishments for violations must also be significant enough that they actually deter companies from breaking rules. Fines must be substantial enough (based on the company size and track record) that they actually affect the perpetrator. Otherwise companies who can afford it are effectively paying for the right to do illegal things rather than being punished for it and have little incentive not to repeat the crime (this is especially true for technology giants like Google and Amazon who provide smart devices only as one subset of their overall goods and services, and generate massive revenue). Europe's GDPR lays out severe penalties – maximum fines per violation are the larger of 4% of a company's global turnover or $20 million (Cross).

Limiting curing periods might also force companies to comply. Some suggest curing periods be allowed for first offenses – a sort of second chance policy. However, this means that until they get caught, companies can do whatever they want, as the first time they get caught they can avoid a fine or other punishment by curing the issue instead. Therefore, curing periods can be granted on a case-by-case basis rather than promised as a blanket rule.

Severe penalties might also move companies to adopt stronger privacy practices not mandated by the law in order to better avoid breaking rules – things like differential privacy, homomorphic encryption, and federated learning.

7.2.2.2. Who can enforce laws?

The two main forms of enforcement are government enforcement and private right of action; both are necessary.

In the US, the FTC and Attorney General's office are currently empowered to enforce existing privacy laws. However, it is imperative that consumers also have a comprehensive private right of action. Allowing only some government agencies to enforce laws is not efficient and allows many violations to go unpunished – the Attorney General's office says it can only tackle roughly 3 cases a year due to resource constraints (Trujillo). Also, and perhaps more importantly, consumers are the ones directly affected. It only makes sense that they should be able to sue for noncompliance.

8. Conclusion

It is clear that completely giving up privacy is not some inescapable 'cost' of using smart home devices. In this thesis, I provided a framework for how make smart home devices that are still beneficial to individuals but preserve privacy better by explaining how they work, how and why the subsequent privacy concerns arise, technical methods to address those concerns (and technical methods that are not enough), and policy recommendations to ensure the implementation of the above and protect consumer privacy from all angles.

I used smart voice assistants, smart thermostats, and smart doorbells as examples with which to explain the functioning of smart home devices and the subsequent privacy concerns that arise. Technologies like speech recognition, NLP, motion detection, geofencing, facial recognition, video analytics, and more learning algorithms, combined with the use of the cloud and Wi-Fi, enable individuals to run their homes and live their lives much more conveniently, efficiently, and safely. These technologies, and thus these smart home devices, rely on immense amounts of data about the user they are working for – data which is constantly transferred from sensor to microcontroller to server to local database to third-party application to more, as shown in sections 2-4. As a result, despite the many proven benefits that users gain from adopting smart home devices, smart home devices raise a significant number of privacy concerns, many of which are quite severe.

I categorized and described these privacy concerns in section 5. Smart home devices collect data, including personally identifiable data, from many sources – cameras, microphones, sensors, interactions, device setup and connected accounts, companion applications, integrated devices, and more. There are concerns about the inferences that can be drawn from this data, especially when it is combined with data available from external sources; the creation of scarily accurate user profiles, used for invasive and often unnerving targeted advertising; unauthorized or unwanted access to confidential data; data sharing with third parties, spreading user data around the world to entities that the user has not consented to share data with; government surveillance; hacking and database breaches that expose data to bad actors who can use it against consumers (for theft, stalking, blackmail, etc.); and spying through smart home devices within users homes. Additionally, there are concerns that consumers are no longer owners of their own data and cannot control it once it's collected – that by deciding to use smart devices, they will never have privacy again.

Having established the privacy concerns, I categorized and analyzed technical methods for mitigating them, showing that privacy protection and smart home devices do not have to be mutually exclusive. There are a number of technical strategies that, if implemented, enable smart home device companies to make smart home devices more privacy-preserving without compromising meaningfully on functionality. These include methods to reduce data collection without abandoning prediction accuracy (data minimization, synthetic data, federated learning); methods to improve data security so there is less risk of storing, using, and sharing (encryption in use, differential privacy, on-device processing); and methods to improve device security (offline processing). Using them in combination with each other is even more promising – for example, using asymmetric encryption of symmetric keys that are stored locally in a smart device with strong security measures against hacking.

Technical methods are not enough alone. An effective framework for improving privacy preservation in smart devices needs to address both technical solutions *and* policy. This is because companies may not implement existing technical methods without incentive, not

because they don't care about hackers or data leaks but because they generate a lot of profit from using data in ways that violate privacy, and are unlikely to prioritize privacy over profit without being made to. Furthermore, a significant number of privacy concerns (though not all) arise not as a consequence of some technology used in smart devices but because of the way that companies collect and manage data. Here, policy must play a role in protecting consumer privacy. Therefore I analyzed the current US policy landscape regarding smart home devices and consumer data privacy and provided subsequent policy recommendations to complete the framework.

Privacy regulations in the US need a lot of work. There is no comprehensive federal data privacy legislation; weak sector- and state-specific laws with too many exemptions and loopholes to meaningfully protect user privacy; and only the FTC with any real regulatory power. I therefore divided recommendations into two buckets. First, I recommended regulations – I categorized proposed regulations according to the categorization of privacy concerns in order to directly address them and layer them on top of the corresponding technical methods that can be implemented to meet them. I also made recommendations regarding user consent and control to address privacy concerns that are symptoms of companies' profit motives and not byproducts of smart device technology. Second, I made recommendations for enforcement, without which those who violate the aforementioned regulations cannot be held accountable – namely, setting severe enough punishments and giving consumers a private right of action so they can defend themselves.

Some might argue that such a framework makes it harder for smart device companies to do their jobs and hampers the functionality they offer to some extent. However, the trade-off seems worth it. The potential decrease in functionality will apply to all companies, therefore competition will still exist; it will potentially drive innovation leading to better privacy-preserving technology, which is a desired outcome. A clear standard that everyone must follow is also easier to abide than hundreds of unique privacy laws differing by state, data type, and entity. Ultimately, "privacy isn't about not using tech" (Klosowski). We should be able to live in smart homes – we just shouldn't have to give up our rights to do it.

Works Cited

Ahmed, Biddwan. "Chatbot vs conversational AI – What's the Difference?" *Yellow*, 2 Apr. 2024, yellow.ai/blog/chatbot-vs-conversational-ai/.

"Amazon Echo Studio." *Privacy Not Included*, 1 Nov. 2023, *Mozilla*, foundation.mozilla.org/en/privacynotincluded/amazon-echo-studio/.

"Amazon Ring Video Doorbell." *Privacy Not Included*, 1 Nov. 2023, *Mozilla*, foundation.mozilla.org/en/privacynotincluded/ring-video-doorbell/.

"Apple Homepod." *Privacy Not Included*, 1 Nov. 2023, *Mozilla*, foundation.mozilla.org/en/privacynotincluded/apple-homepod/.

"Arlo Video Doorbell." *Privacy Not Included*, 1 Nov. 2023, *Mozilla*, foundation.mozilla.org/en/privacynotincluded/arlo-video-doorbell/.

ARO. "IoT devices 'to generate nearly 80 zettabytes of data' by 2025. *ARO*, 13 Apr. 2023, aro.tech/iot-devices-to-generate-nearly-80-zettabytes-of-data-by-2025.

"Art. 4 GDPR." *EU General Data Protection Regulation (GDPR)*, 27 Apr. 2016, gdpr-info.eu/art-4-gdpr/.

Baterna, Quina. "What Is a Nest Thermostat and How Does It Work?" *Make Use Of*, 25 Aug. 2022, www.makeuseof.com/how-does-nest-thermostat-work/.

Bertoli, Emi. "Demand Response." *International Energy Agency*, 11 Jul. 2023, www.iea.org/energy-system/energy-efficiency-and-demand/demand-response.

Biddle, Sam. "AMAZON ADMITS GIVING RING CAMERA FOOTAGE TO POLICE WITHOUT A WARRANT OR CONSENT." *The Intercept*, 13 Jul. 2022, theintercept.com/2022/07/13/amazon-ring-camera-footage-police-ed-markey/.

Bitner, John. "Should you buy a wired or wireless video doorbell?" *DigitalTrends*, 16 Feb. 2024, www.digitaltrends.com/home/wired-vs-wireless-video-doorbell/.

Bolton, Tom, et al. "On the Security and Privacy Challenges of Virtual Assistants." *Sensors (Basel)*, vol. 21, 26 Mar. 2021. doi:10.3390/s21072312.

"Bose Smart Speakers." *Privacy Not Included*, 1 Nov. 2023, *Mozilla*, foundation.mozilla.org/en/privacynotincluded/bose-smart-speakers/.

Bureau of Justice Assistance. "Electronic Communications Privacy Act of 1986 (ECPA)." *U.S. Department of Justice*, bja.ojp.gov/program/it/privacy-civil-liberties/authorities/statutes/1285.

Burgess, Matt. "All the Data Amazon's Ring Cameras Collect About You." *Wired*, 5 Aug. 2022, www.wired.com/story/ring-doorbell-camera-amazon-privacy/.

---. "Google Home's data leak proves the IoT is still deeply flawed." *Wired*, 20 Jun. 2018, www.wired.com/story/google-home-chromecast-location-security-data-privacy-leak/.

California Consumer Privacy Act. California Legislative Information, 2018, leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5.

"Children's Online Privacy Protection Rule ("COPPA")." *Federal Trade Commission*, www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa.

Clauser, Grant. "Amazon's Alexa Never Stops Listening to You. Should You Worry?" *Wirecutter, The New York Times*, 8 Aug. 2019, www.nytimes.com/wirecutter/blog/amazons-alexa-never-stops-listening-to-you/.

"Committee Chairs Rodgers, Cantwell Unveil Historic Draft Comprehensive Data Privacy Legislation." *Energy & Commerce*, 7 Apr. 2024,

energycommerce.house.gov/posts/committee-chairs-rodgers-cantwell-unveil-historic-draft-comprehensive-data-privacy-legislation.
"Consumer Data Privacy Laws." *Bloomberg Law,* pro.bloomberglaw.com/insights/privacy/consumer-data-privacy-laws/#the-need-for-privacy-laws.
Crosling, Mark. "Video Analytics for Smart Video Doorbells – Here's Why You Need it." *Xailient*, 14 Oct. 2022, xailient.com/blog/video-analytics-for-smart-video-doorbells-heres-why-you-need-it/.
Cross, R.J. "Why lots of states are passing bad state privacy laws." *U.S. PIRG Education Fund*, 31 Jan. 2024, pirg.org/edfund/resources/state-privacy-laws/.
de Looper, Christian. "What is the Ring Video Doorbell and how does it work?" *Amazon*, 25 Apr. 2023, www.aboutamazon.com/news/devices/ring-video-doorbell.
Dempsey, James X., and John P. Carlin. *Cybersecurity Law Fundamentals*. International Association of Privacy Professionals (IAPP), 2024.
Denham, Hannah, and Jay Greene. "Did you say, 'Hey, Siri'? Apple and Amazon curtail human review of voice recordings." *The Washington Post*, 2 Aug. 2019, www.washingtonpost.com/technology/2019/08/02/apple-says-its-contractors-will-stop-listening-users-through-siri/.
"Dialogue Manager (DM)." *Hyro*, www.hyro.ai/glossary/dialogue-manager-dm/.
"Differential Privacy." *Harvard University Privacy Tools Project*, privacytools.seas.harvard.edu/differential-privacy.
Dwork, Cynthia, and Aaron Roth. "The Algorithmic Foundations of Differential Privacy." *Foundations and Trends in Theoretical Computer Science*, vol. 9, nos. 3-4, pp. 211-407, 2014, doi:10.1561/0400000042.
"An Early History of Comfort Heating." *The NEWS Magazine*, 6 Nov. 2001, www.achrnews.com/articles/87035-an-early-history-of-comfort-heating.
"Ecobee Smart Thermostat Premium." *Privacy Not Included*, 1 Nov. 2023, *Mozilla*, foundation.mozilla.org/en/privacynotincluded/ecobee-smart-thermostats-premium/.
Edelman, Gilad. "Don't Look Now, but Congress Might Pass an Actually Good Privacy Bill." *Wired*, 21 Jul. 2022, www.wired.com/story/american-data-privacy-protection-act-adppa/.
"18 U.S. Code § 2710 - Wrongful disclosure of video tape rental or sale records." *Legal Information Institute*, Cornell Law School, www.law.cornell.edu/uscode/text/18/2710.
"Encryption." *National Institute of Standards and Technology*, csrc.nist.gov/glossary/term/encryption.
Ertl, Bob. "What Is End-To-End Encryption & How Does It Work?" *Kiteworks*, 23 Jan. 2024, www.kiteworks.com/secure-email/end-to-end-encryption/.
"Eufy Video Doorbells." *Privacy Not Included*, 9 Nov. 2022, *Mozilla*, foundation.mozilla.org/en/privacynotincluded/eufy-video-doorbells/.
Federal Trade Commission. "Careful Connections: Keeping the Internet of Things Secure." Federal Trade Commission, Sep. 2023, www.ftc.gov/business-guidance/resources/careful-connections-keeping-internet-things-secure.
---. *The Federal Trade Commission 2023 Privacy and Data Security Update*. Federal Trade Commission, 2023, www.ftc.gov/system/files/ftc_gov/pdf/2024.03.21-PrivacyandDataSecurityUpdate-508.pdf.

"Federal Trade Commission Act – Section 5: Unfair or Deceptive Acts or Practices." *Consumer Compliance Handbook*, Jun. 2008, www.federalreserve.gov/boarddocs/supmanual/cch/200806/ftca.pdf.

Hindi, Rand. "Private Smart Contracts Using Homomorphic Encryption." *Zama*, 23 May 2023, www.zama.ai/post/private-smart-contracts-using-homomorphic-encryption.

"Family Educational Rights and Privacy Act (FERPA)." *U.S. Department of Education*, 25 Aug. 2021, www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html.

Federal Trade Commission. "FTC Says Ring Employees Illegally Surveilled Customers, Failed to Stop Hackers from Taking Control of Users' Cameras." *Federal Trade Commission*, 31 May 2023, www.ftc.gov/news-events/news/press-releases/2023/05/ftc-says-ring-employees-illegally-surveilled-customers-failed-stop-hackers-taking-control-users.

"FERPA: What it means and how it works." *Student Press Law Center*, splc.org/ferpa-what-it-means-and-how-it-works/.

Fingas, Roger. "How does Alexa work? The tech behind Amazon's virtual assistant, explained." *Android Authority*, 21 Dec. 2023, www.androidauthority.com/how-does-alexa-work-3209316/.

Folks, Andrew. "US State Privacy Legislation Tracker." *International Association of Privacy Professionals*, 22 Apr. 2024, iapp.org/resources/article/us-state-privacy-legislation-tracker/.

Garbar, Dzmitry. "How to develop a voice assistant like Siri." *Belitsoft*, 29 Jul. 2019, belitsoft.com/speech-recognition-software-development/how-to-develop-a-voice-assistant-like-siri.

Garfinkle, Alexandra. "Amazon has sold more than 500 million Alexa-enabled devices, drops 4 new Echo products." *Yahoo!Finance*, 17 May 2023, finance.yahoo.com/news/amazon-has-sold-more-than-500-million-alexa-enabled-devices-drops-4-new-echo-products-140013808.html.

Gatt, Albert, and Emiel Krahmer. "Survey of the state of the art in natural language generation: Core tasks, applications, and evaluation." *Journal of Artificial Intelligence Research*, vol. 61, 27 Jan. 2018. *AI Access Foundation*, doi:10.1613/jair.5477.

Gibbs, Samuel. "Nest Hello review: Google's smart facial-recognition video doorbell." *The Guardian*, 20 Sep. 2018, www.theguardian.com/technology/2018/sep/20/nest-hello-review-google-smart-facial-recognition-video-doorbell.

"Global Smart Doorbell Market Size to Worth USD 13.7 Billion by 2033 | CAGR of 15.6%." *Spherical Insights*, Feb. 2024, www.sphericalinsights.com/press-release/smart-doorbell-market.

Gonfalonieri, Alexandre. "How Amazon Alexa works? Your guide to Natural Language Processing (AI)." *Toward Data Science*, 21 Nov. 2018, *Medium*, towardsdatascience.com/how-amazon-alexa-works-your-guide-to-natural-language-processing-ai-7506004709d3.

Google. "Your voice & audio data stays private while Google Assistant improves." *Google Assistant Help*, 2024, support.google.com/assistant/answer/10176224?hl=en#zippy=%2Chow-google-assistant-improves-with-federated-learning%2Cturn-off-federated-learning-for-the-hey-google-model-delete-your-on-device-voice-recordings.

"Google Nest Learning Thermostat." *Privacy Not Included*, 1 Nov. 2023, *Mozilla*, foundation.mozilla.org/en/privacynotincluded/google-nest-learning-thermostat/.

"Google Nest Mini." *Privacy Not Included*, 1 Nov. 2023, *Mozilla*, foundation.mozilla.org/en/privacynotincluded/google-nest-mini/.
Grant, M. B. "What is geofencing and how do smart thermostats use it?" *Smart Thermostat Guide*, 21 Nov. 2018, smartthermostatguide.com/what-is-geofencing-and-how-do-smart-thermostats-use-it/.
Grind Success. "300+ Powerful Smart Home Automation Quotes to Inspire." *Grind Success*, 21 Nov. 2023, grindsuccess.com/smart-home-automation-quotes/?expand_article=1.
Hannemann, Anika, and Erik Buchmann. "Is Homomorphic Encryption Feasible for Smart Mobility?" Leipzig University Center for Scalable Data Analytics and Artificial Intelligence, 2 Jun. 2023, Germany, arxiv.org/pdf/2306.04195.
Hao, Karen. "How Apple personalizes Siri without hoovering up your data." *MIT Technology Review*, 11 Dec. 2019, www.technologyreview.com/2019/12/11/131629/apple-ai-personalizes-siri-federated-learning/.
Harsh. "The Rise of IoT: How Connected Devices Are Transforming Our World." *Medium*, 8 Jul. 2023, medium.com/@harsh247tech/the-rise-of-iot-how-connected-devices-are-transforming-our-world-b493a3159f0a.
Hassel, Kristin. "What Is End-To-End Encryption & Why Does It Matter?" *Private Internet Access*, 22 Dec. 2023, www.privateinternetaccess.com/blog/what-is-end-to-end-encryption/.
"Health Insurance Portability and Accountability Act of 1996 (HIPAA)." *Centers for Disease Control and Prevention*, 27 Jun. 2022, www.cdc.gov/phlp/publications/topic/hipaa.html.
Hern, Alex. "'Anonymised' data can never be totally anonymous, says study." *The Guardian*, 23 Jul. 2019, www.theguardian.com/technology/2019/jul/23/anonymised-data-never-be-anonymous-enough-study-finds/
---. "Apple overhauls Siri to address privacy concerns and improve performance." *The Guardian*, 7 Jun. 2021, www.theguardian.com/technology/2021/jun/07/apple-overhauls-siri-to-address-privacy-concerns-and-improve-performance.
Hernandez, Grant, et al. "Smart Nest Thermostat: A Smart Spy in Your Home." *University of Central Florida*, 2014.
"How Alexa works: Wake word." *Amazon*, www.amazon.com/b?ie=UTF8&node=23608571011.
"How Does a Doorbell Work? Parts and Components." *WM Henderson*, www.wmhendersoninc.com/blog/how-does-a-doorbell-work-parts-and-components/.
"Improving Siri's privacy protections." *Apple*, 28 Aug. 2019, www.apple.com/in/newsroom/2019/08/improving-siris-privacy-protections/.
"IoT Cybersecurity: regulating the Internet of Things." *Thales*, Jan. 2024, www.thalesgroup.com/en/markets/digital-identity-and-security/iot/inspired/iot-regulations.
Ireland, Corydon. "Alan Turing at 100." *The Harvard Gazette*, 13 Sep. 2012, news.harvard.edu/gazette/story/2012/09/alan-turing-at-100/.
Iyer, Srinivasan. "Overcoming design obstacles in video doorbells." *EDN*, 4 Nov. 2019, www.edn.com/overcoming-design-obstacles-in-video-doorbells/.
Jiřík, Pavel. "The Future of Voice Assistants." *Phonexia*, 25 Apr. 2022, www.phonexia.com/blog/the-future-of-voice-assistants/.
Klosowski, Thorin. "The State of Consumer Data Privacy Laws in the US (And Why It Matters)." *Wirecutter*, *The New York Times*, 6 Sep. 2021, www.nytimes.com/wirecutter/blog/state-of-privacy-laws-in-us/.

Kodé, Anna. "Unwanted Connection: Who Has Control of Your Smart Home?" *The New York Times*, 17 Feb. 2023, www.nytimes.com/2023/02/17/realestate/smart-home-devices.html.
Lu, Jiakang, et al. "The smart thermostat: using occupancy sensors to save energy in homes." *Association for Computing Machinery,* 3 Nov. 2010. *ACM Digital Library*, doi: 10.1145/1869983.1870005.
Lynskey, Dorian. "'Alexa, are you invading my privacy?' – the dark side of our voice assistants." *The Guardian*, 9 Oct. 2019, www.theguardian.com/technology/2019/oct/09/alexa-are-you-invading-my-privacy-the-dark-side-of-our-voice-assistants.
Magee, Steven. "Smart people avoid living in 'Smart' homes." *QuoteFancy*, 2024, quotefancy.com/quote/2719475/Steven-Magee-Smart-people-avoid-living-in-Smart-homes.
Markowitz, Judith. "Toys That Have a Voice." *SpeechTechMag*, 6 Mar. 2003, www.speechtechmag.com/Articles/PrintArticle.aspx?ArticleID=30031.
Mason, Paul. "We can't allow the tech giants to rule smart cities." *The Guardian*, 25 Oct. 2015, www.theguardian.com/commentisfree/2015/oct/25/we-cant-allow-the-tech-giants-to-rule-smart-cities.
Meola, Andrew. "IoT Healthcare in 2023: Companies, medical devices, and use cases." *eMarketer,* 12 Jan. 2023, www.emarketer.com/insights/iot-healthcare/.
Miller, Tracy C. "Protecting Children Online: Evaluating Possible Reforms in the Law and the Application of COPPA." *Mercatus Center*, George Mason University, 20 Feb. 2023, www.mercatus.org/research/policy-briefs/protecting-children-online-evaluating-possible-reforms-law-and-application.
Molla, Rani. "Amazon Ring sales nearly tripled in December despite hacks." *Vox*, 21 Jan. 2020, www.vox.com/recode/2020/1/21/21070402/amazon-ring-sales-jumpshot-data.
Morrison, Sara. "HIPAA, the health privacy law that's more limited than you think, explained." *Vox*, 30 Jul. 2021, www.vox.com/recode/22363011/hipaa-not-hippa-explained-health-privacy.
Moskvitch, Katia. "The machines that learned to listen." *BBC*, 15 Feb. 2017, www.bbc.com/future/article/20170214-the-machines-that-learned-to-listen.
"Motion Sensor Thermostats: Do They Save You Money?" *Carrier*, staycomfyminnesota.hubspotpagebuilder.com/blog/motion-sensor-thermostats.
Mutchler, Ava. "Voice Assistant Timeline: A Short History of the Voice Revolution." *VoiceBot*, 14 Jul. 2017, voicebot.ai/2017/07/14/timeline-voice-assistants-short-history-voice-revolution/.
Ng, Alfred. "The privacy loophole in your doorbell." *Politico*, 7 Mar. 2023, www.politico.com/news/2023/03/07/privacy-loophole-ring-doorbell-00084979.
O'Neill, Sarah. "Chatbots vs Conversational AI vs Virtual Assistants: What's the Difference?" *LXA*, 18 Oct. 2021, www.lxahub.com/stories/chatbots-vs-conversational-ai-vs-virtual-assistants-whats-the-difference.
Office of Private and Civil Liberties. "Overview of the Privacy Act: 2020 Edition." *U.S. Department of Justice*, 15 Dec. 2022, www.justice.gov/opcl/overview-privacy-act-1974-2020-edition/disclosures-third-parties.
Office of Private and Civil Liberties. "Privacy Act of 1974." *U.S. Department of Justice*, 4 Oct. 2022, www.justice.gov/opcl/privacy-act-1974.
Oracle. "What is IoT?" *Oracle*, www.oracle.com/in/internet-of-things/what-is-iot/.

Owen, Louis. “Intelligent virtual assistants – Differentiators and features.” *Yellow*, 1 Feb. 2024, yellow.ai/blog/intelligent-virtual-assistants/.
Pasternack, Alex. “Homomorphic encryption could revolutionize privacy – so what is it?” *Fast Company*, 30 Aug. 2022, www.fastcompany.com/91088808/is-innovation-in-americas-healthcare-system-a-pipe-dream.
Peterson, Hayley. “Wisconsin couple describe the chilling moment that a hacker cranked up their heat and started talking to them through a Google Nest camera in their kitchen.” *Business Insider*, 25 Sep. 2019, www.businessinsider.com/hacker-breaks-into-smart-home-google-nest-devices-terrorizes-couple-2019-9.
Peterson, Tim. “WTF is the Video Privacy Protection Act?” *Digiday*, 26 Jul. 2022, digiday.com/future-of-tv/wtf-is-the-video-privacy-protection-act/.
Pittman, F. Paul. “US Data Privacy Guide.” *White & Case LLP*, 25 Apr. 2024, www.whitecase.com/insight-our-thinking/us-data-privacy-guide.
Porter, Michael E., and James E. Heppelmann. “How Smart, Connected Products Are Transforming Competition.” *Harvard Business Review*, Nov. 2014, hbr.org/2014/11/how-smart-connected-products-are-transforming-competition.
“The Privacy Act of 1974.” *U.S. Office of Special Counsel*, osc.gov/Pages/Privacy-Act.aspx.
Quinlan, Keely. “The American Privacy Rights Act could undercut state privacy efforts.” *StateScoop*, 12 Apr. 2024, statescoop.com/american-privacy-rights-act-state-laws-data/.
Ramos, Diana. “Voice Assistants: How Artificial Intelligence Assistants Are Changing Our Lives Every Day.” *Smartsheet*, 26 Jul. 2021, www.smartsheet.com/voice-assistants-artificial-intelligence.
Ramotion. “Voice User Interface: Introduction, Benefits, and Trends.” *Ramotion*, 12 Feb. 2024, www.ramotion.com/blog/voice-user-interface/.
“Recital 26: Not Applicable to Anonymous Data.” *EU General Data Protection Regulation (GDPR)*, 27 Apr. 2016, gdpr-info.eu/recitals/no-26/.
Rhode Island Energy. “Make your smart thermostat work smarter with ConnectedSolutions.” *Rhode Island Energy*, www.rienergy.com/RI-Home/ConnectedSolutions/Thermostat-Program.
Rjaibi, Walid. “The next frontier for data security: Protecting data in use.” *SecurityIntelligence*, 18 Apr. 2022, securityintelligence.com/posts/next-frontier-data-security-use/.
Romero, Andrew. “ChatGPT has gone full virtual assistant with voice and image recognition.” *9To5Google*, 25 Sep. 2023, 9to5google.com/2023/09/25/chatgpt-has-gone-full-virtual-assistant-with-voice-and-image-recognition/.
“17 new privacy laws around the world and how they’ll affect your analytics.” *Piwik*, piwik.pro/privacy-laws-around-globe/.
“A Short History of Speech Recognition.” *Sonix*, sonix.ai/history-of-speech-recognition.
Stelitano, Luisa. “A quick and easy guide to voice assistants.” *Miquido*, 16 Jul. 2021, www.miquido.com/blog/what-are-voice-assistants/.
Stroink-Skillrud, Donata. “Sharing vs. Selling Personal Information (Data).” *Termageddon*, 13 Sept. 2023, termageddon.com/sharing-vs-selling-personal-information-data/.
Thakkar, Abhi K., and Vijay Ukani. “A Proficient and Economical Approach for IoT-Based Smart Doorbell System.” *Advances in Data-Driven Computing and Intelligent Systems*, *Lecture Notes in Networks and Systems*, vol. 698, pp. 69-81, 4 Aug. 2023, *Springer*, doi:10.1007/978-981-99-3250-4_6.

---. "IoT-Based Smart Doorbell: A Review on Technological Developments." *Next Generation of Internet of Things*, *Lecture Notes in Networks and Systems*, vol. 445, pp. 219-229, 27 Sep. 2022, *Springer*, doi:10.1007/978-981-19-1412-6_18.
"Thermostat." *New World Encyclopedia*, www.newworldencyclopedia.org/entry/Thermostat.
Thormundsson, Bergur. "Virtual Assistant Technology – statistics & facts." *Statista*, 10 Jan. 2024, www.statista.com/topics/5572/virtual-assistants.
Tierie, Gerrit. "Mechanical instruments made by Drebbel." *Cornelis Drebbel (1572-1633)*, H. J. Paris, 1932, *Google Books*, books.google.com/books/about/Cornelis_Drebbel_1572_1633.html?id=d00CzgEACAAJ.
Trethewey, Ross. "How to Choose a Smart Doorbell." *This Old House*, www.thisoldhouse.com/shop-smart-home-tech/22750960/how-to-choose-a-smart-doorbell.
Trujillo, Mario. "Americans Deserve More Than the Current American Privacy Rights Act." *Electronic Frontier Foundation*, 16 Apr. 2024, www.eff.org/deeplinks/2024/04/americans-deserve-more-current-american-privacy-rights-act.
Tuohy, Jennifer Pattison. "The CSA is addressing smart home data privacy, and it's about damn time." *The Verge*, 22 Feb. 2023, www.theverge.com/2023/2/22/23610081/smart-home-matter-data-privacy-certification-csa.
"US Smart Home Statistics (2019-2028)." *Oberlo*, www.oberlo.com/statistics/smart-home-statistics.
van Schendel, Olenka. "Data masking: Anonymisation or pseudonymization?" *GRC World Forums*, www.grcworldforums.com/data-management/data-masking-anonymisation-or-pseudonymisation/12.article.
Velimirovic, Andreja. "Data Encryption in Use Explained." *PhoenixNAP*, 16 Nov. 2023, phoenixnap.com/blog/encryption-in-use.
Vigderman, Aliza, and Gabe Turner. "How Do Motion Detectors Work?" *Security*, 26 Jan. 2024, www.security.org/home-security-systems/motion-detectors/.
Viljoen, Salomé. "The Promise and Pitfalls of the California Consumer Privacy Act." *Cornell Tech*, 19 Feb. 2021, www.dli.tech.cornell.edu/post/the-promise-and-pitfalls-of-the-california-consumer-privacy-act.
Weizenbaum, Joseph. "ELIZA--A Computer Program for the Study of Natural Language Communication Between Man and Machine." *Communications of the ACM*, vol. 9, no. 1, 1 Jan. 1966, pp. 36-45. *ACM Digital Library*, doi:10.1145/365153.365168.
"What are the Differences Between Anonymisation and Pseudonymisation?" *Privacy Company*, 6 Mar. 2023, www.privacycompany.eu/blog/what-are-the-differences-between-anonymisation-and-pseudonymisation.
"What is an AI voice assistant?" *LivePerson*, 8 Dec. 2022, www.liveperson.com/blog/what-is-an-ai-voice-assistant.
Wixted, Sam. "How Does Geofencing Technology Work?" *WebFX*, www.webfx.com/blog/marketing/how-does-geofencing-technology-work.
Yang, Guang. "An Overview of Current Solutions for Privacy in the Internet of Things." *Frontiers in Artificial Intelligence*, vol. 5, 3 Mar, 2022, doi:10.3389/frai.2022.812732.
Yasar, Kinza. "virtual assistant (AI assistant)." *TechTarget*, Oct. 2023, www.techtarget.com/searchcustomerexperience/definition/virtual-assistant-AI-assistant.

Zahn, Max. “Collection of voice data for profit raises privacy fears.” *ABC News*, 18 Jan. 2023, abcnews.go.com/Technology/collection-voice-data-profit-raises-privacy-fears/story?id=96363792.